RbA: Segmenting Unknown Regions Rejected by All

Nazir Nayal[1] Mısra Yavuz[1] João F. Henriques[2] Fatma Güney[1]
[1]KUIS AI Center and Department of Computer Engineering, Koç University [2]University of Oxford
{nnayal17, myavuz21, fguney}@ku.edu.tr, joao@robots.ox.ac.uk

Abstract

Standard semantic segmentation models owe their success to curated datasets with a fixed set of semantic categories, without contemplating the possibility of identifying unknown objects from novel categories. Existing methods in outlier detection suffer from a lack of smoothness and objectness in their predictions, due to limitations of the per-pixel classification paradigm. Furthermore, additional training for detecting outliers harms the performance of known classes. In this paper, we explore another paradigm with region-level classification to better segment unknown objects. We show that the object queries in mask classification tend to behave like one vs. all classifiers. Based on this finding, we propose a novel outlier scoring function called RbA by defining the event of being an outlier as being rejected by all known classes. Our extensive experiments show that mask classification improves the performance of the existing outlier detection methods, and the best results are achieved with the proposed RbA. We also propose an objective to optimize RbA using minimal outlier supervision. Further fine-tuning with outliers improves the unknown performance, and unlike previous methods, it does not degrade the inlier performance. Project page: https://kuis-ai.github.io/RbA

1. Introduction

We address the problem of semantic segmentation of unknown categories. Detecting novel objects, for example, in front of a self-driving vehicle, is crucial for safety yet very challenging. The distribution of potential objects on the road has a long tail of unknowns such as wild animals, vehicle debris, litter, etc., manifesting in small quantities on the existing datasets [73, 7, 18]. The diversity of unknowns in terms of appearance, size, and location adds to the difficulty. In addition to the challenges of data, deep learning has evolved around the closed-set assumption. Most existing models for category prediction owe their success to curated datasets with a fixed set of semantic categories. These models fail in the open-set case by over-confidently assigning the labels of known classes to unknowns [33, 58].

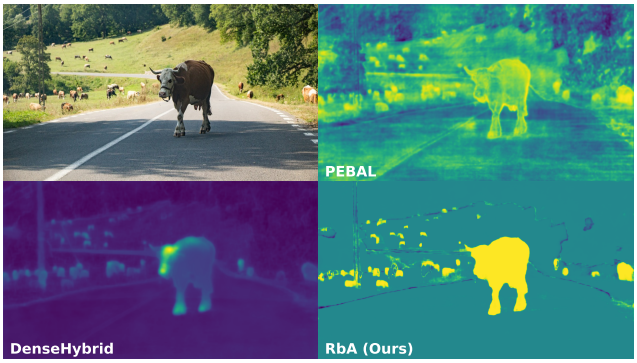


Figure 1: **Preserving objectness and eliminating noise.** While state-of-the-art methods PEBAL [65] and DenseHybrid [25] suffer from a lack of smoothness and objectness with high false positive rates, our method RbA clearly segments the unknown objects and reduces false positives by eliminating uncertainty at semantic boundaries and in ambiguous background regions.

The existing approaches to segmenting unknowns can be divided into two depending on whether they use supervision for unknown objects or not. In either case, the model has access to known classes during training, i.e. inlier or in-distribution, and the goal is to identify the pixels belonging to an unknown class, i.e. anomalous, outlier, or out-of-distribution (OoD). Earlier approaches resort to an ensemble of models [41] or Monte Carlo dropout [23] which require multiple forward passes, therefore costly in practice. More recent approaches use the maximum class probability [35] predicted by the model as a measure of its confidence. However, this approach requires the probability predictions to be calibrated, which is not guaranteed [64, 58, 26, 54, 39]. In the supervised case, the model can utilize outlier data to learn a discriminative representation, however, outlier data is limited. Typically, another dataset from a different domain is used for this purpose [12], or outlier objects are artificially added to driving images [25, 65].

The existing methods in outlier detection suffer from a lack of smoothness and objectness in the OoD predictions as shown in Fig. 1. This is mainly due to the limitations

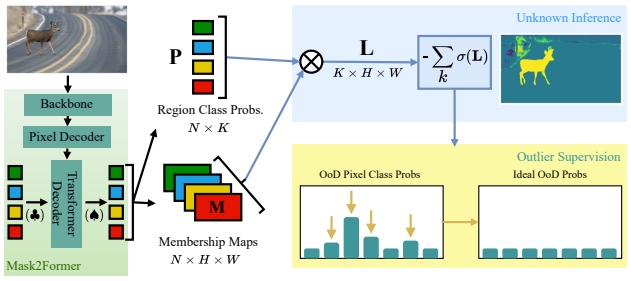


Figure 2: **Overview.** This figure provides an illustration of our proposed outlier scoring function RbA and the objective to optimize it as defined in (6). The class logit scores $\mathbf{L}$ are aggregated as the product of region class probabilities $\mathbf{P}$ and mask predictions $\mathbf{M}$ pooled over all regions. We define the RbA as the probability of not being assigned to any of the known classes. With the proposed objective, we push the probabilities of known classes down, in the outlier pixels.

of the per-pixel classification paradigm that previous OoD methods are built on. In this paper, we explore another paradigm with region-level classification to better segment objects. To that end, we use mask-classification models, such as Mask2Former [15] that are trained to predict regions and then classify each region rather than individual pixels. This endows our method with spatial smoothness, learned by region-level supervision. We discover the properties of this family of models which allow better calibration of confidence values. Then, we exploit these properties to boost the performance of the existing OoD methods that rely on predicted class scores such as max logit [34] and energy-based ones [25, 65, 50].

The existing methods also suffer from high false positive rates due to failing to separate the sources of uncertainty, especially on datasets in the wild such as Road Anomaly [49]. For example, on the boundaries, segmentation models typically predict weak scores for the two inlier classes separated by the boundary, causing these regions to be confused as OoD by score-based methods [34]. Based on exploring the behavior of object queries in mask classification, we find that most of the object queries tend to behave like one vs. all classifiers. Consequently, we propose a novel outlier scoring function based on this one vs. all behavior of object queries. We define the event of a pixel being an outlier as being rejected by all known classes. In other words, we define being an outlier as a complementary event whose probability can be expressed in terms of the known class probabilities. We show that this scoring function can eliminate irrelevant sources of uncertainty as in the case of boundaries, resulting in a considerably lower false positive rate on all datasets.

The state-of-the-art methods in OoD [25, 65] utilize outlier data for supervision. While better unknown segmentation can be achieved, it comes at the expense of lower closed-set performance. Unfortunately, this unintended consequence is not desirable since the primary objective of unknown segmentation is to identify unknowns while still accurately recognizing known classes without compromising the inlier performance.

We propose an objective to optimize the proposed outlier scoring function using a limited amount of outlier data. By fine-tuning a very small portion of the model with this objective, our method outperforms the state-of-the-art on challenging datasets with high distribution shifts such as Road Anomaly [49] and SMIYC [11]. Notably, we achieve this without affecting the closed-set performance. Our contributions can be summarized as follows:

- We postulate and study the inherent ability of mask classification models to express uncertainty, and use this strength to boost the performance of several existing OoD segmentation methods.
- Based on our finding that object queries behave approximately as one vs. all classifiers, we propose a novel outlier scoring function that represents the probability of being an outlier as not being any of the known classes. The proposed scoring function helps to eliminate uncertainty in ambiguous inlier regions such as semantic boundaries.
- We propose a loss function that directly optimizes our proposed scoring function using minimal outlier data. The proposed objective exceeds the state-of-the-art by only fine-tuning a very small portion of the model without affecting the closed-set performance.

2. Related Work

Semantic Segmentation Paradigms: Since the success of Fully Convolutional Networks (FCN) [62], semantic segmentation architectures have revolved around the per-pixel classification paradigm. This paradigm has been extensively studied to increase the closed-set performance with various convolution and pooling operations [13, 14, 19, 80, 71], and by aggregating multi-scale contextual information [74, 75]. Recent work shifted towards transformer-based architectures [70, 63, 76, 82, 72] and attention mechanisms [29, 43, 81, 30, 44, 37, 22].

On the other hand, mask classification has been mainly adopted by instance segmentation and object detection models [31, 28, 8] since it allows pixels to belong to multiple proposals and provides the flexibility to detect a variable number of objects in the scene. Max-DeepLab [67] employs mask classification for panoptic segmentation but with many auxiliary losses. Although some earlier efforts have been made to apply mask classification to semantic segmentation [9, 28], they were quickly outperformed by the per-pixel methods until recently. MaskFormer variants [16, 15] apply

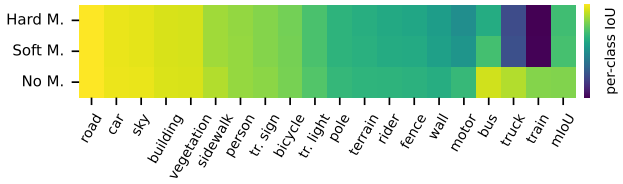


Figure 3: **Masking object queries.** We show the impact on per-class IoU on Cityscapes [18] when using two types of masking: hard masking without any interactions between the queries (top) and soft masking by allowing interactions in the transformer decoder (middle) compared to the original model without masking (bottom). The constant color in most columns shows that most of the object queries can independently segment their particular classes.

query-based mask classification and attention to obtain a unified segmentation model which shows competitive performance with specialized semantic and instance segmentation architectures across benchmarks [47, 18, 83, 57].

Unsupervised Anomaly Segmentation: Unsupervised methods utilize their knowledge about inlier data to detect anomalies at inference time. Early work measures uncertainty based on the observation that anomaly samples typically result in low-confidence predictions. The uncertainty of a model can be estimated through maximum softmax probabilities [35, 46], ensembles [41], Bayesian approximation [55], Monte Carlo dropout [23], or by learning to estimate its confidence [20]. However, posterior probabilities of a closed-set model are not necessarily calibrated, leading to overconfident predictions on unseen categories [64, 58, 26, 54, 39]. Therefore, follow-up work focuses on making a clear distinction between inliers and outliers by using true class probabilities [17], unnormalized logits instead of softmax probabilities [34], standardized class-wise logits [40], and the distance to learned prototypes of known classes [10]. Overall, unsupervised approaches are typically efficient without any extra training but they are inherently limited to which extent they can separate inliers and outliers due to a lack of supervision with outlier data.

Deep generative models are also used for unsupervised anomaly segmentation. Early methods are primarily based on density estimation [42, 59] while subsequent works focus on reconstruction. Several works rely on the predicted segmentation maps to resynthesize [49, 68, 27, 66] or inpaint the inputs [48] and measure discrepancy with comparison networks. Others apply localized adversarial attacks [1], synthesize negatives using normalizing flows [24], or combine Gaussian mixture models with discriminative representation learning [45]. Generative methods are typically impractical for real-time safety-critical applications due to high computational costs and long inference times. Additional comparison modules and the change in input distributions require extra training. Moreover, synthesized unknowns often do not generalize well to real anomalies [24]. Several works [56, 61, 79] show that generative models tend to estimate high likelihoods on out-of-distribution samples.

Anomaly Segmentation with Outlier Supervision: Out-of-distribution data can be used to regularize the model's feature space by learning a representation of unknowns. With the increase in the availability of wide-range datasets, initial approaches utilize generic datasets such as ImageNet [60] for OoD. Given data, OoD detection can be simply treated as binary classification [2, 3]. Outlier data can also be used to estimate the distributional uncertainty of OoD samples [52] or to fine-tune parametrized OoD detectors [36]. The energy score has been proposed as a better alternative to softmax in terms of separation [50]. SynBoost [5] is a supervised image resynthesis method that treats void regions as anomalies to obtain an uncertainty signal.

Recent work uses a subset of COCO [47] or ADE20K [83], either as entire images [12] or after cut-and-paste into the inlier scenes [65, 25]. Meta-OoD [12] maximizes the entropy on outliers, whereas PEBAL [65] learns adaptive energy-based penalties by abstention learning. Combining likelihood and posterior evaluation, DenseHybrid [25] achieves state-of-the-art results. However, for each benchmark, different models are fine-tuned using multiple datasets [83, 57, 78] with high distribution shifts, resulting in a higher degree of supervision and variety. Our model, on the other hand, can achieve better performance across benchmarks by using the same model and only a small subset of COCO [47] for fine-tuning.

One vs. All Classification: Previous work trains one vs. all classifiers for unknown segmentation [21, 4] in addition to a standard multi-class classifier. The class probabilities obtained by the one vs. all classifiers are merely used for calibrating the multi-class classifier. The outlier scoring function is calculated as the negative maximum class probability of the calibrated probabilities. On the contrary, our method utilizes the implicit one vs. all behavior of mask classifiers for explicitly defining the probability of being an outlier.

3. Methodology

In this work, we address the limitations of the existing OoD methods by using mask classification. We first perform an analysis of the mask classification models. Then, based on our analysis, we propose a novel scoring function to exploit the implicit one vs. all behavior in these models. We mathematically define the probability of being an outlier probability as the "none of the above" option for the model. Finally, we propose a training objective to optimize our proposed scoring function with minimal outlier data.

3.1. Mask Classification

We build our method on top of the Mask2Former architecture [15], which is an improved version of the initial MaskFormer [16]. We give only a brief overview to make the discussion self-contained; please refer to Cheng et al. [15] for details. Mask2Former consists of three main parts: the backbone, the pixel decoder, and the transformer decoder. The backbone processes the input image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ to extract features at multiple scales. Then, the pixel decoder further processes the multi-scale features to produce high-resolution per-pixel features $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^{C_p\times H\times W}$. The transformer decoder takes the resulting multi-scale features $\{\mathbf{f}_i\}_{i=1}^{D}$, where D is number of scales, as well as N learnable object queries $\mathbf{Q} \in \mathbb{R}^{N\times C_q}$, where C_p and C_q denote the embedding dimensions. At each layer of the transformer decoder, object queries are refined by interacting with each other and with one of the scales $\mathbf{f}_i$ in a round-robin order.

The refined object queries are first processed with a 3-layer MLP, resulting in $\mathbf{Q}_p \in \mathbb{R}^{N\times C_p}$ to predict N regions. The binary masks for all regions are obtained by multiplying $\mathbf{Q}_p$ with pixel features $\mathbf{F}$ and applying a sigmoid σ to the result:

$$\mathbf{M}(\mathbf{x}) = \sigma(\mathbf{Q}_p\, \mathbf{F}(\mathbf{x})) \tag{1}$$

$\mathbf{M}(\mathbf{x}) \in \mathbb{R}^{N\times H\times W}$ represents the membership score of each pixel belonging to a region. In parallel, refined object queries are fed to a linear layer followed by softmax to produce posterior class probabilities $\mathbf{P}(\mathbf{x}) \in [0,1]^{N\times K}$ of K classes.

In contrast to per-pixel semantic segmentation, the ground truth masks are partitioned into multiple binary masks such that each mask contains all the pixels that belong to a class. Then, bipartite matching is used to match every ground truth mask to an object query using region prediction and classification losses as the cost. For region prediction, a weighted combination of dice loss [53] and binary cross-entropy is applied to the binary mask predictions. For classification, cross-entropy loss is used. In inference, the class scores or logits $\mathbf{L}(\mathbf{x}) \in \mathbb{R}^{K\times H\times W}$ are calculated as the product of mask predictions with class predictions by broadcasting the class prediction to all the pixels within the region:

$$\mathbf{L}(\mathbf{x}) = \sum_{n=1}^{N} \mathbf{P}_n(\mathbf{x})\mathbf{M}_n(\mathbf{x}) \tag{2}$$

3.2. Independence of Object Queries

The logit term $\mathbf{L}$ as defined in Eq. 2 has a deeper interpretation because of its structure. In essence, $\mathbf{L}$ aggregates weighted votes over all object queries to decide whether the pixel belongs to a certain class. During training, the ground truth binary map of each class is matched to an object query using bipartite matching. Therefore, we find that object queries specialize in predicting a specific class after convergence. We empirically verify this behavior on another driving dataset (after training on Cityscapes), the validation set of BDD100K [73]. We identify which class each object query specializes in by counting how many times it predicts a certain class with high confidence, e.g. greater than 98%, see Supplementary for visualization of this specialized behavior.

After identifying which object query predicts which class, we test their independence, i.e. the ability of each object query to predict its class without relying on other object queries. To evaluate the predictions of class k, we mask out all but its specialized query. We do this in one of two ways: 1) before the transformer decoder (♣ in Fig. 2), where each object query only interacts with the image features and not with each other (*hard masking*), or 2) after the transformer decoder (♠ in Fig. 2), allowing queries to interact with each other weakly (*soft masking*). In both cases, only the specialized query is used to predict the mask and class.

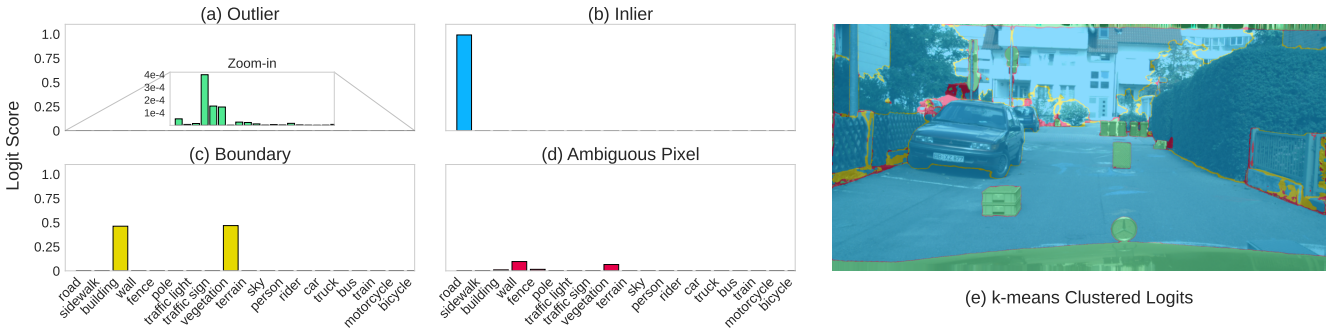
Fig. 3 shows per-class IoU scores on Cityscapes [18] using both strategies compared to the original model without any masking. We observe that the performance in most of the classes is not affected compared to the original model. The drop in performance occurs only in rare classes, such as train, truck, or bus, indicating that their object queries rely on other queries in prediction, which explains the slightly better performance of soft masking than hard masking. This behavior of object queries resembles multiple independent binary classifiers, implicitly embedded in a single model. Note that this is only for analysis purposes and not part of the proposed method.

3.3. Rejected by All (RbA) Scoring Function

Inspired by the independent behavior of object queries, we propose to model the prediction of each class as an independent binary classification problem. We consider it as K one vs. all classifiers where the predicted score for each class is independently modeled as follows:

$$p(\mathbf{y} = k|\mathbf{x}) = \sigma(\mathbf{L}_k(\mathbf{x})) \tag{3}$$

where $\mathbf{y} \in \mathcal{K}^{H\times W}$ is a random variable representing the predicted class label over a predefined set of known classes $\mathcal{K} = \{1, \ldots, K\}$ and σ is a normalization function applied to per class logits to map them to a probability, i.e. a value between 0 and 1. Based on this definition, we assume that the latent space is partitioned into $K + 1$ mutually exclusive and exhaustive regions, such that the label $K + 1$ represents the region where the outliers reside, rejected by all other known classes. By assuming the mutual exclusiveness of per-class probabilities for a given input, we can define the



(e) k-means Clustered Logits

Figure 4: **Categorizing the behavior of logits.** Outlier pixels receive extremely low votes from object queries (a). Inlier pixels receive a high vote from a single object query (b). Boundary pixels separating two inlier classes receive moderate votes from both object queries (c). Ambiguous regions receive weak votes from multiple object queries (d). Clustering clearly outlines these four behaviors of logits (e). Pixels in (e) are color-coded with the same colors of the respective histograms (a-d).

probability map of an input $\mathbf{x}$ being an outlier as follows:

$$\begin{aligned} p(\mathbf{y}=K+1|\mathbf{x}) &= 1-p\left(\cup_{k=1}^{K}\mathbf{y}=k|\mathbf{x}\right) \\ &= 1-\sum_{k=1}^{K}p(\mathbf{y}=k|\mathbf{x}) \\ &= 1-\sum_{k=1}^{K}\sigma(\mathbf{L}_k(\mathbf{x})) \end{aligned} \tag{4}$$

Dropping the constant 1 (which does not affect the optimization), we define our outlier scoring function RbA:

$$\mathrm{RbA}(\mathbf{x}) = -\sum_{k=1}^{K}\sigma(\mathbf{L}_k(\mathbf{x})) \tag{5}$$

We choose σ to be the `tanh` function to map $\mathbf{L}_k > 0$ more uniformly to the range $[0, 1]$.

3.4. Fine-tuning with Minimal Outlier Supervision

We propose to regularize our scoring function with supervision from a small amount of synthetically created outlier data. Our goal is to improve the OoD segmentation while preserving the closed-set performance. Without retraining the entire model, we only fine-tune the mask prediction MLP and classification layer after the transformer decoder (see Supplementary), which constitutes only 0.21% of the total model parameters. For OoD data, we use a modified version of Anomaly Mix proposed in [65], where objects from COCO dataset [47] are randomly cut and pasted on Cityscapes images [18]. We regularize the scores by *maximizing RbA for outlier pixels*, with a squared hinge loss. This is also equivalent to suppressing high-confidence probabilities of known classes for outlier pixels as shown in Fig. 2.

The loss is formally defined as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{RbA}} &= \sum_{\mathbf{x}\in\Omega_{out}}\left(\max(0, \alpha - \mathrm{RbA}(\mathbf{x}))\right)^2 \\ &= \sum_{\mathbf{x}\in\Omega_{out}}\max\left(0, \alpha + \sum_{k=1}^{K}\sigma(\mathbf{L}_k(\mathbf{x}))\right)^2 \end{aligned} \tag{6}$$

where Ω_{out} is the set of outlier pixels. We experimentally set the hyper-parameter α to 5 but we found that any $\alpha > 0$ works well in practice. See Supplementary for an ablation.

3.5. Analyzing RbA

The term $\mathbf{L}$ in Eq. 2 aggregates the independent decisions of object queries about whether a pixel belongs to a certain class. Based on this behavior, we can identify several distinct modes of $\mathbf{L}$. We cluster the logits over classes at each pixel, i.e. K-dimensional vector, using k-means to characterize the modes, visualized in Fig. 4e. For an inlier pixel, only a single object query votes for it with high confidence (Fig. 4b), whereas true outlier pixels do not receive any votes from any object query (Fig. 4a). These two modes, especially the outliers in Fig. 4a, due to the one vs. all behavior, reduces the overconfidence issue in the existing outlier scoring functions used in max logit [34] and energy-based methods [25, 65, 50], therefore improve their results (Table 3).

However, there are pixels that disrupt the separability between the inliers and the outliers which max logit and energy-based methods fail to capture. For example, pixels on a boundary between two inlier classes (Fig. 4c) or ambiguous background pixels (Fig. 4d) end up with a higher anomaly score than the inliers, causing them to be mistaken as an outlier. Boundary and ambiguous regions are commonly characterized by having more than one weak vote from object queries. Since RbA aggregates votes from all classes, summing these weak votes results in a lower outlier score and hence reduces the false positive rate. Fig. 5 high-

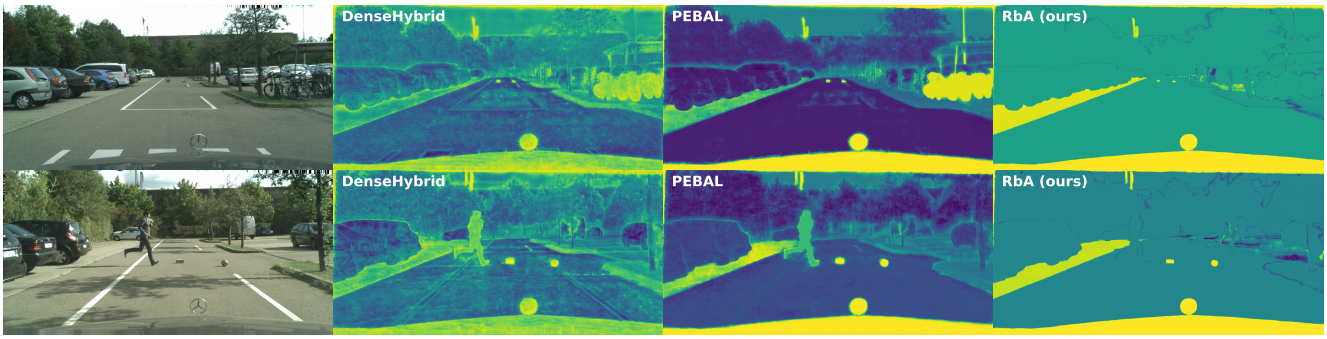


Figure 5: **Visual comparison to the state-of-the-art**. We show visualizations of outlier score maps predicted by our method, RbA compared to the ones predicted by state-of-the-art methods PEBAL [65] and DenseHybrid [25] trained using the same architecture as the RbA for a fair comparison. The other two methods falsely identify the inlier classes such as person and bike, which are correctly ignored by the proposed RbA. It is noteworthy that RbA also eliminates the false positives in the background region, especially at the boundaries separating inliers and better preserves the smoothness of the outlier map compared to other methods despite being also trained with mask classification.

Method	OoD Data	Extra Net.	Anomaly Track					Obstacle Track				
			AP ↑	FPR ↓	sIoU gt ↑	PPV ↑	mean F1 ↑	AP ↑	FPR ↓	sIoU gt ↑	PPV ↑	mean F1 ↑
Emb. Density[6]	✗	✗	37.5	70.8	33.9	20.5	7.9	0.8	46.4	35.6	2.9	2.3
JSRNet[66]	✗	✓	33.6	43.9	20.2	29.3	13.7	28.1	28.9	18.6	24.5	11.0
Road Inpain.[48]	✗	✓	-	-	-	-	-	54.1	47.1	**57.6**	39.5	36.0
Image Resyn.[49]	✗	✓	52.3	25.9	39.7	11.0	12.5	37.7	4.7	16.6	20.5	8.4
ObsNet[1]	✗	✓	75.4	26.7	44.2	**52.6**	**45.1**	-	-	-	-	-
NFlowJS[24]	✗	✓	56.9	34.7	36.9	18.0	14.9	85.6	**0.4**	45.5	49.5	**50.4**
RbA (Ours)	✗	✗	**86.1**	**15.9**	**56.3**	41.4	42.0	**87.8**	3.3	47.4	**56.2**	**50.4**
Max. Entropy[44]	✓	✗	85.5	15.0	49.2	39.5	28.7	85.1	0.8	47.9	**62.6**	48.5
DenseHybrid[25]	✓	✗	78.0	**9.8**	54.2	24.1	31.1	87.1	**0.2**	45.7	50.1	50.7
PEBAL[65]	✓	✗	49.1	40.8	38.9	27.2	14.5	5.0	12.7	29.9	7.6	5.5
SynBoost[5]	✓	✓	56.4	61.9	34.7	17.8	10.0	71.3	3.2	44.3	41.8	37.6
RbA (Ours)	✓	✗	**90.9**	11.6	**55.7**	**52.1**	**46.8**	**91.8**	0.5	**58.4**	58.8	**60.9**

Table 1: **Results on the SMIYC benchmark.** We report results on both the anomaly and the obstacle track. Both tracks cover a wide variety of scenarios and unknown objects. We report both pixel-level (AP, FPR@95) and component-level metrics (sIoU, PPV, mean F1). We show the results with (lower part) and without (upper part) outlier supervision with the best in bold and the second best underlined for each. We mark the methods that use OoD data during training or an extra network auxiliary to the segmentation network.

lights the differences between the anomaly maps predicted by RbA and the state-of-the-art methods, also trained using Mask2Former. Note that RbA assigns low outlier scores at boundaries separating known classes.

4. Experiments

4.1. Datasets

We train the model on Cityscapes [18], which consists of 2975 training and 500 validation images. It contains 19 classes which are considered as inliers in anomaly segmentation benchmarks. The classes in the dataset can be seen in Fig. 3. For evaluation, we consider multiple datasets. First, Segment Me If You Can (SMIYC) benchmark [11] with two datasets: anomaly track and obstacle track. The anomaly track has 100 images that contain unknown objects of various sizes in diverse environments. The obstacle track contains 412 images with typically small unknown objects on the road, 85 of which are taken at night and in adverse weather conditions. Both datasets are characterized by a high domain shift compared to Cityscapes, making them particularly challenging. Road Anomaly [49] is an earlier and smaller version of SMIYC. It consists of 60 images with diverse objects in diverse environments. We also report results on the Fishyscapes Lost&Found [6], which has 100 validation and 275 test images. The domain of this dataset is similar to that of Cityscapes, and the anomalous objects are mostly small and less diverse compared to other datasets.

Method	OoD Data	Extra Net.	Road Anomaly			FS LaF		
			AUC ↑	AP ↑	FPR ↓	AUC ↑	AP ↑	FPR ↓
MSP (R101) [35]	✗	✗	73.76	20.59	68.44	86.99	6.02	45.63
Entropy (R101) [35]	✗	✗	75.12	22.38	68.15	88.32	13.91	44.85
Mahalanobis [42]	✗	✗	76.73	22.85	59.20	92.51	27.83	30.17
SML [40]	✗	✗	81.96	25.82	49.74	96.88	36.55	14.53
GMMSeg (SF) [45]	✗	✗	89.37	57.65	44.34	**97.83**	50.03	12.55
SynthCP [68]	✗	✓	76.08	24.86	64.69	88.34	6.54	45.95
RbA (Ours)	✗	✗	**95.60**	**78.45**	**11.83**	96.43	**60.96**	**10.63**
Maximized Entropy [12]	✓	✗	-	-	-	93.06	41.31	37.69
PEBAL [65]	✓	✗	88.85	44.41	37.98	98.52	64.43	6.56
SynBoost (WRN38) [5]	✓	✓	81.91	38.21	64.75	96.21	60.58	31.02
RbA (Ours)	✓	✗	**97.99**	**85.42**	**6.92**	**98.62**	**70.81**	**6.30**

Table 2: **Results on Road Anomaly and Fishyscapes LaF.** We show the results with (lower part) and without (upper part) outlier supervision with the best in bold and the second best underlined for each. We report the results of RbA both with and without outlier supervision. Our method RbA notably improves the results in all metrics on both datasets.

4.2. Experimental Setup

Implementation Details: We follow the setup of [15] for closed-set training on Cityscapes. We use the Swin-B [51] architecture as the backbone. Differently, we use only one decoder layer in the transformer decoder instead of nine (see Supplementary). For outlier supervision, we fine-tune the mask prediction MLP and the classification layer for 2K iterations with a batch size of 16 using the standard loss functions used in [15] in addition to the RbA loss defined in Eq. 6. Previous work [65] samples 300 new images every epoch out of 40K COCO images with objects different than Cityscapes inliers. Differently, we sample 300 images only at the beginning and fix them, then at each iteration, an image is randomly chosen and pasted on inlier images with probability p_{out}. We experimentally set p_{out} to 0.1.

Evaluation Metrics: For comparison to previous methods on the Road Anomaly and the Fishyscapes, we report Average Precision (AP), Area under ROC Curve (AuROC), and False Positive rate at the threshold of 95% True Positive Rate (FPR@95). On SMIYC, the public benchmark reports AP and FPR@95 for per-pixel metrics as well as component-level metrics that are designed to measure the statistics of detected objects [11]. Specifically, the proposed metrics aim at quantifying true positives (TP), false negatives (FN), and false positives (FP) of detected unknown objects. Please see the benchmark paper [11] for more details on these metrics.

4.3. Quantitative Results

4.3.1 Segment Me If You Can Benchmark

Table 1 shows the results on anomaly and obstacle tracks of the public SMIYC benchmark. Without outlier supervision, RbA outperforms all the models, including those trained with outlier supervision, in AP while maintaining a competitive FPR@95. In terms of component metrics, the gains with RbA are more pronounced, which is due to an improved ability to characterize objectness, compared to the previous work. With outlier supervision, the performance gap improves with respect to the previous best method consistently across both tracks: +5.4% and +4.7% in AP and +1.7% and +10.2% in mean F1 for anomaly and obstacle tracks respectively. DenseHybrid [25] achieves a slightly better FPR@95 on the anomaly and obstacle tracks, but RbA achieves significantly better AP, +12.9% and +4.7% respectively, and better performance in all component-level metrics. ObsNet [1] has impressive performance at the component-level, however, not at the pixel-level. RbA consistently performs well across both tracks in both pixel and component-level metrics.

SMIYC is characterized by high domain shift and diversity of objects in terms of size and appearance, making it particularly challenging. While some methods, like DenseHybrid [25], rely on highly diverse data when fine-tuning, RbA with mask classification shows that outlier supervision is not necessary to perform well under domain shift, thereby surpassing the limitations of the existing methods.

4.3.2 Road Anomaly & Fishyscapes LaF

Table 2 shows the results on the Road Anomaly [49] and the Fishyscapes Lost and Found (LaF) validation set [6]. Without outlier supervision, RbA improves the state-of-the-art significantly in almost all metrics on both datasets, even outperforming methods with outlier supervision in some metrics. With minimal supervision from a limited number of outlier objects, we obtain significant performance gains without hurting the closed-set performance (Table 3).

5. Ablation Study

We ablate our contributions to justify our decision choices with the scoring function, loss function, and backbone. First, we show that mask classification improves the performance

of the existing methods in OoD, but RbA better utilizes its potential. We then report the performance of the squared hinge loss compared to alternative loss functions. Lastly, we experiment with different backbones and show that optimizing for RbA improves the results with different backbones.

Other Methods with Mask Classification: To clearly demonstrate the effectiveness of our method and decouple it from the gains obtained by the Mask2Former, we report the results of other SOTA methods using Mask2Former, including PEBAL [65], DenseHybrid [25], and Max Logit [34] in Table 3. The existing OoD methods perform well with Mask2Former, for example, the performance of PEBAL significantly improves compared to the official results reported in Table 2. As discussed in Section 3.2, the improvement comes from reducing the overconfidence issue owing to the independent behavior of object queries. Our method, RbA, performs better than the other methods in all metrics. More importantly, we achieve this performance in OoD without affecting the closed-set performance, unlike the other methods such as PEBAL causing a significant drop in mIoU. This experiment shows that we can better utilize the properties of mask classification with RbA.

Alternative Loss Functions: We verify our choice of loss function which is a squared hinge loss by optimizing our method with other commonly used loss functions. As can be seen in Table 4, squared hinge loss outperforms other loss functions. Mean Squared Error (MSE) and L1 result in a higher false positive rate. We define the OoD as a binary classification problem and optimize it with BCE by using the outlier score given by the RbA as the positive class logit. While it improves the FPR compared to MSE and L1, it performs worse than the squared hinge loss in all metrics. Using KL Divergence, we minimize the distance between class probabilities of outlier pixels from a fixed distribution with maximum entropy. It performs comparably in FPR but poorly in AP, especially on the Fishyscapes LaF. Detailed formulations can be found in Supplementary.

Method	mIoU ↑	Road Anomaly		FS LaF	
		AP ↑	FPR ↓	AP ↑	FPR ↓
Max Logit [34]	**82.25**	77.31	16.90	58.52	22.14
PEBAL [65]	75.32	79.01	7.21	62.67	25.60
DH [25]	80.27	78.57	12.28	36.94	21.12
RbA (Ours)	82.20	**85.42**	**6.92**	**70.81**	**6.30**

Table 3: **Other methods with Mask2Former.** We show the performance of the state-of-the-art methods with Mask2Former. Our method RbA achieves the best results in all metrics with a clear margin, without affecting the closed-set performance, unlike previous methods. The mIoU before fine-tuning is shown in the first row of the table. The rest of the models are fine-tuned from the same checkpoint.

Method	Road Anomaly		FS LaF	
	AP ↑	FPR ↓	AP ↑	FPR ↓
KL Div.	79.91	11.33	63.58	8.78
MSE	80.71	15.79	69.14	22.06
L1	80.94	15.75	67.19	20.44
BCE	80.66	10.29	64.90	6.89
RbA (Ours)	**85.42**	**6.92**	**70.81**	**6.30**

Table 4: **Ablation study on alternative loss functions.** We compare our loss function based on the squared hinge loss to other commonly used loss functions. The results show that our method with squared hinge loss (RbA) performs the best in terms of OoD segmentation.

Backbone	Road Anomaly		FS LaF	
	AP ↑	FPR ↓	AP ↑	FPR ↓
R101 [32]	38.1 / 61.9	82.7 / 37.2	30.1 / 47.1	26.3 / 12.7
WR38 [77]	21.6 / 52.0	90.0 / 43.8	24.8 / 44.6	76.3 / 13.4
MViT [69]	57.2 / 73.1	85.8 / 24.9	47.8 / 63.7	59.8 / **6.2**
MixT [70]	65.7 / 78.1	24.6 / 12.4	40.3 / 51.3	23.0 / 17.1
Swin-B [51]	**78.5 / 85.4**	**11.8 / 6.9**	**61.0 / 70.8**	**10.6** / 6.3

Table 5: **Ablation study on the backbone.** We show the effect of varying the backbone used for feature extraction on the OoD performance. Comparing the results before and after fine-tuning with the proposed method, we observe clear improvements in the performance in all backbones.

Different Backbones: We use the same Mask2Former model with Swin-B backbone [51] in all our experiments. In Table 5, we report the results with different backbones including transformer-based Multiscale ViT (MViT) [69] and Mix Transformer (MixT) [70] as well as convolutional WideResnet38 (WR38) [77] and ResNet101 (R101) [32] backbones. We keep all the other parameters the same as the default version for a fair comparison. Fine-tuning with RbA brings consistent improvements in all metrics for all backbones. While Swin-B performs the best, R101, MViT, and MixT can still outperform previous methods on Road Anomaly and achieve competitive results on Fishyscapes LaF. This experiment shows that the proposed scoring function improves the performance regardless of the backbone.

6. Open-Set Panoptic Segmentation

We show that RbA can be extended to open-set panoptic segmentation with slight modifications. We follow the setup proposed in [38]. We train Mask2Former with ResNet50 backbone for 280K iterations using a batch size of 18.

Outlier Scoring: In order to calculate the RbA score map for a given input sample in the panoptic setting, we first calculate RbA the same way as in (5), then we apply dilation and erosion operations of kernel size 3×3 to reduce the noise

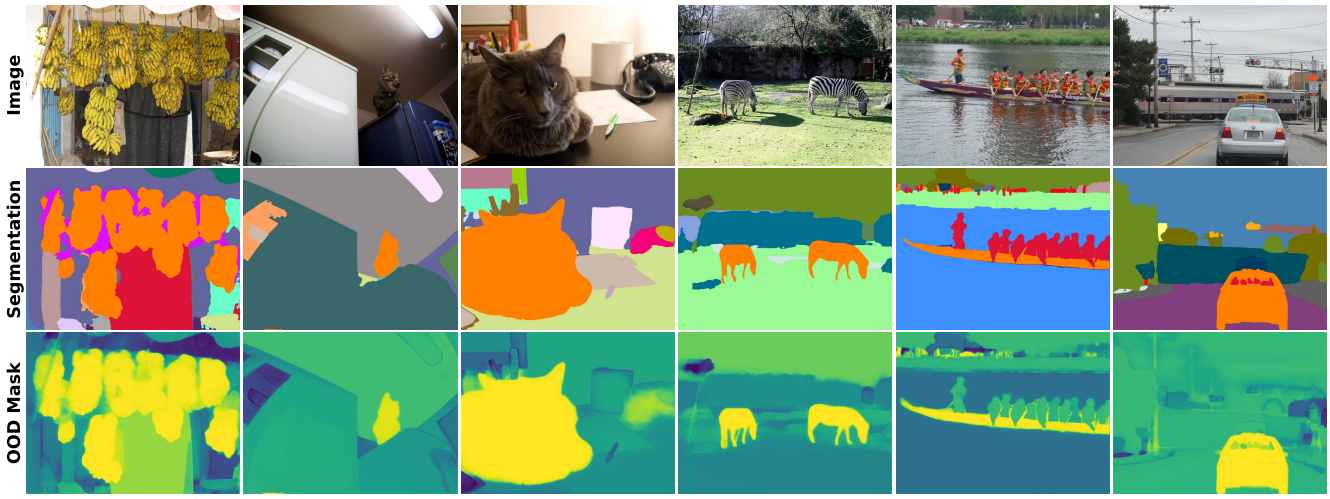


Figure 6: **Qualitative Results on COCO Open-Set Split**

Model	Known Metrics			Unknown Metrics		
	PQ	SQ	RQ	PQ	SQ	RQ
EOSPN [38]	37.4	76.2	46.2	11.3	73.8	15.3
RbA (Ours)	**47.0**	**82.2**	**56.3**	**24.8**	**79.2**	**31.4**

Table 6: **Open-set panoptic segmentation results.** We show the results for known and unknown metrics on the open-set split of COCO dataset. RbA outperforms the baseline method considerably in all metrics.

from the map. Finally, we extract the connected components from the score map and consider each component as an unknown instance.

Dataset: COCO [47] dataset with panoptic annotations is used for evaluating open-set panoptic segmentation. The dataset consists of 118K training and 5K validation samples, containing 80 thing classes and 54 stuff classes. To create the open-set setting, some classes from the known thing classes are removed from the training set and used for evaluation with the unknown metrics. The authors in [38] propose three different splits depending on the percentage of known classes removed from the thing classes. We perform the evaluation on the most difficult split which removes 20% of the known classes to use as the unknown set.

Results: Following [38], we report the Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ) computed on known things and stuff classes and unknown classes separately as shown in Table 6. RbA outperforms the baseline method EOSPN [38] significantly in both known and unknown metrics. Fig. 6 shows qualitative examples on the validation set of open-set COCO.

7. Conclusion and Future Work

In this work, we explore the potential of mask classification to segment unknown classes. We show that object queries behave like one vs. all classifiers and their independent behavior reduces the overconfidence issue in the predicted scores, resulting in improvements in the performance of the existing scoring-based methods such as max logit and energy-based methods. By treating the result of mask classification as multiple one vs. all classifiers, we propose a novel outlier scoring function called RbA defined in terms of known class probabilities. We also propose an objective to optimize the RbA with limited outlier data, obtaining significant performance gains without affecting the closed-set performance. We show that the RbA eliminates irrelevant sources of uncertainty, such as inlier boundaries and ambiguous background regions, leading to a considerable decrease in false positive rates. Moreover, our proposed method can preserve objectness and smoothness due to the region-level inductive biases learned by the mask classifier.

As this work represents an initial attempt to utilize mask classification for unknown segmentation, its properties can be further explored with potential improvements. Given the increased ability to preserve objectness, open-world incremental learning is one step closer, as unknown masks are more reliable as a source of supervision. While current efforts are limited to static image datasets, temporal or depth information can provide important cues to detect unknowns.

Acknowledgements: We thank A. Kaan Akan, Ali Safaya, Görkay Aydemir, Shadi Hamdan, Mert Çökelek, and Merve Rabia Barın for their valuable feedback. This project is funded by the Royal Society Newton Fund Advanced Fellowship (NAF\R\2202237). We also thank KUIS AI Center, the Royal Academy of Engineering (RF\201819\18\163), and Unvest R&D Center for their support.

References

[1] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In *ICCV*, 2021. 3, 6, 7

[2] Petra Bevandic, Ivan Kreso, Marin Orsic, and Sinisa Segvic. Discriminative out-of-distribution detection for semantic segmentation. *arXiv.org*, 1808.07703, 2018. 3

[3] Petra Bevandic, Ivan Kreso, Marin Orsic, and Sinisa Segvic. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *GCPR*, 2019. 3

[4] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Dense outlier detection and open-set recognition based on training with noisy negative images. *arXiv preprint arXiv:2101.09193*, 2021. 3

[5] Giancarlo Di Biase, Hermann Blum, Roland Y. Siegwart, and César Cadena. Pixel-wise anomaly detection in complex driving scenes. In *CVPR*, 2021. 3, 6, 7

[6] Hermann Blum, Paul-Edouard Sarlin, Juan I. Nieto, Roland Y. Siegwart, and César Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *IJCV*, 129:3119–3135, 2021. 6, 7

[7] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. 1

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[9] João Carreira, Rui Caseiro, Jorge P. Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 2

[10] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *ICCV*, 2021. 3

[11] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Y. Siegwart, Mathieu Salzmann, P. Fua, and Matthias Rottmann. SegmentMeIfYouCan: A benchmark for anomaly segmentation. *arXiv.org*, 2104.14812, 2021. 2, 6, 7

[12] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *ICCV*, 2021. 1, 3, 7

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40:834–848, 2018. 2

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[15] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 4, 7

[16] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 2, 4

[17] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *NeurIPS*, 2019. 3

[18] 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. In *CVPR*, 2016. 1, 3, 4, 5, 6

[19] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 2

[20] Terrance Devries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv.org*, 1802.04865, 2018. 3

[21] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, and Isabelle Bloch. One versus all for deep neural network incertitude (ovnni) quantification. *IEEE Access*, 2022. 3

[22] J. Fu, J. Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2

[23] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 1, 3

[24] Matej Grcic, Petra Bevandi'c, and Sinivsa vSegvi'c. Dense anomaly detection by robust learning on synthetic negative data. *arXiv.org*, 2112.12833, 2021. 3, 6

[25] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7, 8

[26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 1, 3

[27] David Haldimann, Hermann Blum, Roland Y. Siegwart, and César Cadena. This is not what I imagined: Error detection for semantic segmentation through visual dissimilarity. *arXiv.org*, 1909.00676, 2019. 3

[28] Bharath Hariharan, Pablo Arbeláez, Ross B. Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 2

[29] Adam W. Harley, Konstantinos G. Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *ICCV*, 2017. 2

[30] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *ICCV*, 2019. 2

[31] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 8

[33] 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 *CVPR*, 2019. 1

[34] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Xiaodong Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022. 2, 3, 5, 8

[35] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1, 3, 7

[36] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 3

[37] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, Humphrey Shi, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 2

[38] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *CVPR*, 2021. 8, 9

[39] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *NeurIPS*, 2018. 1, 3

[40] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *ICCV*, 2021. 3, 7

[41] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 1, 3

[42] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 3, 7

[43] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. In *BMVC*, 2018. 2

[44] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 2019. 2, 6

[45] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. GMMSeg: Gaussian mixture based generative semantic segmentation models. *arXiv.org*, 2210.02025, 2022. 3, 7

[46] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 3

[47] 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 *ECCV*, 2014. 3, 5, 9

[48] Krzysztof Lis, Sina Honari, P. Fua, and Mathieu Salzmann. Detecting road obstacles by erasing them. *arXiv.org*, 2012.13633, 2020. 3, 6

[49] Krzysztof Lis, Krishna Kanth Nakka, Pascal V. Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *ICCV*, 2019. 2, 3, 6, 7

[50] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *NeurIPS*, 2020. 2, 3, 5

[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 7, 8

[52] Andrey Malinin and Mark John Francis Gales. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018. 3

[53] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 4

[54] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *NeurIPS*, 2021. 1, 3

[55] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv.org*, 1811.12709, 2018. 3

[56] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2018. 3

[57] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 3

[58] Anh M Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015. 1, 3

[59] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *NeurIPS*, 2019. 3

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 3

[61] Joan Serrà, David Álvarez, V. Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv.org*, 1909.11480, 2020. 3

[62] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[63] Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 2

[64] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv.org*, 1312.6199, 2013. 1, 3

[65] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and G. Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7, 8

[66] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly de-

tection by partial image reconstruction with segmentation coupling. In *ICCV*, 2021. 3, 6

[67] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Loddon Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. 2

[68] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *ECCV*, 2020. 3, 7

[69] Hang Xiao, Ya Zhang, Kristin Dana, and Jianping Shi. Multi-scale vision transformers. In *CVPR*, 2021. 8

[70] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2, 8

[71] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 2

[72] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NeurIPS*, 2021. 2

[73] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 1, 4

[74] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 2

[75] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 2

[76] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. HRFormer: High-resolution transformer for dense prediction. *arXiv.org*, 2110.09408, 2021. 2

[77] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 8

[78] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *ECCV*, 2018. 3

[79] Lily H. Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *ICML*, 2021. 3

[80] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[81] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 2

[82] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2

[83] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 3