

# RbA: Segmenting Unknown Regions Rejected by All

Nazir Nayal<sup>1</sup> Mı̄sra Yavuz<sup>1</sup> João F. Henriques<sup>2</sup> Fatma Güney<sup>1</sup>

<sup>1</sup> KUIS AI Center, Koç University, <sup>2</sup> 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.*

## 1. Introduction

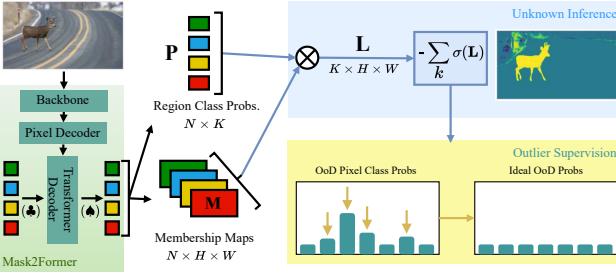
We address the problem of semantic segmentation of unknown categories. Detecting novel objects, for example, in front of a self-driving vehicle, is crucial for safety yet very challenging. The distribution of potential objects on the road has a long tail of unknowns such as wild animals, vehicle debris, litter, etc., manifesting in small quantities on the existing datasets [73, 6, 17]. 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 [40] or Monte Carlo dropout [22] 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, 38]. 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 [11], 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 [14] 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, 49].

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 [48]. 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.

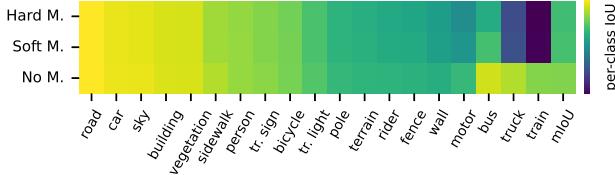
We propose an objective to optimize the proposed outlier scoring function using a limited amount of outlier data. By fine-tuning a very small portion of the model with this objective, our method outperforms the state-of-the-art on challenging datasets with high distribution shifts such as Road Anomaly [48] and SMIYC [10]. 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 [12, 13, 18, 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, 42, 81, 30, 43, 37, 21].

On the other hand, mask classification has been mainly adopted by instance segmentation and object detection models [31, 28, 7] 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 [8, 28], they were quickly outperformed by the per-pixel methods until recently. MaskFormer variants [15, 14] apply



**Figure 3: Masking object queries.** We show the impact on per-class IoU on Cityscapes [17] 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 [46, 17, 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, 45], ensembles [40], Bayesian approximation [55], Monte Carlo dropout [22], or by learning to estimate its confidence [19]. However, posterior probabilities of a closed-set model are not necessarily calibrated, leading to overconfident predictions on unseen categories [64, 58, 26, 54, 38]. Therefore, follow-up work focuses on making a clear distinction between inliers and outliers by using true class probabilities [16], unnormalized logits instead of softmax probabilities [34], standardized class-wise logits [39], and the distance to learned prototypes of known classes [9]. 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 [41, 59] while subsequent works focus on reconstruction. Several works rely on the predicted segmentation maps to resynthesize [48, 68, 27, 66] or inpaint the inputs [47] 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 [44]. 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 [49]. SynBoost [4] is a supervised image resynthesis method that treats void regions as anomalies to obtain an uncertainty signal.

Recent work uses a subset of COCO [46] or ADE20K [83], either as entire images [11] or after cut-and-paste into the inlier scenes [65, 25]. Meta-OoD [11] 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 [46] for fine-tuning.

### 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 [14], which is an improved version of the initial MaskFormer [15]. We give only a brief overview to make the discussion self-contained; please refer to Cheng et al. [14] 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 de-

coder 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  $\sigma$  to the result:

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

$\mathbf{M}(\mathbf{x}) \in \mathbb{R}^{N \times H \times W}$  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:

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

### 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 [17] 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:

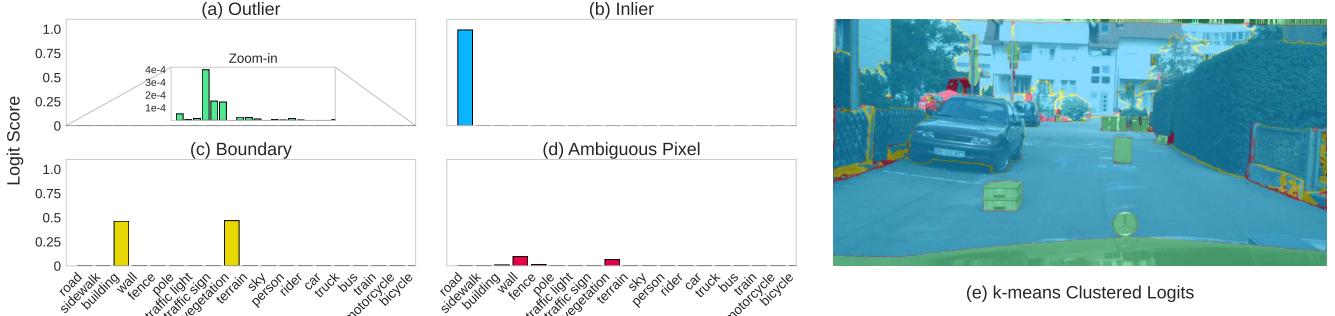
$$p(\mathbf{y} = k | \mathbf{x}) = \sigma(\mathbf{L}_k(\mathbf{x})) \quad (3)$$

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  $\mathcal{K} = \{1, \dots, K\}$  and  $\sigma$  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 probability map of an input  $\mathbf{x}$  being an outlier as follows:

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

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

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



**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).

We choose  $\sigma$  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 [46] are randomly cut and pasted on Cityscapes images [17]. 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:

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

where  $\Omega_{out}$  is the set of outlier pixels. We experimentally set the hyper-parameter  $\alpha$  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, 49], 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 highlights 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 [17], 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 [10] 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

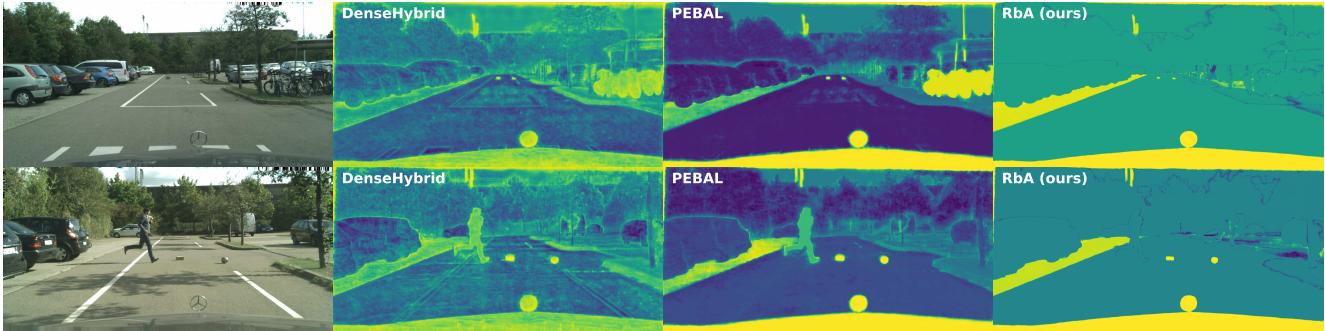


Figure 5: **Visual comparison to the state-of-the-art.** We show visualizations of outlier score maps predicted by our method, RbA compared to the ones predicted by state-of-the-art methods PEBAL [65] and DenseHybrid [25] trained using the same architecture as the RbA for a fair comparison. The other 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	Extra	Anomaly Track						Obstacle Track			
	Data	Net.	AP ↑	FPR ↓	sIoU gt ↑	PPV ↑	mean F1 ↑	AP ↑	FPR ↓	sIoU gt ↑	PPV ↑	mean F1 ↑
Emb. Density[5]	✗	✗	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 Inpaint.[47]	✗	✓	-	-	-	-	-	54.1	47.1	<b>57.6</b>	39.5	<u>36.0</u>
Image Resyn.[48]	✗	✓	52.3	<u>25.9</u>	39.7	11.0	12.5	37.7	4.7	16.6	20.5	8.4
ObsNet[1]	✗	✓	<u>75.4</u>	26.7	<u>44.2</u>	<b>52.6</b>	<b>45.1</b>	-	-	-	-	-
NFlowJSI[24]	✗	✓	56.9	34.7	36.9	18.0	14.9	<u>85.6</u>	<b>0.4</b>	45.5	<u>49.5</u>	<b>50.4</b>
RbA (Ours)	✗	✗	<b>86.1</b>	<b>15.9</b>	<b>56.3</b>	<u>41.4</u>	<u>42.0</u>	<b>87.8</b>	<u>3.3</u>	<u>47.4</u>	<b>56.2</b>	<b>50.4</b>
Max. Entropy[43]	✓	✗	<u>85.5</u>	15.0	49.2	<u>39.5</u>	28.7	85.1	0.8	<u>47.9</u>	<b>62.6</b>	48.5
DenseHybrid[25]	✓	✗	78.0	<b>9.8</b>	<u>54.2</u>	24.1	<u>31.1</u>	<u>87.1</u>	<b>0.2</b>	45.7	50.1	<u>50.7</u>
PEBAL[65]	✓	✗	49.1	40.8	38.9	27.2	14.5	5.0	12.7	29.9	7.6	5.5
SynBoost[4]	✓	✓	56.4	61.9	34.7	17.8	10.0	71.3	3.2	44.3	41.8	37.6
RbA (Ours)	✓	✗	<b>90.9</b>	<u>11.6</u>	<b>55.7</b>	<b>52.1</b>	<b>46.8</b>	<b>91.8</b>	<u>0.5</u>	<b>58.4</b>	<u>58.8</u>	<b>60.9</b>

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.

weather conditions. Both datasets are characterized by a high domain shift compared to Cityscapes, making them particularly challenging. Road Anomaly [48] 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 [5], 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.

## 4.2. Experimental Setup

**Implementation Details:** We follow the setup of [14] for closed-set training on Cityscapes. We use the Swin-B [50] 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 [14] 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

Method	OoD	Extra	Road Anomaly			FS LaF		
	Data	Net.	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 [41]	✗	✗	76.73	22.85	59.20	92.51	27.83	30.17
SML [39]	✗	✗	81.96	25.82	49.74	<u>96.88</u>	36.55	14.53
GMMSeg (SF) [44]	✗	✗	<u>89.37</u>	<u>57.65</u>	<u>44.34</u>	<b>97.83</b>	<u>50.03</u>	<u>12.55</u>
SynthCP [68]	✗	✓	76.08	24.86	64.69	88.34	6.54	45.95
RbA (Ours)	✗	✗	<b>95.60</b>	<b>78.45</b>	<b>11.83</b>	96.43	<b>60.96</b>	<b>10.63</b>
Maximized Entropy [11]	✓	✗	-	-	-	93.06	41.31	37.69
PEBAL [65]	✓	✗	<u>88.85</u>	<u>44.41</u>	<u>37.98</u>	<u>98.52</u>	<u>64.43</u>	<u>6.56</u>
SynBoost (WRN38) [4]	✓	✓	81.91	38.21	64.75	96.21	60.58	31.02
RbA (Ours)	✓	✗	<b>97.99</b>	<b>85.42</b>	<b>6.92</b>	<b>98.62</b>	<b>70.81</b>	<b>6.30</b>

Table 2: **Results on Road Anomaly and Fishscapes 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.

AP and FPR@95 for per-pixel metrics as well as component-level metrics that are designed to measure the statistics of detected objects [10]. 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 [10] 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 Dense-Hybrid [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 & Fishscapes LaF

Table 2 shows the results on the Road Anomaly [48] and the Fishscapes Lost and Found (LaF) validation set [5]. 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).

### 4.4. 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.

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

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.

**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.

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)	<b>85.42</b>	<b>6.92</b>	<b>70.81</b>	<b>6.30</b>

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.

**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.

**Different Backbones:** We use the same Mask2Former model with Swin-B backbone [50] 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]

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 / <b>6.2</b>
MixT [70]	<u>65.7 / 78.1</u>	<u>24.6 / 12.4</u>	40.3 / 51.3	<u>23.0 / 17.1</u>
Swin-B [50]	<b>78.5 / 85.4</b>	<b>11.8 / 6.9</b>	<b>61.0 / 70.8</b>	<b>10.6 / 6.3</b>

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.

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.

## 5. 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 Barin for their valuable feedback. This project is funded by the Royal Society Newton Fund Advanced Fellowship (NAF\R\2202237). Nazir Nayal and Misra Yavuz are supported by the KUIS AI fellowship. We also thank 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] 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](#)
- [5] Hermann Blum, Paul-Edouard Sarlin, Juan I. Nieto, Roland Y. Siegwart, and César Cadena. The fishscapes benchmark: Measuring blind spots in semantic segmentation. *IJCV*, 129:3119–3135, 2021. [6](#), [7](#), [14](#)
- [6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giacarlo 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](#)
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. [2](#)
- [8] João Carreira, Rui Caseiro, Jorge P. Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. [2](#)
- [9] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *ICCV*, 2021. [3](#)
- [10] 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](#), [5](#), [7](#)
- [11] 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](#)
- [12] 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](#)
- [13] 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](#)
- [14] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. [2](#), [3](#), [6](#), [12](#), [13](#), [14](#)
- [15] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. [2](#), [3](#)
- [16] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *NeurIPS*, 2019. [3](#)
- [17] 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](#), [12](#), [13](#)
- [18] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. [2](#)
- [19] Terrance Devries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv.org*, 1802.04865, 2018. [3](#)
- [20] Xianzhi Du, Barret Zoph, Wei-Chih Hung, and Tsung-Yi Lin. Simple training strategies and model scaling for object detection. *arXiv.org*, 2107.00057, 2021. [12](#)
- [21] J. Fu, J. Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. [2](#)
- [22] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. [1](#), [3](#)
- [23] Golnaz Ghiasi, Yin Cui, A. Srinivas, Rui Qian, Tsung-Yi Lin, Ekin Dogus Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. [12](#)
- [24] Matej Grcic, Petra Bevandić, and Siniša Šegvić. 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](#), [16](#), [17](#), [18](#), [19](#)
- [26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. [1](#), [3](#)
- [27] David Heldmann, 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, 7, 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] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *NeurIPS*, 2018. 1, 3
- [39] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungba 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
- [40] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 1, 3
- [41] 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
- [42] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. In *BMVC*, 2018. 2
- [43] 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
- [44] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. GMMSeg: Gaussian mixture based generative semantic segmentation models. *arXiv.org*, 2210.02025, 2022. 3, 7
- [45] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 3
- [46] 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, 12
- [47] Krzysztof Lis, Sina Honari, P. Fua, and Mathieu Salzmann. Detecting road obstacles by erasing them. *arXiv.org*, 2012.13633, 2020. 3, 6
- [48] Krzysztof Lis, Krishna Kanth Nakka, Pascal V. Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *ICCV*, 2019. 2, 3, 6, 7, 14
- [49] 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
- [50] 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. 6, 8, 12
- [51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 12, 13
- [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, Yuan-hong Chen, and G. Carneiro. Pixel-wise energy-biased ab-stention learning for anomaly segmentation on complex urban driving scenes. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7, 8, 12, 16, 17, 18, 19
- [66] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection 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, Wenhui 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, 15
- [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: Pointwise 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
- [84] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 12

# Appendix

## Overview

This supplementary document contains the implementation details that are necessary to reproduce our approach (Section A), additional ablation results to justify hyper-parameter choices (Section B), additional details about analysis and ablation experiments reported in the main paper (Section C), additional qualitative results to showcase our method compared to state-of-the-art (Section D), and some challenging cases that cause failure (Section E).

## A. Implementation Details

### A.1. Architecture

Fig. 6 illustrates the full Mask2Former architecture along with our unknown inference procedure. The main components of the model are the backbone, the pixel decoder, and the transformer decoder. We explain the details of each next.

**Backbone:** We use the Swin-B variant as the backbone [50]. It can take an RGB image with any resolution higher than  $32 \times 32$  as input and outputs feature maps at several resolutions to the pixel decoder. Specifically, the output feature maps are downsampled with strides 4 ( $\mathbf{x}_4$ ), 8 ( $\mathbf{x}_8$ ), 16 ( $\mathbf{x}_{16}$ ), and 32 ( $\mathbf{x}_{32}$ ) with respect to the input image.

**Pixel Decoder:** Following [14], the pixel decoder mainly consists of 6 layers of deformable attention (MSDeformAttn) [84]. The multi-scale feature maps with strides  $\mathbf{x}_8$ ,  $\mathbf{x}_{16}$ , and  $\mathbf{x}_{32}$  are processed with MSDeformAttn layers to produce  $\mathbf{f}_1$ ,  $\mathbf{f}_2$ , and  $\mathbf{f}_3$ , respectively. In [14], the three processed feature maps are passed to 9 transformer decoder layers in a round-robin fashion. However, we found that using a single layer in the transformer decoder works better for unknown detection. Therefore, we only pass the last layer  $\mathbf{f}_3$  to the transformer decoder. The feature map  $\mathbf{x}_4$  is processed with a  $1 \times 1$  filter-size convolutional layer and then added to the processed feature map  $\mathbf{f}_1$  after bilinear upsampling. Finally, the output is passed to a  $3 \times 3$  convolutional layer to produce per-pixel features  $\mathbf{F} \in \mathbb{R}^{C_p \times H \times W}$ , where  $C_p = 256$  is the embedding dimension. The computation can be summarized as follows:

$$\mathbf{F} = \text{Conv}_{3 \times 3}(\text{Conv}_{1 \times 1}(\mathbf{x}_4) + \text{Upsample}(\mathbf{f}_1)) \quad (7)$$

**Transformer Decoder:** Learnable object queries  $\mathbf{Q} \in \mathbb{R}^{N \times C_q}$  are fed to the transformer decoder layer to be processed with the feature maps from the pixel decoder, where  $N = 100$  is the number of object queries and  $C_q = 256$  is the embedding dimension. A single transformer decoder layer consists of a cross-attention layer followed by self-attention and feed-forward network (FFN), each of which is followed by a LayerNorm. The cross-attention operation

is performed with mask-attention, where each object query only attends to regions it predicted in the previous layer. Since we use only a single layer, each object query attends to the region it predicts directly from the input feature map before being processed in the transformer decoder. Learnable positional embeddings are added to the object queries. The transformer decoder outputs a refined set of object queries  $\mathbf{Q}_r$  that predict the regions and classify them.

**Region Class Prediction:** The refined object queries  $\mathbf{Q}_r$  are fed into a single linear layer followed by a softmax to produce the class probability of each region  $\mathbf{P} \in \mathbb{R}^{N \times K}$  where  $K$  is the number of classes.

**Membership Maps Prediction:** The refined object queries  $\mathbf{Q}_r$  are also fed into a 3-layer MLP, so that  $\mathbf{Q}_r$ 's dimensionality matches that of  $\mathbf{F}$ . Then,  $\mathbf{Q}_r$  and  $\mathbf{F}$  are multiplied before being fed into a sigmoid activation to produce the per-pixel membership maps  $\mathbf{M} \in \mathbb{R}^{N \times H \times W}$ .

### A.2. Closed-Set Training

**Loss Functions:** Before applying any loss function, bipartite matching is used to match object queries to ground truth binary masks, where each mask contains all the pixels of a certain class. The matching cost is computed as a weighted sum of the individual losses. The classification is performed with the cross-entropy loss. A weighted combination of dice loss and binary cross-entropy is used to predict regions.

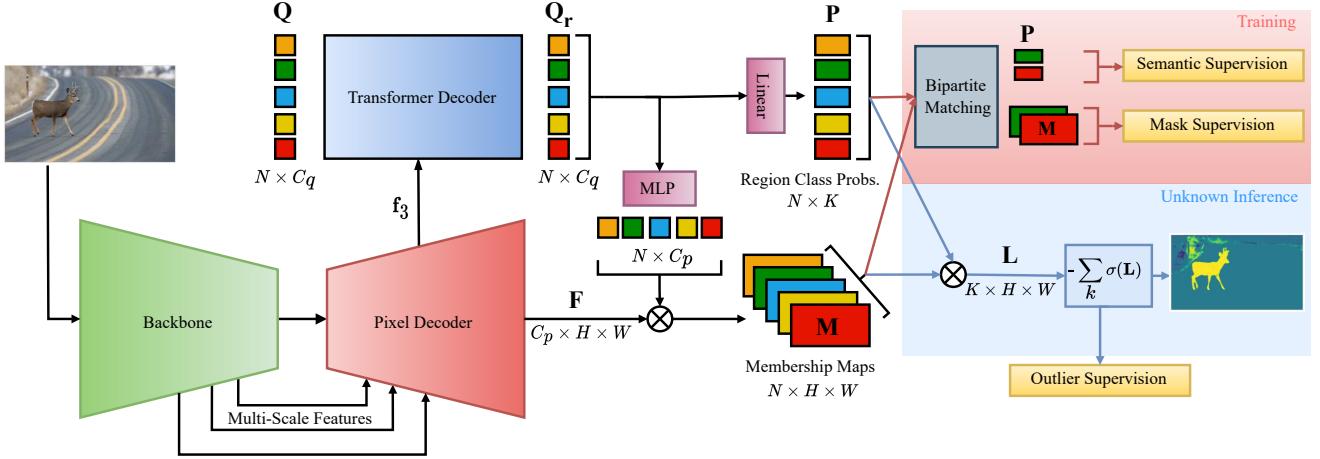
**Hyper-Parameters:** Following [14], the model is trained for 90K iterations using a batch size of 16. AdamW [51] optimizer is used with 0.05 for weight decay and an initial learning rate of  $10^{-4}$ , which is reduced using a polynomial scheduler. The learning rate for the backbone is multiplied by 0.1.

**Data Augmentation:** We use the same augmentations as in [14]. First, the short side of the input image is resized by a scale uniformly chosen between  $[0.5 - 2]$ . Then a random crop of size  $512 \times 1024$  is applied. After that, large-scale jittering augmentation [20, 23] is applied with a random horizontal flip.

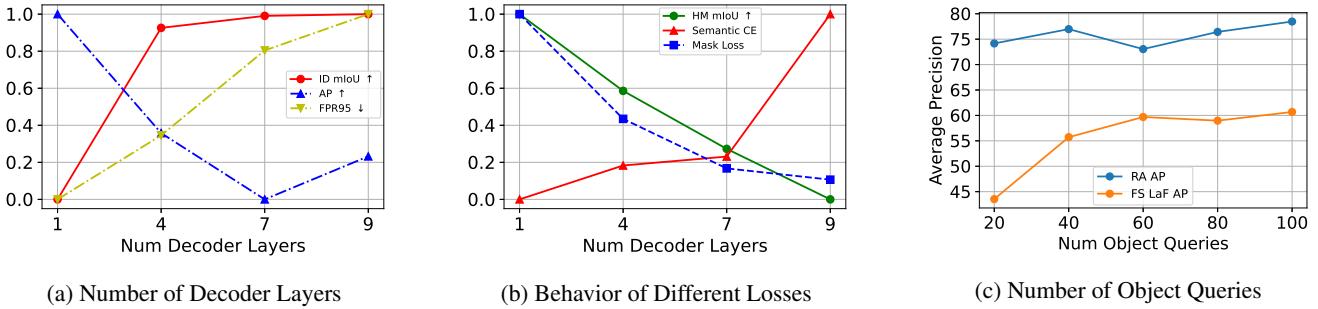
### A.3. Outlier Supervision

**Data Sampling:** We use a slightly modified version of AnomalyMix proposed in [65] for outlier supervision. After eliminating the samples that contain Cityscapes classes [17], around 40K images remain for outlier supervision on the COCO [46] dataset. For a single fine-tuning experiment, we randomly sample 300 images and fix them throughout the entire fine-tuning phase.

**Fine-tuned Components:** For all the fine-tuning experiments, we only fine-tune the 3-layer MLP and linear layers shown in pink in Fig. 6. Their weights together constitute approximately 0.21% of the entire model parameters.



**Figure 6: Detailed architecture.** This figure provides a more detailed view of the Mask2Former [14] architecture, including our modifications and unknown inference computation. We use a single transformer decoder layer as opposed to the original implementation that uses 9 layers. Therefore, only a single scale feature  $f_3$  from the last layer is passed from the pixel decoder to the transformer decoder. For outlier supervision, all modules are frozen except for the MLP and the linear layers shown in pink.



**Figure 7: Ablation study on architectural choices.** With more decoder layers, in-distribution performance on Cityscapes improves but the outlier performance drops in terms of AP and FPR@95 on Road Anomaly a. Good performance is mainly due to better mask prediction at the cost of a higher semantics loss b. The drop in mIoU with hard masking (**HM mIoU**) is another indicator of semantic information loss in the object queries with more decoder layers. The performance improves as the number of object queries increases c.

**Hyper-Parameters:** After the model is trained on the closed-set setting, we fine-tune it for 2000 iterations on Cityscapes [17] using the setting of the closed-set training; AdamW [51] optimizer with 0.05 weight decay and  $10^{-4}$  initial learning with polynomial scheduling. For every Cityscapes image used in fine-tuning, an object from the 300 COCO samples is uniformly chosen and pasted on the Cityscapes image with probability  $p_{out} = 0.1$ , which is independent for each image. The RbA score for outlier pixels is optimized with the squared hinge loss  $\mathcal{L}_{RbA}$  using  $\alpha = 5$ .

## B. Additional Ablation Study

**Number of Transformer Decoder Layers:** As shown in [14], more transformer decoder layers improve the inlier

performance, i.e. mIoU on Cityscapes. However, we found that using fewer decoder layers results in better performance in terms of outliers. Fig. 7a highlights the decrease in performance in terms of the AP and FPR@95 on the Road Anomaly dataset as the number of decoder layers increases. We investigate this behavior by isolating the sources of error with respect to the number of decoder layers. Fig. 7b shows semantic and mask losses of the Mask2Former [14] averaged over the validation samples on Cityscapes. With more decoder layers, we observe that the semantic cross-entropy loss increases while the mask-related BCE and dice losses decrease. This shows that the increase in inlier mIoU with more decoder layers can be attributed to increased performance in detecting masks at the cost of a higher semantic

Num Iter	AP ↑	FPR@95 ↓	$p_{out}$	AP ↑	FPR@95 ↓	mIoU ↑	$\alpha$	AP ↑	FPR@95 ↓
1000	$83.20 \pm 0.11$	$8.69 \pm 0.04$	0.05	$84.34 \pm 0.15$	$9.12 \pm 0.10$	<b>82.28 ± 0.02</b>	-0.1	$84.01 \pm 0.18$	$9.25 \pm 0.06$
2000	<b>85.49 ± 0.08</b>	<b>7.25 ± 0.04</b>	0.1	<b>85.48 ± 0.12</b>	<b>7.24 ± 0.06</b>	<b>82.15 ± 0.09</b>	-0.01	$84.41 \pm 0.14$	$9.06 \pm 0.11$
3000	<b>85.72 ± 0.12</b>	$7.82 \pm 0.11$	0.2	<b>85.74 ± 0.11</b>	$7.90 \pm 0.11$	$81.54 \pm 0.02$	0.0	$84.30 \pm 0.06$	$9.10 \pm 0.07$
4000	$84.89 \pm 0.07$	<b>7.45 ± 0.05</b>	0.4	$84.97 \pm 0.11$	$9.19 \pm 0.05$	$81.27 \pm 0.06$	2	<b>85.30 ± 0.07</b>	$7.80 \pm 0.06$
5000	$84.66 \pm 0.06$	$8.14 \pm 0.03$					5	$85.24 \pm 0.08$	<b>6.95 ± 0.08</b>
							10	<b>85.61 ± 0.10</b>	<b>7.26 ± 0.07</b>

(a) Number of Iterations

(b) Outlier Selection Probability

(c) Outlier Threshold

**Table 6: Ablation study on outlier supervision.** Finetuning with RbA loss for 2000-3000 iterations achieves the best performance and the performance deteriorates after 3000 iterations as shown in **a**. A higher probability of exposure to outlier data results in a consistent decline in the closed-set performance. A probability of 0.1 achieves the best balance between outlier and inlier performance **b**. Finally, we ablate the RbA loss parameter  $\alpha$  in **c** and find that the best results are achieved with  $\alpha > 0$ .

Module	Params (%)	mIoU	Road Anomaly		FS LaF	
			AP ↑	FPR ↓	AP ↑	FPR ↓
Full Model	100	80.81	76.00	<b>9.50</b>	<b>73.88</b>	<b>6.02</b>
Transformer Dec.	1.93	80.24	<b>85.08</b>	10.18	<b>72.6</b>	<b>5.51</b>
Pixel Dec.	4.66	<b>81.59</b>	75.83	10.51	69.8	6.77
MLP+Linear	<b>0.21</b>	<b>82.20</b>	<b>85.42</b>	<b>6.92</b>	70.81	6.30

**Table 7: Ablation study on fine-tuning different modules.** We show the effect of fine-tuning different components of the model on the Road Anomaly and Fishscapes LaF validation sets. Fine-tuning MLP+Linear maintains the best performance in unknown detection without sacrificing the closed-set performance.

error. By using fewer decoder layers, we regulate the semantic confusion, which helps to better align the logit scores, resulting in better outlier performance.

Fig. 7b also shows the mIoU evaluated by applying hard masking on the specialized object queries. Specialized object queries perform worse with more decoder layers. The information loss in the object queries as well as the increase in the semantic loss show the importance of semantics for outlier segmentation compared to precise masks.

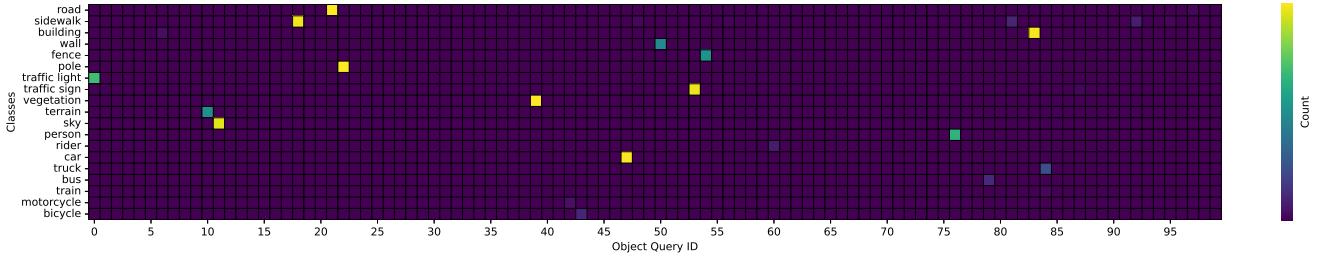
**Number of Object Queries:** The original Mask2Former [14] uses 100 object queries. We train different models by varying the number of object queries to observe its effect on detecting outliers. We focus on the AP values on both Road Anomaly [48] with large objects and on Fishscapes LaF [5] with small objects. Fig. 7c shows that more object queries result in better AP on both datasets. Even if some object queries specialize in predicting a particular class, the other object queries still play a role, especially for rare classes, as shown in Fig.3 in the main paper.

**Outlier Data Exposure:** We perform an experiment to show the effect of the number of iterations required for finetuning with our RbA loss in Table 6a. We report the AP and FPR metrics on the Road Anomaly dataset, averaged over 5

different runs to eliminate the effect of randomness. We can see that the best results can be achieved after around 2000 and 3000 iterations and then begin to degrade. We also evaluate the effect of the amount of outlier data exposed during training, which is controlled by the parameter  $p_{out}$ . We experiment with different values for  $p_{out}$  and report the outlier performance on Road Anomaly and closed-set performance on Cityscapes averaged over 5 different runs in Table 6b. We can see that more exposure to outlier data negatively affects the closed-set performance. Consequently, even the outlier segmentation performance starts to degrade for  $p_{out} > 0.2$ . We choose the  $p_{out} = 0.1$  because it strikes a reasonable balance between outlier and closed-set performance.

**RbA Loss Parameter:** In Table 6c, we report the performance of our loss function  $\mathcal{L}_{\text{RbA}}$  using different values of  $\alpha$ , averaged over 5 different runs. We find that positive values of  $\alpha$  work similarly well and set  $\alpha$  to 5 in our experiments.

**Fine-tuned Component:** In Table 7, we analyze the effect of fine-tuning different parts of the model on validation sets of Road Anomaly [48] and Fishscapes Lost and Found [5] using RbA loss. The alternative components we experimented with are the full model, only the transformer decoder (blue + pink in Fig. 6), only the pixel decoder (red in Fig. 6), and MLP+Linear layers (pink in Fig. 6). On the Road Anomaly dataset, fine-tuning MLP+Linear achieves the best performance in terms of AP and FPR. On Fishscapes LaF, the best AP is achieved by fine-tuning the entire model, while the best FPR is obtained by fine-tuning only the transformer decoder. Both options cost a decrease in performance on Road Anomaly and negatively affect the closed-set performance. Fine-tuning MLP+Linear achieves the best balance between outlier detection and closed-set performance and is the least costly option in terms of the number of parameters finetuned.



**Figure 8: Specialization of Object Queries.** Certain object queries specialize in predicting a specific class. For each query, we show how many times it predicts a region to belong to a class with high confidence. The sparsity in the plot clearly shows the specialization of queries.

## C. Details of Experiments

In this section, we provide further illustrations and detailed settings of our analysis and ablation experiments in the main paper. We first perform an experiment to verify the specialization of object queries. We then provide a detailed formulation of our ablations on the loss functions and other methods using Mask2Former including the hyper-parameters that we use to obtain the results presented in the main paper.

### C.1. Specialization of Object Queries

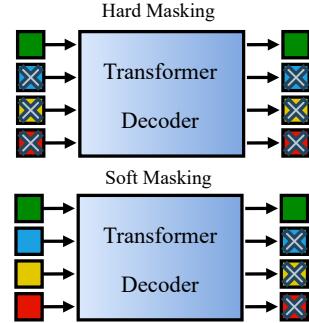
Our method is based on our finding that the object queries in mask classification enjoy a degree of independence from one another and that each object query in a subset specializes in segmenting a specific class from the closed set. Due to bipartite matching being applied between queries and ground truth class masks during training, this behavior can be anticipated. Here, we empirically verify it using a different dataset than the one used in training (BDD100K [73]). For each object query, we count how many times it predicts a certain class with high confidence. Fig. 8 shows the heatmap of counts for the Mask2Former model with 100 object queries. For each of the closed-set classes, there is a single object query dominantly predicting it.

In the main paper, we test the independence of the specialized queries by applying hard masking and soft masking and evaluating per class IoU on Cityscapes. Fig. 9 shows an illustration of hard and soft masking applied.

### C.2. Other Loss Functions

In the main paper, we perform an ablation study with different loss functions in comparison to squared hinge loss. In this section, we provide the formulation and the parameter setting of each loss function. In the following,  $\Omega_{out}$  denotes the set of outlier pixels, and  $\Omega_{in}$ , the set of inlier pixels on an image.

**Mean-Squared Error (MSE):** With MSE loss, we opti-



**Figure 9: Illustration of hard vs. soft masking of object queries.** In hard masking, when predicting class  $k$ , all but the object queries specialized to predict class  $k$  are masked before the transformer decoder so that the specialized query can only interact with the image features. In soft masking, the queries are allowed to interact within the transformer decoder but are dropped after, and only the specialized query is used to predict class  $k$ .

mize the RbA to be closer to  $\alpha = 5$  for outlier pixels:

$$\mathcal{L}_{MSE} = \sum_{\mathbf{x} \in \Omega_{out}} (\text{RbA}(\mathbf{x}) - \alpha)^2 \quad (8)$$

**L1:** Similar to MSE, we optimize the RbA with L1 loss by setting  $\alpha$  to 5:

$$\mathcal{L}_{L1} = \sum_{\mathbf{x} \in \Omega_{out}} |\text{RbA}(\mathbf{x}) - \alpha| \quad (9)$$

**Binary Cross Entropy (BCE):** We formulate the scoring of outliers as a per-pixel binary classification problem, where outliers correspond to the positive class. We use the RbA score as the logit score for the positive class. Assuming that  $y$  corresponds to the binary label (outlier vs. inlier) of a pixel

$\mathbf{x}$ , we optimize the BCE loss as follows:

$$\mathcal{L}_{\text{BCE}} = \sum_{\mathbf{x} \in \Omega_{out}} y \cdot \log(\text{RbA}(\mathbf{x})) + (1-y) \cdot \log(1 - \text{RbA}(\mathbf{x})) \quad (10)$$

**KL Divergence:** We use the KL Divergence to minimize the distance between the predicted class distribution to a fixed distribution. For inlier pixels, we minimize the distance to the Dirac delta function where the correct class has a probability of 1.0. For outlier pixels, we minimize the distance to a uniform distribution where the entropy is maximum. Even though this loss function does not optimize our proposed score function RbA, it helps us estimate the contribution of RbA by pushing the predicted class probabilities toward the ideal distributions for outliers and inliers. Let  $\mathbf{L}_p(\mathbf{x})$  be the class probability distribution of a pixel  $\mathbf{x}$  with ground truth label  $y$ . Let  $\mathbf{P}_{in}(y)$  be a fixed probability distribution where class  $y$  has probability 1.0. Let  $\mathcal{U}$  be a fixed uniform probability distribution. Formally, the loss function is defined as follows:

$$\begin{aligned} \mathcal{L}_{in} &= \sum_{\mathbf{x} \in \Omega_{in}} D_{KL}(\mathbf{L}_p(\mathbf{x}) \parallel \mathbf{P}_{in}(y)) \\ \mathcal{L}_{out} &= \sum_{\mathbf{x} \in \Omega_{out}} D_{KL}(\mathbf{L}_p(\mathbf{x}) \parallel \mathcal{U}) \\ \mathcal{L}_{KL} &= \frac{1}{2} (\mathcal{L}_{in} + \mathcal{L}_{out}) \end{aligned} \quad (11)$$

### C.3. Other Methods with Mask2Former

To disentangle the contribution of our scoring function RbA from the architecture, in the main paper, we present the results of the state-of-the-art methods PEBAL [65] and DenseHybrid [25] using the Mask2Former architecture. Here, we provide the details of this ablation experiment for each method. For a fair comparison, we train both methods by fine-tuning the same components as the RbA, that is the MLP+Linear layers as shown in Fig. 6, and also using the same outlier data supervision method as described in the main paper.

**PEBAL:** The optimized objective as per the official implementation [65] consists of three components. First, there is the pixel-wise anomaly abstention loss (PAL) with the abstention term and penalty defined as follows:

$$\mathcal{L}_{PAL} = - \sum_{\mathbf{x} \in \Omega} \log \left( \mathbf{L}_y(\mathbf{x}) + \frac{\mathbf{L}_{K+1}(\mathbf{x})}{a(\mathbf{x})} \right)$$

where  $\Omega$  is the set of all pixels on a given image,  $y \in 1, \dots, K+1$  is the ground truth class of pixel  $\mathbf{x} \in \Omega$ , and  $K+1$  is the class for the outliers. In Mask2Former, class  $K+1$  is assumed to be the no object class, therefore we avoid dropping it from the region class probabilities term

$\mathbf{P}$  (see Fig. 6) while fine-tuning with the PEBAL objective. The abstention penalty term  $a(\mathbf{x})$  is defined as follows:

$$\begin{aligned} a(\mathbf{x}) &= (-E(\mathbf{x}))^2 \\ E(\mathbf{x}) &= -\log \sum_{k=1}^K \exp(\mathbf{L}_k(\mathbf{x})) \end{aligned} \quad (12)$$

where  $E(\mathbf{x})$  is the free energy function. When the penalty term is high, the prediction is discouraged from abstaining and vice versa.

The second component of the loss optimizes the energy terms such that it is maximized for outlier pixels and minimized for inlier pixels as follows:

$$\begin{aligned} \mathcal{L}_{energy} &= \sum_{\mathbf{x} \in \Omega_{in}} \max(0, E(\mathbf{x}) - m_{in})^2 + \\ &\quad \sum_{\mathbf{x} \in \Omega_{out}} \max(0, m_{out} - E(\mathbf{x}))^2 \end{aligned} \quad (13)$$

where  $m_{in}$  and  $m_{out}$  are hyper-parameters to be set. In our experiments, we experimentally use  $m_{out} = -2.5$  and  $m_{in} = -3.5$ .

The last component is a regularization term for the smoothness and sparsity of the predicted energy map:

$$\mathcal{L}_{reg} = \sum_{\mathbf{x} \in \Omega} \beta_1 |E(\mathbf{x}) - E(\mathcal{N}(\mathbf{x}))| + \beta_2 |E(\mathbf{x})| \quad (14)$$

where  $\mathcal{N}(\mathbf{x})$  is the set of vertical and horizontal neighboring pixels of  $\mathbf{x}$ , and  $\beta_1$  and  $\beta_2$  are hyper-parameters. We use  $\beta_1 = 3 \times 10^{-7}$  and  $\beta_2 = 5 \times 10^{-5}$ .

The full objective is defined as the weighted sum of the three loss functions:

$$\mathcal{L}_{PEBAL} = \mathcal{L}_{PAL} + \beta \mathcal{L}_{energy} + \mathcal{L}_{reg} \quad (15)$$

where we set  $\beta = 0.1$ . Starting from the same checkpoint that we use fine-tuning RbA, we optimize the model with  $\mathcal{L}_{PEBAL}$  for 5K iterations. We set other hyper-parameters to be the same as the ones that we use for RbA.

**DenseHybrid:** Following the official implementation of DenseHybrid [25], we added an additional outlier prediction head  $\mathbf{D}(\mathbf{x}) \in \mathbb{R}^{2 \times H \times W}$  to the Mask2Former model which is defined as follows:

$$\mathbf{D}(\mathbf{x}) = \text{Conv}_{3 \times 3}(\text{ReLU}(\text{BatchNorm}(\mathbf{x}))) \quad (16)$$

The outlier prediction head takes the output feature map of the highest resolution from the pixel decoder and predicts a binary output for every pixel denoting the probability of being an outlier. The objective for DenseHybrid is defined as follows:

$$\mathcal{L}_{DH} = \text{CE}(\mathbf{L}(\mathbf{x}_{in}), \mathbf{Y}_{in}) + \beta_1 \text{CE}(\mathbf{D}(\mathbf{x}), \mathbf{Y}_{out}) + \beta_2 \mathcal{L}_o \quad (17)$$

where CE is short for the cross-entropy loss,  $\mathbf{x}_{in} \in \Omega_{in}$  denotes the set of inlier pixels,  $\mathbf{Y}_{in}$  denotes the ground truth for the closed-set and  $\mathbf{Y}_{out}$  denotes the binary map where the outlier pixels are set to one. We experimentally set the hyper-parameters  $\beta_1 = 0.3$  and  $\beta_2 = 0.03$ .  $\mathcal{L}_o$  is defined as follows:

$$\mathcal{L}_o = \frac{1}{|\Omega_{out}|} \sum_{\mathbf{x} \in \Omega_{out}} \log \sum_{k=1}^K \exp(\mathbf{L}_k(\mathbf{x})) + \text{sg}[\text{mean}(\mathbf{L}(\mathbf{x}))] \quad (18)$$

where sg is short for the stop gradient operation, and mean denotes the mean of all the elements of the input tensor.

## D. Additional Qualitative Results

In Fig. 10 and Fig. 11, we show additional qualitative results of RbA compared to the state-of-the-art methods PEBAL [65] and DenseHybrid [25]. For other methods, we show both the outputs of the models reported in their respective repositories and our implementations of the methods using Mask2Former. The proposed scoring function RbA reduces the false positives on the boundaries of inliers and ambiguous background regions compared to the baselines. These improvements can be observed more prominently on the obstacle track (Fig. 11) under adverse weather and lighting conditions. Moreover, compared to the baselines, RbA results in fewer false positives as a result of reducing confusion with inlier classes.

## E. Failure Cases

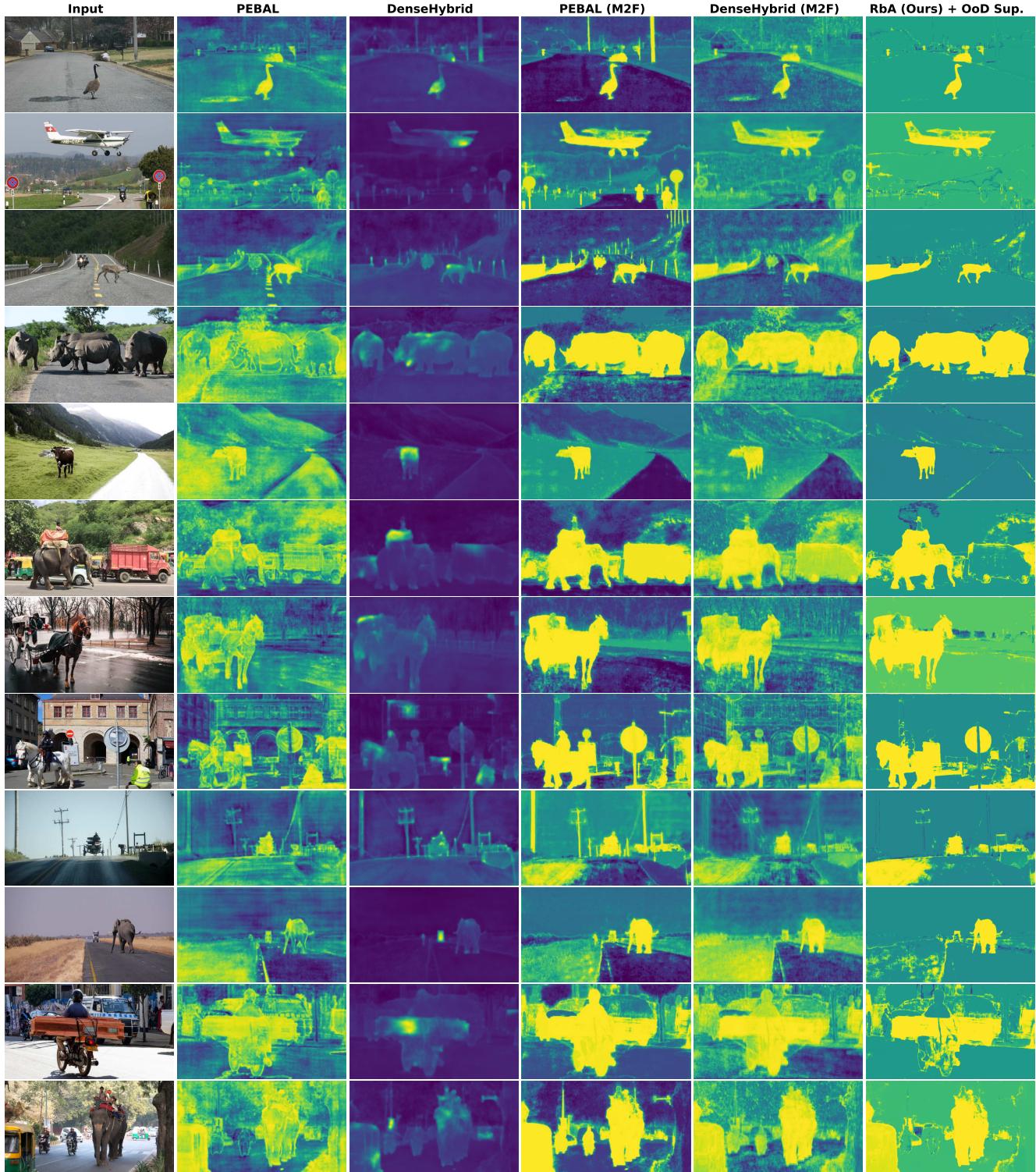
We analyze some failure cases of our method in this section. A common reason for the failure cases is the high similarity to the inlier classes.

**Tractors and Boats:** As shown in Fig. 12, RbA fails to detect tractors and boats as outliers due to their similarity to inlier vehicle instances. Although the objects are partially identified, the model cannot decisively predict the whole object regions as outliers. The existing methods either segment the boats and tractors at the cost of more false positives, as in the case of PEBAL, or also suffer from a lack of smoothness, as in the case of DenseHybrid.

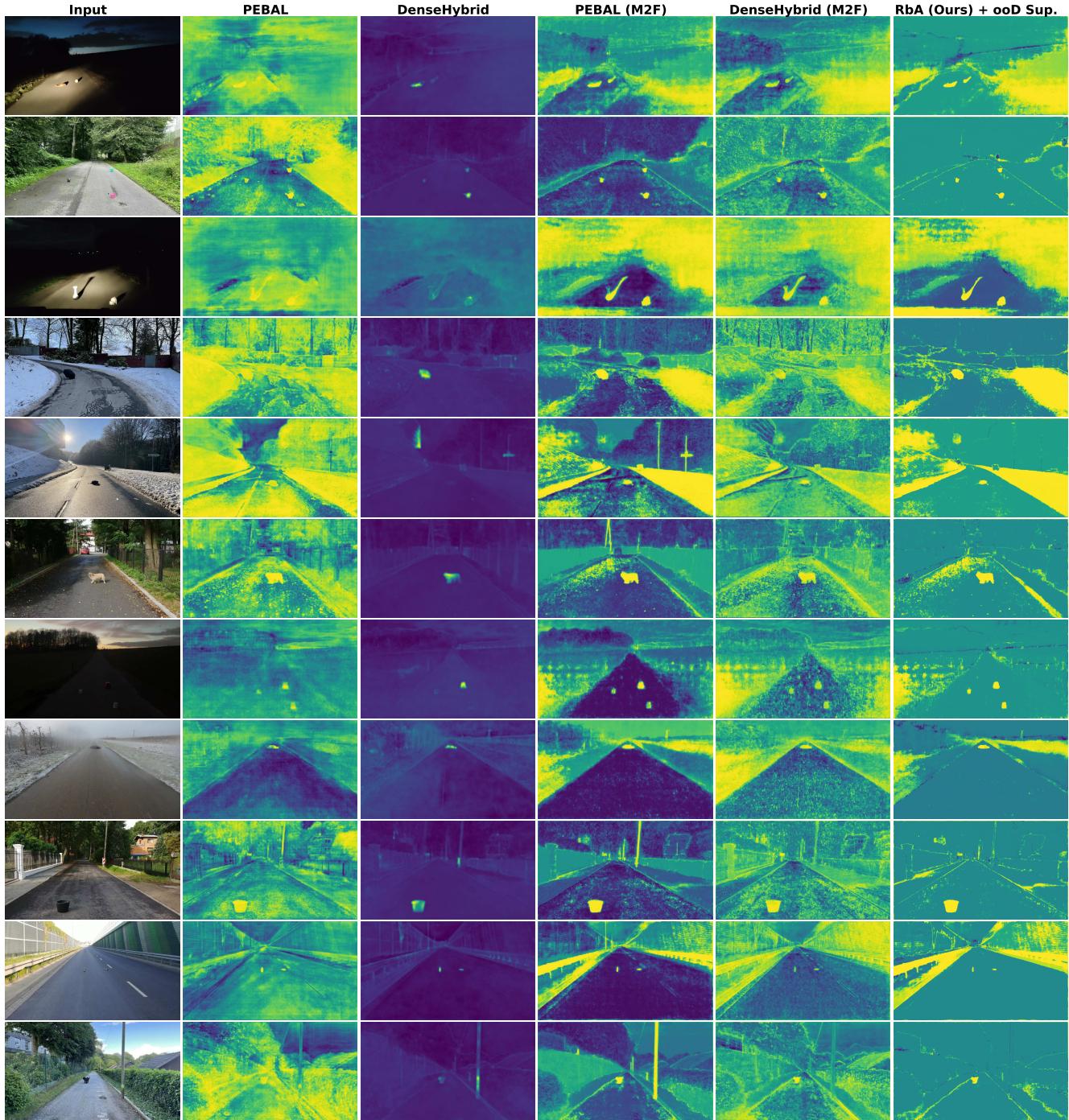
**Far away Animals:** As shown in Fig. 13, animals that are situated relatively far from the camera are confused as the inlier pedestrian class. This can be attributed to the dominance of pedestrian class in the training data as well as the similarity of legged animals to a pedestrian in appearance.

**Toy Cars:** Fig. 14 shows that the model fails to detect a toy car on the road and predicts it confidently as the inlier car class. While the class assignment can be considered semantically correct, it is still a hazard in a real driving scenario. Note that a small car can either be a toy car or a real car that is far away. Therefore, distinguishing real cars

from toy cars might require additional information such as depth or scale.



**Figure 10: Qualitative Results on SMIYC Anomaly Track.** On the anomaly track of the SMIYC benchmark, we compare RbA with outlier (OoD) supervision to the state-of-the-art methods PEBAL [65] and DenseHybrid [25] using the models that were shared in their respective repositories, as well as the versions we trained using Mask2Former (M2F). RbA better distinguishes outliers from inliers and produces more smooth outlier maps with fewer false positives.



**Figure 11: Qualitative Results on SMIYC Obstacle Track.** We compare RbA with outlier supervision to the state-of-the-art methods PEBAL [65] and DenseHybrid [25]. Under adverse weather and difficult lighting conditions, RbA can detect anomalies consistently better compared to DenseHybrid and reduce false positives more compared to PEBAL.

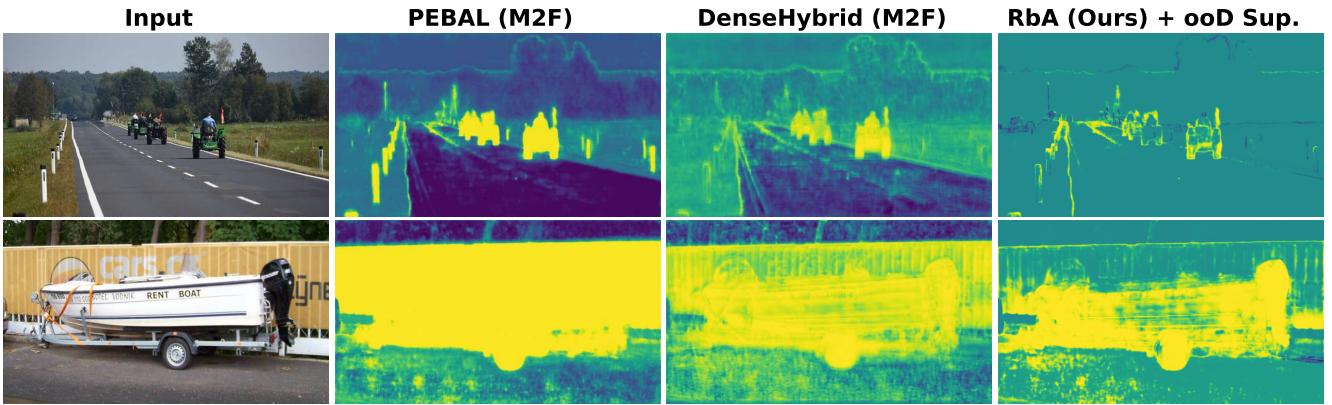


Figure 12: **Failure Cases: Tractors and Boats.** Due to their high similarity to inlier car and truck classes, unknown objects like tractors or boats are sometimes predicted as inliers.

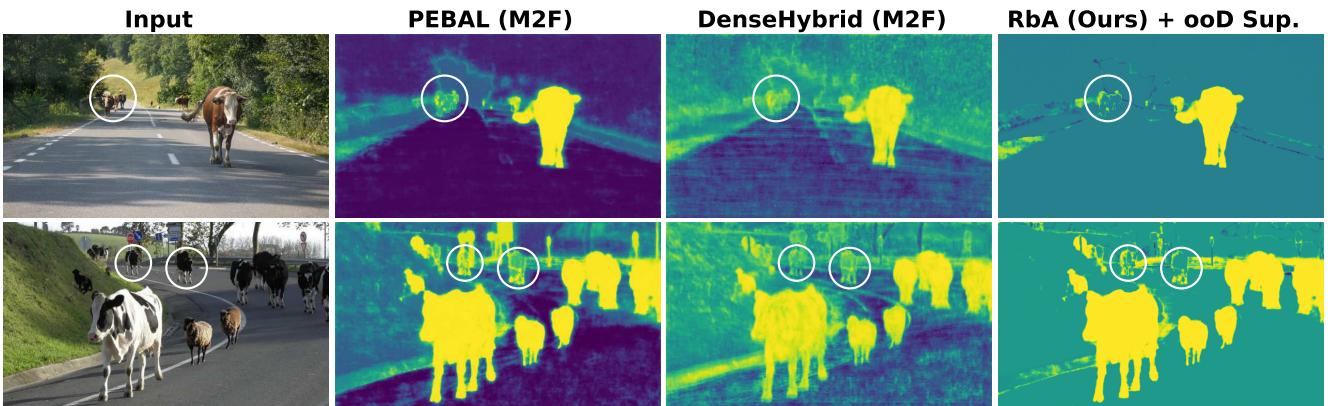


Figure 13: **Failure Cases: Animals Confused As Pedestrians.** As the pedestrian is one of the most frequent classes on Cityscapes, the model sometimes predicts animals that appear at a distance as pedestrians (highlighted in circles) on images from SMIYC Anomaly Track.

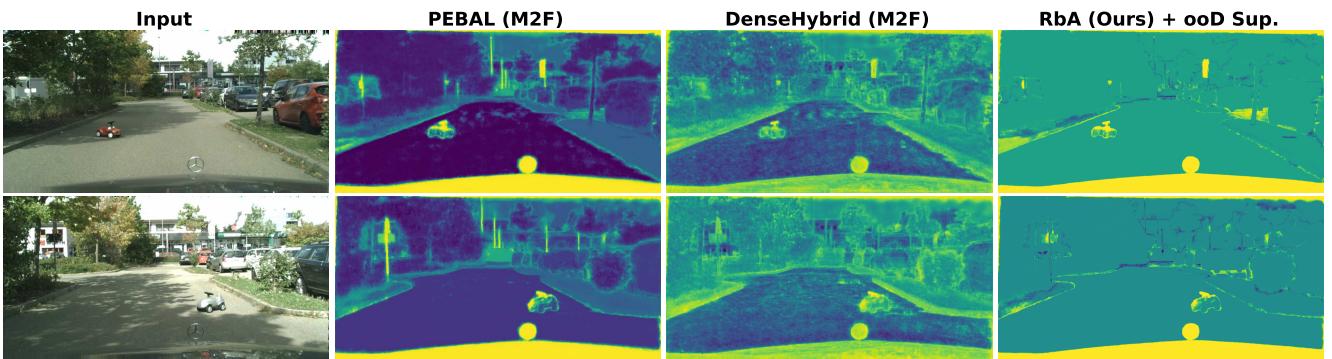


Figure 14: **Failure Cases: Toy Cars Predicted As Real Cars.** One confusing anomaly case for our model is small toy cars placed in front of the vehicle. Even though they can be semantically considered as cars, they are considered obstacles in a real driving scenario.