Out-of-Distribution Segmentation in Autonomous Driving: Problems and State of the Art

Youssef Shoeb
Continental AG
Technische Universität Berlin
youssef.shoeb@continental.com

Azarm Nowzad
Continental AG
azarm.nowzad@continental.com

Hanno Gottschalk
Technische Universität Berlin
gottschalk@math.tu-berlin.de

Abstract

In this paper, we review the state of the art in Out-of-Distribution (OoD) segmentation, with a focus on road obstacle detection in automated driving as a real-world application. We analyse the performance of existing methods on two widely used benchmarks, SegmentMeIfYouCan Obstacle Track and LostAndFound-NoKnown, highlighting their strengths, limitations, and real-world applicability. Additionally, we discuss key challenges and outline potential research directions to advance the field. Our goal is to provide researchers and practitioners with a comprehensive perspective on the current landscape of OoD segmentation and to foster further advancements toward safer and more reliable autonomous driving systems.

1. Introduction

The task of semantic segmentation involves assigning each pixel in an image to a predefined class. In recent years, neural networks have become the state-of-the-art approach for semantic segmentation, achieving remarkable accuracy across applications such as autonomous driving, medical imaging, and industrial inspection. However, these models rely on closed-set assumptions and are only trained to recognize categories seen during training. When deployed in open-world settings, they may encounter unknown objects and produce incorrect predictions with high confidence, lacking mechanisms to detect inputs outside the training distribution [55].

This limitation is particularly critical in autonomous driving, where vehicles must detect and respond to unexpected hazards, such as road debris or wildlife, even if such objects were not part of the training data. Addressing this challenge requires models that not only segment known objects but also identify and flag unknown ones, a task known as Out-of-Distribution (OoD) segmentation.

Similar to OoD segmentation, other problem formulations, such as *anomaly* and *open-world* segmentation, share similar motivations and methodologies but differ in their definitions. While there is no universally accepted definition of OoD in the literature, we adopt the following distinctions in this survey to guide our discussion. Specifically, we define anomalies as rare or unusual instances that occur within a known distribution (e.g., an atypical type of vehicle), while OoD objects stem from entirely unknown semantic categories not presented during training. While anomalies include instances that may or may not be correctly classified by a model, OoD objects represent a fundamental mismatch between the learned categories of the model and the objects encountered during deployment. Open-world segmentation, by contrast, aims to detect both known and unknown objects simultaneously. However, due to a lack of benchmarks with both inlier and unknown labels, the boundary between open-world and OoD segmentation is often blurred. Typically, OoD performance is evaluated on an external dataset, while inlier performance is measured on the original training set. In this survey, we do not explicitly differentiate between OoD and open-world segmentation but acknowledge that they are ideally treated separately: OoD segmentation methods are typically standalone models, while open-world segmentation methods are integrated into the segmentation model itself.

We present a comprehensive overview of recent advances in OoD segmentation, focusing on road obstacle detection as a safety-critical use case. We review state-of-the-art techniques, highlight key challenges, and propose future directions to advance the field and enhance autonomous system safety.

2. Problem Definition, Metrics & Dataset

Problem Definition: The goal of an OoD segmentation method is to assign an OoD score to each pixel in an image, distinguishing in-distribution from OoD pixels. This is achieved by defining a function

$$s : \mathcal{X} \to \mathbb{R}^{H \times W} \tag{1}$$

where $\mathcal{X} \subseteq [0,1]^{H \times W \times 3}$ represents the space of RGB images with normalized pixel values, and $s(x) \in \mathbb{R}^{H \times W}$ is the OoD score map where each pixel in the input image $x \in \mathcal{X}$ is assigned a real-valued score indicating its likelihood of being OoD.

Evaluation Protocol: Evaluating OoD segmentation is more complex than standard image-level OoD detection due to the dense, pixel-wise nature of the task. In particular, performance metrics can either be computed per image and then averaged across the dataset or aggregated across all pixels before computing a global metric. In this work, we adopt the latter strategy: all predictions and ground truth labels are collected across the full test set, and metrics are computed on the resulting global distribution of in-distribution and OoD pixels. This approach reduces the influence of image-level variance and reflects the overall effectiveness of the model across the entire dataset.

Metrics: One of the standard metrics for evaluating the separability of binary classification models is the area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC). However, in OoD segmentation, datasets are often highly imbalanced, with significantly fewer OoD pixels compared to in-distribution pixels. In such cases, the area under the precision-recall curve (AUPRC) is preferred as an alternative evaluation metric. Unlike AUC-ROC, which is based on true positive and false positive rates, AUPRC is computed using precision and recall for OoD pixels. This makes it a more informative measure of how well a model identifies OoD pixels without being disproportionately influenced by the large number of in-distribution pixels. The AUPRC approximates the integral $\int \text{precision}(\delta)\, d\text{recall}(\delta)$ at multiple thresholds. For an input image $x \in \mathcal{X}$, let $\mathcal{Y} \in \{0,1\}^{H \times W}$ denote the binary ground truth mask, where $\mathcal{Y}_{ij} = 1$ indicate and OoD pixel. Then the set of ground-truth OoD pixels are defined as $\mathcal{Y}_1 = \{(i,j) \mid \mathcal{Y}_{i,j} = 1\}$. Similarly, given an OoD score map $s(x)$ and a threshold $\delta \in \mathbb{R}$, the set of predicted OoD pixels is defined as $\hat{\mathcal{Y}}_1(\delta) = \{(i,j) \mid s(x)_{i,j} > \delta\}$ The precision and recall at a threshold δ are then defined as:

$$\text{precision}(\delta) = \frac{|\mathcal{Y}_1 \cap \hat{\mathcal{Y}}_1(\delta)|}{|\hat{\mathcal{Y}}_1(\delta)| + \epsilon}, \quad \text{recall}(\delta) = \frac{|\mathcal{Y}_1 \cap \hat{\mathcal{Y}}_1(\delta)|}{|\mathcal{Y}_1| + \epsilon} \tag{2}$$

where ϵ is a small constant to avoid division by zero.

A more safety-relevant metric is the false positive rate at 95% true positive rate (FPR_{95}). Here, the true positive rate is equal to the recall of the OoD class, and the false positive rate is the number of pixels falsely predicted as OoD over all in-distribution pixels. Formally, FPR_{95} is then calculated as:

$$\text{FPR}_{95} = \min_{\delta \in \mathbb{R}} \left\{ \frac{|\hat{\mathcal{Y}}_1(\delta) \cap \mathcal{Y}_0|}{|\mathcal{Y}_0|} \;\middle|\; \text{recall}(\delta) \geq 0.95 \right\} \tag{3}$$

where: $\mathcal{Y}_0 = \{(i,j) \mid \mathcal{Y}_{i,j} = 0\}$ represents the set of in-distribution pixels.

One disadvantage of pixel-level metrics is that they may be biased against small OoD objects, potentially neglecting their contribution to the evaluation. From a practical perspective, one is interested in detecting all OoD objects, regardless of their size *i.e.*, the number of pixels an object covers. Therefore, component-level metrics provide a more holistic evaluation by assessing model performance at the level of components rather than pixels.

A component is defined as a set of connected pixels that share the same label. Let $\mathcal{K} = \{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_m\}$ denote the set of ground-truth OoD components, where $\mathcal{K}_i$ is the set of connected pixels labelled as OoD in the ground truth mask. Similarly, let $\hat{\mathcal{K}} = \{\hat{\mathcal{K}}_1, \hat{\mathcal{K}}_2, \ldots, \hat{\mathcal{K}}_n\}$ be the set of predicted OoD components. The component-wise intersection over union (sIoU) for a ground truth component $k \in \mathcal{K}$ is defined as:

$$\text{sIoU}(k) := \frac{|k \cap \hat{\mathcal{K}}(k)|}{|(k \cup \hat{\mathcal{K}}(k)) \setminus \mathcal{A}(k)|} \in [0,1] \tag{4}$$

where $\hat{\mathcal{K}}(k) = \bigcup_{\hat{k} \in \hat{\mathcal{K}}, \hat{k} \cap k \neq \emptyset} \hat{k}$ and $\mathcal{A}(k) = \bigcup_{k_0 \in \mathcal{K}, k_0 \neq k} k_0$. The sIoU captures how well a ground-truth component k aligns with predicted components, given a threshold $\tau \in [0,1)$, a target $k \in \mathcal{K}$ is considered a true positive (TP) if $\text{sIoU}(k) > \tau$, and a false negative (FN) otherwise. A predicted component $\hat{k}$ is considered a False Positive (FP) if its Positive Predictive Value (PPV) is $\leq \tau$, meaning it has insufficient overlap with any ground-truth component. The PPV for $\hat{k} \in \hat{\mathcal{K}}$ is defined as:

$$\text{PPV}(\hat{k}) := \frac{|\hat{k} \cap \mathcal{K}(\hat{k})|}{|\hat{k}|} \in [0,1]. \tag{5}$$

To summarize the TPs, FPs, and FNs, the mean F_1-score is averaged across multiple thresholds to summarize the detection performance of a model. The F_1-score at a threshold τ is calculated as follows:

$$F_1(\tau) := \frac{2 \cdot \text{TP}(\tau)}{2 \cdot \text{TP}(\tau) + \text{FP}(\tau) + \text{FN}(\tau)} \in [0,1] \tag{6}$$

As the numbers of TPs, FPs, and FNs depend on the detection threshold τ, the average F1 score over different τ is used to give $\overline{F1}$ as an evaluation metric that is less affected by the detection threshold.

Method	Configuration					Metrics				
						Pixel Level		Component Level		
	OE	M2F	GM	UE	Other	AUPRC ↑	FPR_{95} ↓	$sIoU_{gt}$ ↑	PPV ↑	$\overline{F1}$ ↑
RbA [40]	✓	✓	✗	✗	✗	**95.12**	**0.08**	53.34	59.08	57.44
UEM [41]	✓	✗	✗	✗	✓	<u>94.38</u>	<u>0.10</u>	57.90	41.20	43.30
Mask2Anomaly [46]	✓	✓	✗	✗	✗	93.22	0.20	55.72	75.42	68.15
UNO [10]	✓	✓	✗	✗	✗	93.19	0.16	**70.97**	72.17	<u>77.65</u>
EAM [19]	✓	✓	✗	✗	✗	92.87	0.52	<u>65.85</u>	<u>76.50</u>	75.58
LR [50]	✓	✗	✗	✗	✓	92.00	0.20	62.90	**81.90**	**78.40**
PixOOD [59]	✗	✗	✗	✗	✓	88.90	0.30	42.68	57.49	50.82
RbA	✗	✓	✗	✗	✗	87.85	3.33	47.44	56.16	50.42
CLS [64]	✗	✓	✗	✗	✗	87.10	0.67	44.70	53.13	51.02
DenseHybrid [18]	✓	✗	✓	✗	✗	87.08	0.24	45.74	50.10	50.72
Maximized Entropy [6]	✓	✗	✗	✓	✗	85.07	0.75	47.87	62.64	48.51
DaCUP [57]	✗	✗	✓	✗	✗	81.50	1.13	37.68	60.13	46.01
ATTA [16]	✓	✗	✗	✗	✓	76.26	2.81	43.93	37.66	36.57
FlowEneDet [20]	✗	✗	✓	✗	✗	73.71	0.97	42.62	42.25	39.96
SynBoost [11]	✗	✗	✓	✗	✗	71.34	3.15	44.28	41.75	37.57
DOoD [15]	✗	✗	✓	✗	✗	64.30	2.60	30.30	33.10	24.10
Image Resynthesis [32]	✗	✗	✓	✗	✗	37.71	4.70	16.61	20.48	8.38
JSRNet [56]	✗	✗	✓	✗	✗	28.09	28.86	18.55	24.46	11.02
PGN [35]	✗	✗	✗	✓	✗	16.52	19.69	19.42	14.89	7.39
Maximum Softmax [21]	✗	✗	✗	✓	✗	15.72	16.60	19.72	15.93	6.25
PEBAL [52]	✓	✗	✓	✗	✗	4.98	12.68	29.91	7.55	5.54
MC Dropout [38]	✗	✗	✗	✓	✗	4.88	50.31	5.49	5.77	1.05
Ensemble [29]	✗	✗	✗	✓	✗	1.06	77.20	8.63	4.71	1.28
Embedding Density [2]	✗	✗	✓	✗	✗	0.82	46.38	35.64	2.87	2.31

Table 1. **Quantitative Results on SMIYC-OT**: Comparison of current methods on the SMIYC-OT dataset. Methods are sorted from top to bottom according to their AP. The best result for each metric is highlighted in **bold**, and the second best is <u>underlined</u>. Methods are categorized into four groups: Mask2Former-based (M2F), Uncertainty Estimation (UE), Generative Models (GM), and others.

Dataset: We evaluate the performance of various methods on the SegmentMeIfYouCan Obstacle Track (SMIYC-OT) and LostAndFound-NoKnown (L&F) datasets [5]. Both benchmarks assess the ability to segment OoD objects on the road that do not belong to any of the predefined classes in the Cityscapes dataset [9]. While other OoD segmentation datasets exist, such as Fishyscapes [3] and RoadAnomaly [32], we exclude them from this survey for specific reasons. Fishyscapes Static uses synthetic anomalies overlaid from unrelated datasets, resembling outlier exposure and risking overfitting to artificial features rather than generalizing to real-world unknowns. RoadAnomaly introduces a significant domain shift from Cityscapes due to differing camera viewpoints and road-level perspectives, leading to unrealistic scenarios for standard autonomous driving. We, therefore, focus on SMIYC-OT and L&F, which offer a more balanced and representative evaluation for real-world OoD segmentation.

3. Current State of the Art

Table 1 and 2 summarise the performance of different methods on the SMIYC-OT and LostAndFound-NoKnown benchmarks, respectively. We report both pixel-level and component-level metrics for each method. For comparing the different methods, we categorize the methods into four distinct groups:

1. Mask2Former-based (M2F): Methods that leverage special features from the Mask2former architecture [8].
2. Uncertainty Estimation (UE): Approaches that rely on confidence or uncertainty estimates.
3. Generative Models (GM): Methods that utilize generative modelling for OoD scoring.
4. Other: Methods that do not fall into the above categories but offer unique approaches to OoD detection.

In addition, we also denote if the model used any outlier data to improve OoD detection performance, a common technique referred to as *outlier exposure* (OE) [22].

Method	Configuration					Metrics				
						Pixel Level		Component Level		
	OE	M2F	GM	UE	Other	AUPRC ↑	$FPR \downarrow$	$sIoU_{gt} \uparrow$	$PPV \uparrow$	$\overline{F1} \uparrow$
PixOOD [59]	✗	✗	✗	✗	✓	**85.07**	4.46	30.18	<u>78.47</u>	44.41
LR [50]	✓	✗	✗	✗	✓	<u>83.70</u>	100.00	**49.70**	**95.90**	**72.60**
DaCUP [57]	✗	✗	✓	✗	✗	81.37	7.36	38.34	67.29	51.14
UEM [41]	✓	✗	✗	✗	✓	81.04	**1.45**	39.49	55.77	46.54
FlowEneDet [20]	✗	✗	✓	✗	✗	79.75	2.92	43.79	52.83	48.05
DOoD [15]	✗	✗	✓	✗	✗	79.50	3.70	27.70	40.60	30.5
RbA [40]	✓	✓	✗	✗	✗	79.25	22.57	45.56	68.90	<u>59.31</u>
DenseHybrid [18]	✓	✗	✓	✗	✗	78.67	<u>2.12</u>	<u>46.90</u>	52.14	52.33
Maximized Entropy [6]	✓	✗	✗	✓	✗	77.90	9.70	45.90	63.06	49.92
JSRNet [56]	✗	✗	✓	✗	✗	74.17	6.59	34.28	45.89	35.97
PGN [35]	✗	✗	✗	✓	✗	69.30	9.80	50.00	44.80	45.40
Embedding Density [2]	✗	✗	✓	✗	✗	61.70	10.36	37.75	35.21	27.55
Image Resynthesis [32]	✗	✗	✓	✗	✗	57.08	8.82	27.16	30.69	19.17
MC Dropout [38]	✗	✗	✗	✓	✗	36.78	35.55	17.35	34.71	12.99
Maximum Softmax [21]	✗	✗	✗	✓	✗	30.14	33.20	14.20	62.23	10.32
Ensemble [29]	✗	✗	✗	✓	✗	2.89	82.03	6.66	7.64	2.68

Table 2. **Quantitative Results on L&F**: Comparison of current methods on the L&F dataset. Methods are sorted from top to bottom according to their AP. The best result for each metric is highlighted in **bold**, and the second best is <u>underlined</u>. Methods are categorized into four groups: Mask2Former-based (M2F), Uncertainty Estimation (UE), Generative Models (GM), and others.

3.1. Outlier Exposure

In the setting where one has knowledge about potential OoD objects a network may encounter, one can improve separation between in-distribution and OoD objects using proxy OoD objects. Proxy OoD data, often referred to as *known unknowns*, are samples that do not belong to the in-distribution classes but resemble potential OoD objects the model might encounter in deployment. Proxy OoD data can then be used to regularize the feature space of the model by learning representations for unknown objects; this technique is referred to as *outlier exposure*. The assumption here is that exposure to proxy OoD during training can help the model generalize to unseen OoD objects in deployment.

For OoD segmentation, outlier exposure has been applied in various ways to enhance OoD detection performance. A common approach involves using a subset of objects from COCO [31] or ADE20K [66] as proxy OoD objects. These objects are then pasted into in-distribution images during a fine-tuning stage, either as cutouts pasted on the inlier data or as complete images. Maximized Entropy [6] encourages higher entropy predictions on proxy OoD samples to prevent overconfidence in uncertain regions, while RbA [40] and PEBAL [52] explicitly train the model to produce lower logit scores for these proxy OoD samples. The general idea is to train the model to learn certain heuristics that can later be used to detect out-of-distribution inputs. Other approaches explicitly train the model on the proxy OoD samples to model unknown objects as a separate class.

While proxy outliers in road obstacle detection can be justified by defining OoD objects as entities not present in the training set, potentially capturing a broad notion of "objectness", this strategy raises notable concerns. The primary issue is the risk of overfitting to seen examples, which may limit the model's ability to generalize to truly novel obstacles in real-world scenarios. Moreover, the variability in datasets used for outlier exposure across different methods calls into question how the choice of outliers impacts the effectiveness of these approaches. Despite its importance, little research explores how outlier selection impacts OoD detection, particularly given the limited object categories in current benchmarks. We highlight the need for further exploration to determine optimal strategies for selecting proxy data for outlier exposure.

3.2. Uncertainty Estimation

A neural network with a softmax output layer can be interpreted as a statistical model $f_\theta(y|x)$, which assigns a probability distribution over the n class labels $y \in \mathcal{C} = \{y_1, \ldots, y_n\}$ for each pixel in the image. This distribution depends on the parameters θ and the input x. Some of the first approaches proposed for OoD segmentation were to use uncertainty metrics on these probability scores like: max probability [21], softmax-entropy [6], approxi-

mations over the posterior distribution of the weights of the model [38] or predictive entropy of deep ensembles [29] to estimate the OoD score of pixels.

While these uncertainty-based metrics capture different aspects of model confidence, it is important to distinguish between epistemic, aleatoric, and predictive uncertainty. This decomposition is not universally agreed upon across all disciplines but follows a widely accepted framework in the probabilistic modelling community, as introduced in [13]. In this framework, epistemic uncertainty, also known as model uncertainty, arises from the lack of knowledge about the true model parameters and can be reduced with more training data; this is typically the uncertainty one is interested in for OoD detection, as it can be reduced by collecting more data. The epistemic uncertainty is quantified as the expected information gain between the network parameters θ and the output $\mathbb{I}[Y;\theta|x,\mathcal{D}]$. On the other hand, aleatoric uncertainty represents inherent noise in the data, such as sensor noise, and cannot be reduced by collecting more data. Aleatoric uncertainty $\mathbb{E}_{p(\theta|\mathcal{D})}[\mathbb{H}[Y|x,\theta]]$, represents the expected entropy of the output distribution given fixed model parameters. Predictive uncertainty is the total uncertainty in the model's predictions and is composed of both epistemic and aleatoric uncertainty:

$$\mathbb{H}[Y|x,\mathcal{D}] = \mathbb{I}[Y;\theta|x,\mathcal{D}] + \mathbb{E}_{p(\theta|\mathcal{D})}[\mathbb{H}[Y|x,\theta]]. \quad (7)$$

Eq.7 highlights a key limitation of using predictive entropy for OoD detection: it does not distinguish between epistemic and aleatoric uncertainty. As a result, high predictive entropy does not necessarily indicate an OoD sample; it may also correspond to an ambiguous in-distribution input. Thus, predictive entropy is only a reliable OoD measure in the absence of ambiguous samples. While deep ensembles and Bayesian models improve epistemic uncertainty estimation, predictive entropy alone remains an unreliable OoD detection metric since it conflates epistemic and aleatoric uncertainty. For a single deterministic model, methods like softmax entropy can be used as a proxy for uncertainty. However, empirical findings from [28] suggest that while deep ensembles can effectively capture epistemic uncertainty via mutual information, softmax entropy in a deterministic model fails to do so. In practice, separating the two components of uncertainty is often very difficult, as they are strongly correlated in many practical settings [37].

Training with outlier exposure reduces epistemic uncertainty by incorporating OoD examples into the learning process, effectively shifting uncertainty from epistemic to aleatoric. As a result, softmax and predictive entropy become more reliable uncertainty measures. However, this raises a fundamental question: if a model has been trained with outliers, are these samples still truly OoD? Furthermore, could excessive exposure to outliers lead to overgeneralization, reducing the ability of the model to separate inliers from outliers?

An alternative approach for quantifying pixel-level uncertainty in OoD segmentation was proposed in PGN [35]. Their approach computes gradient-based uncertainty scores at the pixel level by leveraging loss gradients during inference. The intuition behind this approach is that OoD pixels exhibit high gradient magnitudes, as they do not align well with any in-distribution class.

Uncertainty-based methods generally rank at the bottom of the benchmark (with the exception of Maximized Entropy [6] which utilized outlier supervision and a post-processing method called *Meta-Classification* [48] to eliminate false-positives), this is because closed-set models are not necessarily calibrated, often leading to overconfident predictions on unseen classes [43].

3.3. Generative Models

An intuitive approach for OoD detection that does not require any labelled samples is to fit a density model to the in-distribution samples and then use the inverse of the likelihood $p(x)$ as an OoD score. One of the first approaches for OoD segmentation [2] followed this approach and used a normalizing flow [12] to approximate the density of intermediate embeddings of a DeepLabv3+ segmentation model [7]. This approach was later extended in FlowEneDet [20], where the authors use a 2D Glow-like [24] architecture trained using correctly classified pixels as positive samples and the misclassified pixels as negative samples. The proposed architecture in FlowEneDet is an extra network that can be added to any semantic segmentation network. Beyond pure generative methods, an alternative perspective views discriminative models with softmax classifiers as energy-based models for the joint distribution $p(x,y)$ through the LogSumExp scoring function on the logits [17]. This interpretation led to approaches like PEBAL [52], which improved upon simple maximum-logit scoring for OoD segmentation. More recently, hybrid approaches such as DenseHybrid [18] have combined generative modelling of the inlier data with discriminative training of negative samples via outlier exposure to enhance OoD segmentation.

The main disadvantage of using generative models for OoD detection via density estimation is that they do not explicitly differentiate between outliers within the in-distribution data and true OoD samples. Since the in-distribution data itself may contain outlier points (e.g., samples from the tail of the distribution), the model might assign low likelihoods to these points, potentially misclassifying them as OoD. Additionally, it has been empirically shown that generative models trained on one dataset assign a higher likelihood to samples from OoD datasets [26, 39, 47, 65] in image classification tasks. The authors in [63] argue that the principled way of doing OoD detection is

using the likelihood ratios between a proxy OoD estimate and the in-distribution estimate. In one of their formulations, the authors of [41] use the likelihood ratios with the inlier and OoD estimates obtained using a variant of GMM-seg [30]. This method is currently the highest-performing generative model on the benchmark. However, the authors found that using a discriminative model in the likelihood ratios resulted in better performance.

Another approach for using generative models for OoD segmentation involves using generative models like variational autoencoders [25] or conditional generative adversarial networks [60] for re-synthesising an image from its predicted semantic map and comparing it with the original image [32]. The discrepancies between input and generated outputs are then used as an OoD score. This approach was further expanded to integrated uncertainty estimation with image re-synthesizing to refine the OoD segmentation in SynBoost [11]. Building on the idea of detecting OoD objects based on the difference between input and generated output, JSRNeT [56] proposes a generative discriminative reconstruction module that can be added on top of a fixed semantic segmentation network. However, instead of generating a complete image, this reconstruction module learns the appearance of the road. This method was then extended in DaCUP [57], which introduced an embedding bottleneck to better model multi-modal road appearances and improve training data utilization. Additionally, a distance-based embedding scoring feature was introduced to refine anomaly detection, along with an inpainting module that filters out false positives by identifying regions reconstructable from their surroundings. These improvements were integrated into a coupling module that combines multiple anomaly estimation signals to enhance overall performance. Diffusion models have also been explored for OoD segmentation in DOoD [15], where the denoising score is used to compute per-pixel OoD scores.

A key disadvantage of image re-synthesis and reconstruction approaches is their high inference time. These methods require multiple processing steps: predicting a semantic map, generating a synthesized or reconstructed image, and then computing discrepancies between the input and the output. This additional computational overhead makes real-time applications particularly challenging.

3.4. Mask2Former

Five of the top seven methods on this benchmark depend on the Mask2former [8] architecture for detecting OoD objects. Mask2former is a mask-level segmentation architecture that decouples object localization and classification by splitting the task into two steps. Given an $H \times W$ sized image, Mask2former computes N pairs $\{(\mathbf{m}_i, \mathbf{p}_i)\}_{i=1}^{N}$, where $\mathbf{m}_i \in [0,1]^{H \times W}$ are mask predictions associated with some semantically related regions in the input image and $\mathbf{p}_i \in [0,1]^{K+1}$ class probabilities classifying to which semantic category the mask $\mathbf{m}_i$ belongs to. Here, the masks can be assigned to one of the K known Cityscapes classes or to one auxiliary void class. The final semantic segmentation inference is carried out by an ensemble-like approach over the pairs $\{(\mathbf{m}_i, \mathbf{p}_i)\}_{i=1}^{N}$ yielding pixel-wise class scores

$$\mathbf{q}[h,w,k] = \sum_{i=1}^{N} \mathbf{p}_i(k) \cdot \mathbf{m}_i[h,w] \quad \in [0,N] \tag{8}$$

for image pixel locations $h = 1,\dots,H, w = 1,\dots,W$ and classes $k = 1,\dots,K$. The authors of RbA [40] noticed that each object query acts as a one-vs-all classifier and that OoD objects can be detected by determining whether a pixel is associated to any of the k known classes. The OoD score is defined as:

$$\mathbf{RbA}[h,w] = -\sum_{k=1}^{K} \phi(\mathbf{q}[h,w,k]) \quad \in [0,K] \tag{9}$$

with ϕ being the $\tanh$ activation function.

Other methods that utilize the mask-level architecture of Mask2former and the attention-based transformer decoder include Mask2Anomaly [46], EAM [19], and UNO [10]. Mask2Anomaly uses a contrastive loss in the outlier supervision process to separate known masks from proxy OoD masks. EAM ensembles region-wide outlier scores from the mask and pixel classifiers in the network architecture. To boost performance, they utilize the additional void class included in the transformer decoder in the outlier supervision process to instruct all masks to avoid the proxy OoD examples. UNO extends on EAM but uses an additional object query in the transformer decoder to learn the OoD objects. The final OoD score is an ensemble of the void and proxy OoD object scores.

In CSL [64], the authors introduce a class-agnostic segmentation framework that enhances OoD detection by integrating structural constraints into existing methods. CSL extends an adjusted version of Mask2Former with a base teacher network and an MLP, generating per-pixel distributions that are fused with inlier region proposals to produce the final OoD prediction. The framework employs soft assignment and mask-split preprocessing to refine segmentation boundaries and improve generalization to unseen classes.

One of the main disadvantages of these approaches is that they are specific to the Mask2former architecture and the transformer decoder, which acts as a one-vs-all classifier. Therefore, these methods are usually not transferable to other architectures. Additionally, RbA, UNO, and EAM utilized the Swin-L [34] backbone, which has $\approx$ 200M parameter, making them impractical for real-time deployment, particularly in resource-constrained environments like autonomous vehicles.

3.5. Other

One of the main limitations of the methods that utilize outlier supervision is that the fine-tuning step decreases the performance of the original inlier model. In our own experiments, maximized entropy decreased the inlier performance by 1.6 % mIoU points, and RbA decreased by 2.5 % mIoU. In many practical scenarios, this is highly undesirable as OoD objects typically occur far less often than inlier objects. Instead of adjusting the inlier parameters of the model during the outlier supervision process, the authors of UEM [41] proposed an external module explicitly designed for OoD segmentation; this allows to keep the original inlier performance intact while performing very well in detecting OoD objects. The authors also suggested using the Likelihood Ratios between the learned distribution of outliers from the proxy OoD dataset and that of the original inlier model to calculate the OoD score. The authors found that using the feature representation of self-supervised pre-trained models like DinoV2 [44] resulted in better modelling of OoD objects than fully supervised models like Swin [33] or contrastive image-language pre-trained models like CLIP [45]

Motivated by the use of likelihood ratios in UEM, the authors of LR [50] extend the idea and use features from the foundation model Segment Anything Model [27] (SAM) for OoD segmentation. One of the main advantages of SAM is that it moves away from pixel-level reasoning and reasons about segment-level masks, which results in more semantically meaningful segmentations. The main limitation of this approach is that SAM is limited to prompted prediction, and if an object is not prompted, it will not be assigned a mask. Other approaches [42] have also proposed using SAM but prompt the predictions based on an existing OoD segmentation model to refine the pixel predictions; this removes some of the false positives and merges some of the predictions to get more coherent segments. However, if the initial OoD segmentation network misses an object, SAM will not detect it.

In PixOOD [59], the authors propose an OoD segmentation method that does not require any outlier exposure. Their approach models the intra-class in-distribution variability through an online data condensation algorithm and consists of three components: 1) extracting the patch feature representation from a DinoV2 model, 2) building a 2D-projection space for each in-distribution class, and 3) finding calibrated in-distribution/OoD decision strategy. The framework builds on the methodology introduced in [58] but extends the approach for pixel-level predictions. A similar direction is explored in [14], where the authors apply a k-nearest-neighbour-based strategy based on patch feature representations for dense OoD detection.

In ATTA [16], the authors address the challenge of OoD segmentation under domain shift by proposing a dual-level OoD detection framework. Their method simultaneously handles domain and semantic shifts to improve OoD segmentation robustness. ATTA detects domain shifts at the first level by leveraging global low-level features, enabling selective adaptation of the model to unseen domains. At the second level, it identifies OoD pixels caused by semantic shifts using dense high-level feature maps. The framework employs a $(\mathcal{C}+1)$-class probabilistic classifier, where the first $\mathcal{C}$ classes represent the in-distribution categories, and an additional class is designated for OoD objects. ATTA utilizes an anomaly-aware self-training procedure to refine OoD predictions dynamically. This post hoc adaptation can be applied to various OoD segmentation models, though it does introduce additional computational overhead, nearly doubling inference time.

4. Limitations

While the benchmarks provide a structured framework for assessing the OoD segmentation capabilities of different methods, several design choices and constraints raise concerns about their suitability for real-world deployment. In this section, we analyze some of the benchmark's key limitations.

Region of Interest: Both benchmarks limit the evaluation of detection OoD objects to a predefined region of interest (i.e., the road). This design choice ensures that only the objects of interest are evaluated. In practice, this predefined region is often obtained from an auxiliary network that segments the drivable area, effectively reducing the OoD segmentation task to a foreground-background segmentation problem [62]. Some methods explicitly utilize this assumption [36, 50, 56, 57], while others did not make direct use of this predefined region of interest assumption. For road obstacle detection, an open question remains: is it more efficient to use a segmentation network with inherent OoD detection capabilities or to first generate a region of interest and then apply a foreground-background segmentation approach to detect OoD objects?

On Small Objects: Another limitation of the benchmark design is the removal of small predicted regions below a predefined threshold of 50 pixels. While this filtering step may reduce noise and stabilize evaluation metrics, it undermines one of the key advantages of segmentation models, their ability to detect small anomalous objects. OoD objects in autonomous driving scenarios, such as lost cargo or road debris, are often safety-critical, and segmentation models are designed to capture fine-grained spatial details. By discarding small predictions, the benchmark may underestimate the effectiveness of methods capable of detecting such small structures. Moreover, this design choice does not adequately penalize models that produce fragmented false predictions, as small, scattered errors may be ignored rather than contributing to the overall error metric.

Threshold Selection: One limitation of existing benchmarks for OoD segmentation is their reliance on threshold-free metrics (e.g., AUPRC), which, while useful for assessing performance independent of a specific threshold, are not directly applicable to downstream tasks. In practice, a threshold must be selected to make reliable decisions, and ideally, it should effectively separate in-distribution and out-of-distribution regions with high confidence. Furthermore, the SMIYC-OT benchmark assesses component-level performance for practical reasons outlined in Section 2. However, their evaluation derives an optimal threshold from pixel-level metrics using the optimal pixel-level F1 score to assess component-level performance. This approach is troublesome for deployment scenarios, as ground-truth labels for OoD objects are often not available in large numbers for calibrating the OoD score, making it difficult to select an optimal threshold in real-world settings. Moreover, this methodology obscures a critical limitation of some approaches: a narrow range of effective OoD scores. These methods are highly sensitive to threshold selection, which means small variations can significantly impact their performance. This sensitivity is not adequately captured with current metrics, which may misrepresent the robustness of different approaches. Metrics like the Maximal Detection Margin [1], are a step in the right direction to mitigate some of these issues.

5. Future Directions

Although OoD segmentation on benchmark data sets has reached a certain level of maturity and offers a broad choice of performant models, the question arises of how to proceed with the research on OoD segmentation. Here, we provide a non-exhaustive list of topics.

Downstream Tasks: The use case of OoD Segmentation should be sharpened. There seems to be little research on how OoD segmentation influences downstream tasks like building environmental models or planning in automated driving or robotics [61]. Intuitively, it is clear that OoD objects should not be hit. But as there is little work on the suspected behaviour of an OoD object, how to avoid something we don't know might be complicated. Avoiding an animal on the road requires a different response than avoiding lost cargo. Beyond safety concerns in autonomous systems, OoD segmentation can also play a crucial role in self-monitoring AI-based perception algorithms and retrieval of unusual objects [49]. This can lead to a better understanding of the operational domain of deployment (ODD). Research on how to build continuous learning strategies on top of the retrieved data is still in its infancy; see, however, see [51, 53, 54] for first steps.

In addition to detection, context-aware decision-making is essential: not all OoD objects pose the same level of risk. Thus, it becomes crucial to interpret the OoD detection in light of the surrounding scene context and possible future interactions. This requires integrating scene understanding and behaviour prediction with the OoD segmentation pipeline. For instance, distinguishing between a static roadside billboard and a moving pedestrian in an unknown outfit directly influences the vehicle's response. Such scene-level understanding supports more adaptive and risk-sensitive planning strategies in real-world applications.

Real-time Deployment: Recently, academic OoD segmentation research has profited much from large-scale foundation models. However, OoD segmentation in the wild requires the deployment of OoD segmentation algorithms on edge devices. To bridge the gap between academic research and real-world deployment, future work must address key efficiency challenges: reducing inference times, ensuring compatibility with multi-task networks, and optimizing OoD models for edge devices through quantization and knowledge distillation. A balance between computational efficiency and segmentation accuracy is critical for industrial adoption

New Datasets: As presently available benchmark datasets have saturated, new datasets should emerge and lead the field into the future. The present state of OoD detection suffers from under-complex street scenes [4]. This is mainly due to safety requirements in recording OoD data sets. In this way, the effects of immersion of OoD objects into ID scenes, occlusion, and dynamical OoD objects are still to be integrated into contemporary OoD research. The next frontier in OoD segmentation research requires datasets that reflect real-world complexities incorporating temporal dynamics, multimodal sensor inputs with a camera, LiDAR and RaDAR at least, but potentially also with depth cameras, multi-spectral imaging and acoustic sensors. It would be desirable if semantic labels were available to measure domain performance. One way to achieve this at a limited cost would be to utilize pseudo labels from highly performant in domain networks [23]. However, this induces a bias towards utilizing the same network, and this might stand in the way of future improvement of both ID and OoD segmentation.

As we documented in this survey, the rapid progress in OoD segmentation, driven by benchmark datasets, has led to an impressive state-of-the-art. However, the true potential of this field lies in its real-world deployment: ensuring safety in autonomous systems, improving AI-driven decision-making, and pushing the boundaries of open-world perception. Looking into the future, integrating multimodal learning, continuous adaptation, and real-time deployment will be key to unlocking the next generation of OoD segmentation models used in automated driving systems.

Acknowledgement: The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project "just better DATA".

References

[1] Jan Ackermann, Christos Sakaridis, and Fisher Yu. Maskomaly: Zero-shot mask anomaly segmentation. In *The British Machine Vision Conference (BMVC)*, 2023. 8

[2] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3, 4, 5

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129(11):3119–3135, 2021. 3

[4] Daniel Bogdoll, Svenja Uhlemeyer, Kamil Kowol, and J Marius Zöllner. Perception datasets for anomaly detection in autonomous driving: A survey. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–8. IEEE, 2023. 8

[5] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 3

[6] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, 2021. 3, 4, 5

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 6

[9] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[10] Anja Delić, Matej Grcić, and Siniša Šegvić. Outlier detection by ensembling uncertainty with negative objectness, 2024. 3, 6

[11] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021. 3, 6

[12] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. 5

[13] Yarin Gal et al. Uncertainty in deep learning. 2016. 5

[14] S. Galesso, M. Argus, and T. Brox. Far away in the deep space: Dense nearest-neighbor-based out-of-distribution detection. In *IEEE International Conference on Computer Vision (ICCV), Workshops*, 2023. 7

[15] Silvio Galesso, Philipp Schröppel, Hssan Driss, and Thomas Brox. Diffusion for out-of-distribution detection on road scenes and beyond. In *ECCV*, 2024. 3, 4, 6

[16] Zhitong Gao, Shipeng Yan, and Xuming He. Atta: anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. *Advances in Neural Information Processing Systems*, 36:45150–45171, 2023. 3, 7

[17] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020. 5

[18] Matej Grcic, Petra Bevandic, and Sinisa Segvic. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *17th European Conference on Computer Vision ECCV 2022*. Springer, 2022. 3, 4, 5

[19] Matej Grcić, Josip Šarić, and Siniša Šegvić. On advantages of mask-level recognition for outlier-aware segmentation. In *CVPRW*, 2023. 3, 6

[20] Denis Gudovskiy, Tomoyuki Okuno, and Yohei Nakata. Concurrent misclassification and out-of-distribution detection for semantic segmentation via energy-based normalizing flow. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. JMLR.org, 2023. 3, 4, 5

[21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 3, 4

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 3

[23] Christoph Hümmer, Manuel Schwonberg, Liangwei Zhou, Hu Cao, Alois Knoll, and Hanno Gottschalk. Strong but simple: A baseline for domain generalized dense perception by clip-based transfer learning. In *Proceedings of the Asian Conference on Computer Vision*, pages 4223–4244, 2024. 8

[24] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 5

[25] Diederik P Kingma and Max Welling. Stochastic gradient vb and the variational auto-encoder. In *Second international conference on learning representations, ICLR*, page 121, 2014. 6

[26] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33:20578–20589, 2020. 5

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 7

[28] A Kirsch, J Mukhoti, J van Amersfoort, PHS Torr, and Y Gal. On pitfalls in ood detection: Entropy considered harmful. In *ICML: Uncertainty Robustness in Deep Learning Workshop*, 2021. 5

[29] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3, 4, 5

[30] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. In *Advances in Neural Information Processing Systems*, 2022. 6

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[32] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161, 2019. 3, 4, 6

[33] 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

[34] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[35] Kira Maag and Tobias Riedlinger. Pixel-wise gradient uncertainty for convolutional neural networks applied to out-of-distribution segmentation. In *VISIGRAPP (2): VISAPP*, 2024. 3, 4, 5

[36] Samuel Marschall and Kira Maag. Multi-scale foreground-background confidence for out-of-distribution segmentation. *arXiv preprint arXiv:2412.16990*, 2024. 7

[37] Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in Neural Information Processing Systems*, 37:50972–51038, 2025. 5

[38] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 3, 4, 5

[39] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019. 5

[40] Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *ICCV*, 2023. 3, 4, 6

[41] Nazir Nayal, Youssef Shoeb, and Fatma Güney. A likelihood ratio-based approach to segmenting unknown objects. *arXiv preprint*, arXiv:2409.06424, 2024. 3, 4, 6, 7

[42] Alexey Nekrasov, Alexander Hermans, Lars Kuhnert, and Bastian Leibe. UGainS: Uncertainty Guided Anomaly Instance Segmentation. In *GCPR*, 2023. 7

[43] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 5

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 7

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 7

[46] S. Rai, F. Cermelli, D. Fontanel, C. Masone, and B. Caputo. Unmasking anomalies in road-scene segmentation. In *ICCV*, 2023. 3, 6

[47] Jie Jessie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *NeurIPS*, 2019. 5

[48] Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020. 5

[49] Youssef Shoeb, Robin Chan, Gesina Schwalbe, Azarm Nowzad, Fatma Güney, and Hanno Gottschalk. Have we ever encountered this before? retrieving out-of-distribution road obstacles from driving scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7396–7406, 2024. 8

[50] Youssef Shoeb, Nazir Nayal, Azarm Nowzard, Fatma Güney, and Hanno Gottschalk. Segment-level road obstacle detection using visual foundation model priors and likelihood ratios. *arXiv preprint arXiv:2412.05707*, 2024. 3, 4, 7

[51] Youssef Shoeb, Azarm Nowzad, and Hanno Gottschalk. Adaptive neural networks for intelligent data-driven development. *arXiv preprint*, arXiv:2502.10603, 2025. 8

[52] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex

urban driving scenes. In *European Conference on Computer Vision*, pages 246–263. Springer, 2022. 3, 4, 5

[53] Svenja Uhlemeyer, Matthias Rottmann, and Hanno Gottschalk. Towards unsupervised open world semantic segmentation. In *Uncertainty in Artificial Intelligence*, pages 1981–1991. PMLR, 2022. 8

[54] Svenja Uhlemeyer, Julian Lienen, Youssef Shoeb, Eyke Hüllermeier, and Hanno Gottschalk. Unsupervised class incremental learning using empty classes. In *Proceedings of British Machine Vision Conference Workshops*, 2024. 8

[55] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. 1

[56] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15651–15660, 2021. 3, 4, 6, 7

[57] Tomáš Vojíř and Jiří Matas. Image-consistent detection of road anomalies as unpredictable patches. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5491–5500, 2023. 3, 4, 6, 7

[58] Tomáš Vojíř, Jan Šochman, Rahaf Aljundi, and Jiří Matas. Calibrated out-of-distribution detection with a generic representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4507–4516, 2023. 7

[59] Tomáš Vojíř, Jan Šochman, and Jiří Matas. PixOOD: Pixel-Level Out-of-Distribution Detection. In *ECCV*, 2024. 3, 4, 7

[60] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 6

[61] Julian Wörmann, Daniel Bogdoll, Christian Brunner, Etienne Bührle, Han Chen, Evaristus Fuh Chuo, Kostadin Cvejoski, Ludger van Elst, Philip Gottschall, Stefan Griesche, et al. Knowledge augmented machine learning with applications in autonomous driving: A survey. *arXiv preprint arXiv:2205.04712*, 2022. 8

[62] Zuyao You, Lingyu Kong, Lingchen Meng, and Zuxuan Wu. FOCUS: Towards universal foreground segmentation. In *AAAI*, 2025. 7

[63] Andi Zhang and Damon Wischik. Falsehoods that ml researchers believe about ood detection. In *NeurIPS ML Safety Workshop*, 2022. 5

[64] Hao Zhang, Fang Li, Lu Qi, Ming-Hsuan Yang, and Narendra Ahuja. Csl: Class-agnostic structure-constrained learning for segmentation including the unseen. *AAAI*, 2024. 3, 6

[65] Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pages 12427–12436. PMLR, 2021. 5

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 4