Context-Aware Outlier Exposure for Anomaly Segmentation in Road Scenes

Stefano Cardella Domenico Scalera Carlo Di Pantaleo Riccardo Mozzicato
Polytechnic of Turin
{s341749, s333304, s333106, s347492}@studenti.polito.it
https://github.com/AMLAnomaly/AnomalySegProject

Abstract

Anomaly segmentation in road scenes aims to identify unknown objects, which is really important for safety in autonomous driving. In this paper, we combine post-hoc methods with a context-aware outlier exposure (OE) fine-tuning strategy. We generate training data by pasting out-of-distribution objects into Cityscapes images. To avoid unrealistic situations that could mislead the model, we propose three context constraints: placing objects only on roads or sidewalks, scaling object dimensions according to perspective, and combining both strategies. Benchmarks on SMIYC [1], RoadAnomaly [2] and Fishyscapes [3] datasets show that our Outlier Exposure technique significantly improves anomaly detection, and that implementing realistic placement and scaling helps the model generalize better without needing additional real anomaly annotations.

1. Introduction

Semantic segmentation is a core task for autonomous driving, aiming to assign a label to every pixel in an image. Modern road-scene systems are usually trained in a closed-set setting, where the model is exposed only to a fixed set of semantic classes (e.g., road, sidewalk, vehicle, pedestrian) using specific datasets such as Cityscapes [4]. However, the real world is typically open-set: the scenes may contain unexpected objects and anomaly events never seen during training. This leads to anomaly segmentation, where the goal is to identify the pixels belonging to an unknown class, i.e. anomalies, while at the same time performing accurate semantic segmentation on known categories. Since anomalies and outliers can be very diverse across different scenes, anomaly segmentation in road scenes remains challenging. To detect potential outliers, different post-hoc methods are used to detect anomalies, based on uncertainty scores computed from the model's output. However, modern deep networks are often overconfident even when they are wrong [5], and this issue becomes particularly severe under distribution shift. Often, relying only on the model's confidence can be misleading: normal image regions (such as edges, shadows, or unusual textures) may be mistaken for anomalies, while truly out-of-distribution content may still receive high confidence. In this paper, we explore outlier exposure (OE) for anomaly segmentation [6]. Standard OE techniques help the model to become more robust by training it to give low scores to samples from an external dataset. However, a simple cut-and-paste objects technique can lead to unrealistic scenarios: pasted objects may appear in implausible positions or not respect the typical perspective of driving images. Since real anomalies often appear on the road and follow a perspective rule (small when far from the camera, larger when near), we propose a context-aware outlier exposure technique, placing outlier objects only in realistic locations and respecting perspective, with the aim of improving the model's ability to detect real anomalies.

Our main contributions are:

- We apply a fine-tuning strategy: first unfreezing just the segmentation head and queries, then progressively updating the last encoder layers.
- We evaluate model calibration using temperature scaling and fine-tune the model analyzing different hyperparameters. [5].
- We propose three OE implementations that improve realism in road scenes: (i) road-only placement, (ii) perspective-aware scaling, and (iii) a combined strategy. We evaluate these techniques using a unified hyperparameter configuration.

2. Related Work

2.1. From Pixel Classification to Mask Classification

Over the years, segmentation architectures have evolved significantly. Older approaches like ERFNet [7] treat semantic segmentation as a per-pixel classification task, where each pixel is treated independently. While ERFNet allows fast real-time computation, it often leads to inaccurate predictions because the architecture processes pixels independently and does not reason about whole objects.

MaskFormer [8] introduced a different paradigm: in-

stead of asking "what class is this pixel?", the model identifies complete object masks and classifies them. This approach naturally captures object shapes, leading to better and more coherent predictions. Mask2Former [9] improved this concept by introducing masked attention, allowing object queries to focus only on specific image regions. During training, a matching process assigns each real object mask to a specific query, so that every query learns to detect a particular object type.

More recent work demonstrates that Vision Transformers work very well in this paradigm. EoMT [10] shows that a simple ViT encoder, without particular architectural modifications, works effectively as a backbone and, when combined with a lightweight decoder, produces mask-based outputs. These masks are then combined to obtain pixel-level predictions by weighting class probabilities according to the mask confidence at each pixel. EoMT also uses mask annealing: masked attention is very important at the beginning of training to encourage query specialization, but it is gradually reduced. This strategy preserves segmentation quality and at the same time makes inference without masked attention simpler and faster. Additionally, DINOv2 [11] has shown that self-supervised pretraining on large amounts of unlabeled images can produce robust visual features that transfer very well to tasks like segmentation and anomaly detection.

2.2. Rejected-by-All Anomaly Detection

A key improvement is introduced by RbA [12], which takes advantage of the mask classification idea. The authors observed that object queries in models like Mask2Former act as independent detectors for specific classes. Based on this intuition, RbA defines an anomaly as a pixel that is "rejected by all" known classes. Instead of using standard softmax probabilities, the method combines the independent scores of all queries. This makes the method more reliable near object boundaries. While traditional methods often confuse uncertain edges with anomalies, RbA correctly identifies these regions as "known" because they receive multiple weak votes from different queries. In contrast, true outliers receive no significant votes at all.

2.3. Synthetic Anomalies through Outlier Exposure

Since it is very difficult to collect real examples of all possible anomalies, Outlier Exposure is used to create synthetic training data. Methods like RbA [12] use a cut-and-paste strategy, where objects from datasets such as COCO [13] are placed onto road scenes to fine-tune the model. However, a simple cut-and-paste approach can create unrealistic scenarios, where objects appear in implausible positions or do not respect perspective rules.

This creates the risk that the model learns to rely on these unrealistic artifacts instead of the anomalous objects themselves. For this reason, we explore a context-aware version of OE. Our idea is that if synthetic anomalies follow realistic physical rules, the model will learn stronger visual features that generalize better to real-world anomalies.

3. Method

In this section, we present our approach for anomaly segmentation using the EOMT architecture. We begin by detailing the inference pipeline.

3.1. Inference Pipeline and Logit Reconstruction

To handle high-resolution input images effectively while respecting to the memory constraints of the architecture, we employ a sliding window inference strategy followed by a spatial reconstruction mechanism. This pipeline transforms the raw model outputs into a cohesive semantic segmentation map.

3.1.1. Sliding Window Tessellation

Given an input image $I \in \mathbb{R}^{H_{\text{orig}} \times W_{\text{orig}} \times 3}$, the preprocessing stage first resizes the image to an intermediate scale (H', W') to align with the model's design specifications. To process high-resolution inputs under memory constraints, we decompose it into a set of N overlapping crops (patches), denoted as $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, where each crop x_k has a fixed spatial resolution.

3.1.2. Per-Pixel Logit Computation via EOMT

Each crop x_k is passed through the EOMT architecture [10]. Consistent with the Mask Transformer [8] [9] paradigm, the model predicts a set of mask logits $\mathbf{M} \in \mathbb{R}^{Q \times H \times W}$ and corresponding class logits $\mathbf{C} \in \mathbb{R}^{Q \times (K+1)}$ for the Q queries, where K represents the number of semantic categories.

To derive the dense per-pixel semantic logits $\mathbf{L}_{\text{crop}} \in \mathbb{R}^{K \times H \times W}$ for a specific crop, we perform a probability-mask product. Formally, this operation is implemented via Einstein summation (`einsum`) as follows:

$$\mathbf{L}_{\text{crop}} = \sum_{q=1}^{Q} \sigma(\mathbf{M}_q) \cdot \text{softmax}(\mathbf{C}_q)_{1\ldots K} \tag{1}$$

Here, $\sigma(\cdot)$ denotes the sigmoid activation function applied to the mask logits, and $\text{softmax}(\cdot)$ is applied to the class logits. The class logits are sliced $(1 \ldots K)$ to exclude the "no-object" (or void) category.

3.1.3. Spatial Reconstruction (Stitching)

The final stage involves reconstructing the segmentation map relative to the original image geometry. We initialize two tensors, $\mathbf{S}_{\text{sum}} \in \mathbb{R}^{K \times H' \times W'}$ and $\mathbf{S}_{\text{count}} \in \mathbb{R}^{H' \times W'}$, having the spatial dimensions of the scaled image. For each processed crop, its semantic score map $\mathbf{L}_{\text{crop}}$ is accumulated into $\mathbf{S}_{\text{sum}}$ at the corresponding spatial coordinates derived

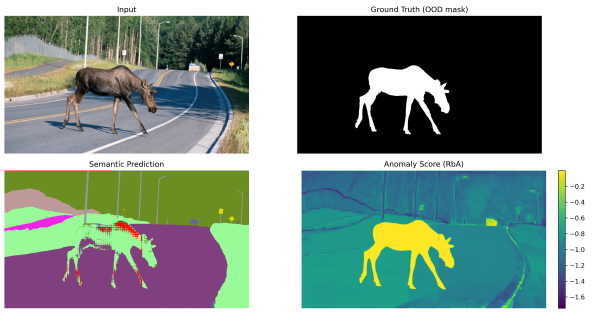

Figure 1. **Qualitative anomaly segmentation example on the Road Anomaly dataset.** Top-left: input image. Top-right: ground-truth OOD mask. Bottom-left: semantic prediction produced by the segmentation model. Bottom-right: anomaly score map computed using the RbA scoring function. High anomaly responses are concentrated on the unknown object while the in-distribution regions remain low.

from the origin indices, while $\mathbf{S}_{\text{count}}$ tracks how many times each spatial location has been covered by the sliding window.

The unified logit map is obtained by averaging the overlapping regions:

$$\mathbf{L}_{\text{unified}} = \frac{\mathbf{S}_{\text{sum}}}{\mathbf{S}_{\text{count}}} \tag{2}$$

Finally, $\mathbf{L}_{\text{unified}}$ is interpolated via bilinear upsampling to restore the original input resolution $H_{\text{orig}} \times W_{\text{orig}}$, yielding the final output tensor $\mathbf{P} \in \mathbb{R}^{1\times K\times H_{\text{orig}}\times W_{\text{orig}}}$.

.

3.2. Anomaly Scoring Functions

Once the dense pixel logits $\mathbf{P} \in \mathbb{R}^{1\times K\times H_{\text{orig}}\times W_{\text{orig}}}$ are reconstructed, the final step is to derive an anomaly score map $S \in \mathbb{R}^{H_{\text{orig}}\times W_{\text{orig}}}$. This map assigns a scalar value to each pixel, where higher values indicate a higher probability of the pixel being an outlier (anomaly).

For a given pixel i at spatial location (h, w), let $\mathbf{l}_i \in \mathbb{R}^K$ denote the vector of semantic logits extracted from $\mathbf{P}$.

3.2.1. Standard Post-Hoc Methods

We consider three established baselines that rely on the predictive uncertainty of the model:

Maximum Softmax Probability (MSP). [14] A classical baseline that interprets the complement of the maximum class probability as the anomaly score. The intuition is that outliers result in low-confidence predictions across all known classes:

$$S_{\text{MSP}}(i) = 1 - \max_k \left(\text{softmax}(\mathbf{l}_i)_k\right) \tag{3}$$

MaxLogit. [14] Unlike MSP, MaxLogit utilizes raw logits to avoid the calibration effects introduced by the softmax normalization. It operates on the assumption that in-distribution pixels tend to have higher maximum logit values than out-of-distribution ones:

$$S_{\text{MaxLogit}}(i) = -\max_k(\mathbf{l}_{i,k}) \tag{4}$$

MaxEntropy. [1] This method quantifies uncertainty via the Shannon entropy of the softmax distribution. High entropy implies a uniform distribution across classes, suggesting the model is uncertain and the pixel is likely an anomaly:

$$S_{\text{Entropy}}(i) = -\sum_{k=1}^{K} p_k \log(p_k), \quad \text{where } p = \text{softmax}(\mathbf{l}_i) \tag{5}$$

3.2.2. Rejected by All (RbA)

We adopt the *Rejected by All* (RbA) [12] scoring function, which leverages the one-vs-all behavior of object queries in Mask Transformers. In this view, a pixel is considered anomalous if it is rejected by all known classes.

Each logit is mapped to a bounded score using $\tanh$, and the anomaly score is computed as the negative sum over all classes:

$$S_{\text{RbA}}(i) = -\sum_{k=1}^{K} \tanh(\mathbf{l}_{i,k}) \tag{6}$$

As shown in Fig. 1, inliers produce a strong activation for at least one class, yielding low anomaly scores, while outliers exhibit low responses across all classes, resulting in higher scores. This aggregation reduces ambiguity compared to max-based scoring methods.

3.3. Calibration via Temperature Scaling

Deep neural networks are often poorly calibrated and tend to produce overconfident predictions [5]. To control the sharpness of the logit distributions used for anomaly scoring, we introduce Temperature Scaling as a logit rescaling strategy.

Given a logit vector $\mathbf{z}$, the scaled logits are defined as:

$$\tilde{z}_k = \frac{z_k}{T} \tag{7}$$

where $T > 0$ controls the distribution sharpness.

We evaluate two application points within the pipeline:

Variant A — Class-Logit Scaling. Temperature scaling is applied to the query classification logits $\mathbf{C}_q$ before softmax and before the logit reconstruction step:

$$\mathbf{C}_q^{\text{scaled}} = \frac{\mathbf{C}_q}{T}. \tag{8}$$

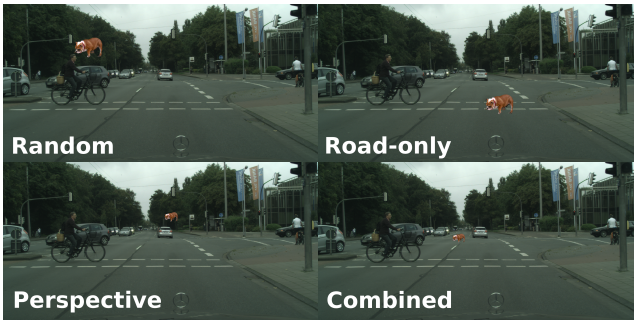

Figure 2. **Outlier Exposure placement strategies.** Examples of synthetic anomalies generated by pasting COCO objects onto Cityscapes images under different constraints. **Top-left:** Random placement. **Top-right:** Semantically constrained placement (Road/Sidewalk only). **Bottom-left:** Perspective-aware scaling based on vertical position. **Bottom-right:** Combined constraints (semantic + perspective).

Variant B — Pixel-Logit Scaling. Temperature scaling is applied to the reconstructed dense pixel logits $\mathbf{P}$ immediately before computing the anomaly scores:

$$\mathbf{P}^{\text{scaled}} = \frac{\mathbf{P}}{T}. \tag{9}$$

We test $T \in \{0.5, 0.75, 1.0, 1.1\}$, where $T = 1.0$ represents the unscaled baseline.

3.4. Outlier Exposure and Training Objective

To enable the model to distinguish between known semantic classes and anomalies, we introduce a supervision signal derived from outlier exposure (OE) [6]. We adopt a synthetic training strategy inspired by RbA [12], which involves pasting foreign objects from an external dataset onto in-distribution scenes.

3.4.1. Data Curation and Placement Strategies

We use the COCO dataset [13] as the source for anomalies. To ensure the semantic integrity of the outlier supervision, we implement a strict curation protocol starting with class filtering and single-object selection.

1. Curation. We filter COCO to exclude images containing any class that overlaps with Cityscapes [4] (e.g., *person*, *car*). From the remaining images, we select exactly one object instance by choosing the largest valid annotation (by area). To maximize clear object visibility, we enforce size constraints: the selected object must have area > 1000 pixels and cover $< 40\%$ of the image area.

2. Placement Variants. As shown in Fig. 2, standard copy-paste methods often place objects randomly, leading to physical impossibilities (e.g., a giant boat in the sky) that the model might exploit as shortcuts. To address this, we evaluate four variants obtained by enabling/disabling two constraints: semantic (road-only) and perspective scaling:

- **Variant A: Random Placement (Baseline).** The object is pasted at a random spatial location (v_x, v_y) with no constraints on the underlying background class or depth consistency.
- **Variant B: Semantically Constrained (Road/Sidewalk Only).** To align synthetic anomalies with the realistic distribution of road obstacles, we restrict placement to drivable surfaces. During the pasting loop, we verify the underlying semantic labels of the Cityscapes ground truth. A placement is considered valid only if at least 50% of the outlier object's pixel overlaps with pixels labeled as *Road* or *Sidewalk*.
- **Variant C: Perspective-Aware Scaling.** To prevent unrealistic scale cues (e.g., large objects far from the camera), we implement a geometric constraint using the vertical position v_y used as an indicator of depth. We apply a linear scaling factor s defined as:

$$s = 0.3 + 0.9 \cdot \left(\frac{v_y}{H}\right) \tag{10}$$

  where H is the image height. This maps objects from $0.3\times$ scale at the horizon ($v_y = 0$) to $1.2\times$ scale in the immediate foreground ($v_y = H$).
- **Variant D: Combined Constraints.** We apply both the semantic restriction (Road/Sidewalk overlap $\geq 50\%$) and the perspective-aware scaling simultaneously to generate the most physically plausible anomalies.

3.4.2. RbA-Based Outlier Loss

We align our training objective directly with the Rejected-by-All (RbA) inference mechanism. We compute a pixel-level anomaly score $S(p)$ defined as the Negative Logit Sum (NLS) of the logits.

First, dense per-pixel logits $\mathbf{l}_p \in \mathbb{R}^K$ are reconstructed by combining the softmax-normalized class predictions and sigmoid-normalized mask predictions via an Einstein summation. We then apply the hyperbolic tangent to map these scores to $[-1, 1]$ and aggregate them:

$$S(p) = -\sum_{k=1}^{K} \tanh(\mathbf{l}_{p,k}) \tag{11}$$

In this formulation, anomalies (rejected by all classes) tend to have negative logits across all dimensions, resulting in a high positive $S(p)$. Conversely, in-distribution pixels (accepted by at least one class) yield lower scores.

To optimize separation, we employ a Squared Hinge Loss. This objective enforces a margin between known classes and synthetic anomalies by penalizing inliers that exceed an upper threshold τ_{in} and outliers that fall below a lower threshold τ_{out}:

$$\mathcal{L}_{outlier} = \frac{1}{2}\Bigg[\frac{1}{|\mathcal{P}_{ID}|}\sum_{p\in\mathcal{P}_{ID}}\max(0, S(p)-\tau_{in})^2 + \frac{1}{|\mathcal{P}_{OOD}|}\sum_{p\in\mathcal{P}_{OOD}}\max(0,\tau_{out}-S(p))^2\Bigg] \tag{12}$$

The total training loss is a weighted sum of the standard segmentation terms and this outlier penalty:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{cls}}\mathcal{L}_{\text{CE}} + \lambda_{\text{OOD}}\mathcal{L}_{\text{outlier}} \tag{13}$$

These segmentation losses follow the standard Mask2Former training objective [9].

3.4.3. Fine-tuning Protocol

We fine-tune the model using a progressive unfreezing strategy to adapt the learned representations to anomaly detection. To systematically evaluate the effect of this choice, we combine each of the four placement variants (Random, Road-Only, Perspective, Combined) with four unfreezing configurations.

Specifically, for *each* of the four placement variants (Random, Road-Only, Perspective, Combined), we performed fine-tuning using four distinct unfreezing configurations:

1. **Head Only:** Unfreezing only the prediction head and the learnable object queries.
2. **Head + L Layers:** Unfreezing the head and queries plus the last L encoder layers, with $L \in \{1, 2, 3\}$.

This results in a total of 16 experimental settings. For all experiments, we implement a dual-metric early stopping mechanism: we primarily monitor the Area Under Precision-Recall Curve (AuPRC) using RbA method on the Road Anomaly dataset to maximize anomaly detection performance, while monitoring the Mean Intersection over Union (mIoU) on the Cityscapes validation set as a safeguard to prevent degradation of the in-distribution classes.

3.5. Hyperparameter Sensitivity Analysis

To validate the configuration adopted during fine-tuning, we conducted a targeted hyperparameter sensitivity analysis over the parameters governing the NLS-based outlier loss. Our training objective relies on the NLS score with $\tanh$ activation (Section 3.4), making the choice of score margins particularly important.

We explored a grid of 9 configurations (Table 1) to assess the influence of the outlier sampling probability (p_{out}), the outlier loss weight (λ_{OOD}), the margin thresholds (τ_{in}, τ_{out}), and the optimization schedule (learning rate and weight decay).

Exp	λ_{OOD}	p_{out}	τ_{in}	τ_{out}	LR	WD	Objective / Description
0 (Base)	1.0	0.20	-0.75	-0.10	$\mathbf{10^{-4}}$	0.05	-
1	2.0	0.20	-0.75	-0.10	10^{-4}	0.05	Increase OOD loss penalty
2	0.5	0.20	-0.75	-0.10	10^{-4}	0.05	Reduce OOD penalization
3	1.0	0.30	-0.75	-0.10	10^{-4}	0.05	Higher OOD sampling freq.
4	1.0	0.10	-0.75	-0.10	10^{-4}	0.05	ID-focused training
5	**1.0**	**0.20**	**-0.60**	**-0.20**	$\mathbf{10^{-4}}$	**0.05**	**Reduced margin**
6	1.0	0.20	-0.90	-0.05	10^{-4}	0.05	Larger margin
7	2.0	0.30	-0.80	-0.10	$5 \cdot 10^{-5}$	0.10	Strong OOD + Stabilized LR
8	1.5	0.20	-0.80	-0.05	$7.5 \cdot 10^{-5}$	0.10	Regularized & Balanced

Table 1. Hyperparameter sensitivity grid for the NLS-based outlier loss.

Outcome. The empirical evaluation on the Road Anomaly dataset revealed that **Exp 5 consistently provided the best trade-off** between anomaly detection performance (AUPRC) and preservation of closed-set segmentation quality (Cityscapes mIoU). In particular, reducing the margin between ID and OOD scores ($\tau_{in} = -0.60$, $\tau_{out} = -0.20$) proved more effective than the default parameters.

Consequently, we adopt the parameters of Exp 5 for all quantitative experiments presented in the main results section.

4. Experiments

4.1. Datasets

To validate our approach, we use distinct datasets for closed-set training, outlier exposure, and anomaly segmentation benchmarking.

4.1.1. Training and Outlier Source

- **In-Distribution (ID) Training: Cityscapes.** We train the segmentation model on the Cityscapes dataset [4], which serves as the source for in-distribution (ID) semantic classes. It contains 2,975 training images and 500 validation images of urban street scenes, annotated with 19 semantic categories considered as "known".
- **Outlier Exposure (OE): COCO.** To generate synthetic anomalies during fine-tuning, we use the COCO validation dataset [13] as a pool of anomalies. As detailed in Section 3.4, we curate a subset of objects that do not overlap with Cityscapes classes, ensuring that the model learns to identify general anomalies rather than specific categories.

4.1.2. Evaluation Benchmarks

We evaluate the performance of our method on five commonly used anomaly segmentation benchmarks. For all datasets, we use the **validation sets** to compute our metrics.

- **Segment Me If You Can (SMIYC).** This benchmark [1] is designed to test robustness against unexpected objects. We report results on the validation sets of two distinct tracks:

  - **Road Anomaly 2021 (RA-21) Chan et al. [1]:** Consists of images capturing rare and diverse anomalous objects (e.g., wild animals, helicopters, specialized vehicles) in various environments.
  - **Road Obstacle 2021 (RO-21) [1]:** Contains images focused on small obstacles located on the road surface.
- **Road Anomaly.** This dataset (often referred to as the original Road Anomaly) serves as a precursor to SMIYC. It comprises 60 validation images featuring a wide diversity of anomalous objects located in environments that differ significantly from the Cityscapes domain.
- **Fishyscapes.** We utilize the Fishyscapes benchmark [3], reporting performance on the validation sets of both tracks:
  - **Lost & Found (L&F) [3]:** A real-world dataset containing 100 validation images of small anomalous objects (*e.g.* crates, toys) on the road surface. It tests the model's ability to detect challenging real anomalies in a familiar environment.
  - **Static [3] :** A synthetic dataset containing 30 validation images constructed by compositing foreign objects onto Cityscapes images with careful blending to minimize visible copy–paste artifacts.

4.2. Evaluation Metrics

Following standard practice in anomaly segmentation literature, we quantify performance using pixel-level metrics that measure the separability between in-distribution and out-of-distribution pixels.

- **Area Under Precision-Recall Curve (AUPRC):** Also referred to as Average Precision (AP), this metric summarizes the precision-recall trade-off across all possible thresholds. It is widely considered the most robust metric for anomaly segmentation due to the severe class imbalance between inliers and outliers.
- **False Positive Rate at 95% True Positive Rate (FPR@95):** This metric measures the percentage of in-distribution pixels incorrectly classified as anomalies when the threshold is set to detect 95% of the true outlier pixels. A lower FPR@95 indicates a method that generates fewer false alarms while maintaining high safety standards.

5. Results

5.1. Baseline Comparison: Pixel-based vs. Mask-based Anomaly Detection

We compare the pixel-based ERFNet against the mask-based EoMT across five anomaly segmentation benchmarks (Table 2).

EoMT consistently outperforms ERFNet across all datasets, with improvements ranging from 5× on Fishyscapes L&F (3.30% → 18.36% AuPRC) to nearly 20× on SMIYC RO21 (4.63% → 93.85% AuPRC). The most substantial gains appear on RoadAnomaly, where EoMT achieves 73.89% AuPRC versus 15.58% for ERFNet, while reducing FPR95 from 73.25% to 14.88%. On SMIYC RO21, EoMT reaches near-perfect separation (FPR95 = 0.33%). The only dataset where both architectures struggle is Fishyscapes Static, where FPR95 remains above 40% for both models probably due to synthetic blending artifacts.

Regarding scoring functions, MaxLogit performs best for ERFNet, while MSP, MaxLogit, and MaxEntropy yield comparable results on EoMT, with MaxEntropy achieving a slight edge on anomaly tracks. The RbA score shows inconsistent behavior: while competitive on RoadAnomaly and SMIYC RO21, it fails catastrophically on SMIYC RA21 (FPR95 = 96.04%).

5.2. Temperature Scaling

We evaluated the effect of Temperature Scaling (T) on the EoMT architecture, applying the scaling to the final per-pixel logits and applying it to the class logits (Table 3).

Pixel vs. Class Logit Scaling

The adoption of temperature scaling to the final pixel logits resulted in minimal variations in performance metrics. AuPRC and FPR95 values remained virtually constant across the tested range $T \in [0.5, 1.1]$. Scaling the class logits instead induced slightly larger variations in the output distribution, though the closed-set performance remained stable (mIoU varying only between 81.66% and 81.79%). The overall impact on anomaly detection, however, is not that significant.

5.3. Outlier Exposure

We evaluate four configurations: Base OE (random placement), Road Only (placement restricted to road/sidewalk), Perspective (size scaling based on vertical position), and Road + Perspective (both constraints).

For each configuration, we experimented with unfreezing head + queries and progressively encoder blocks (0 to 3 blocks) and report the best-performing checkpoint. We use RoadAnomaly as validation set for checkpoint selection, so **results on RoadAnomaly are optimistically biased** and should be interpreted with caution.

Table 4 reports results for all scoring functions across the five benchmarks.

Comparison with Baseline

Most OE configurations substantially improve anomaly detection over the baseline without outlier supervision. On Fishyscapes Static, AuPRC increases from 61.75% to 97.86%, while FPR95 drops dramatically from 43.42% to below 0.19%. On Fishyscapes Lost & Found, AuPRC

Model	Method	SMIYC RA21		SMIYC RO21		FS L&F		FS Static		RoadAnomaly		mIoU
		AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	
ERFNet	MSP	29.10	62.55	2.71	65.23	1.75	50.60	7.47	41.84	12.42	82.58	72.17
	MaxLogit	**38.32**	**59.34**	**4.63**	**48.44**	**3.30**	**45.49**	**9.50**	**40.30**	**15.58**	**73.25**	
	MaxEntropy	30.97	62.66	3.04	65.91	2.58	50.16	8.84	41.55	12.67	82.75	
EoMT	MSP	68.90	**30.32**	93.78	0.35	16.47	13.99	59.42	**43.11**	71.77	15.61	81.68
	MaxLogit	68.23	31.58	**93.85**	0.35	16.46	13.78	**59.56**	46.56	71.08	15.20	
	MaxEntropy	**69.13**	30.53	93.84	**0.33**	**18.36**	13.81	58.33	43.42	**73.89**	**14.88**	
	RbA	63.43	96.04	93.53	0.44	16.62	**9.78**	61.75	73.66	70.53	15.26	

Table 2. Anomaly segmentation results comparing ERFNet and EoMT across five benchmarks. Best results per architecture are shown in **bold**.

Strategy	Temp.	SMIYC RA21		SMIYC RO21		FS L&F		FS Static		RoadAnomaly		mIoU
		AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	
Pixel Logits	$T = 0.5$	68.75	30.31	93.74	0.35	16.45	13.99	59.42	43.10	71.52	15.83	81.68
	$T = 0.75$	68.85	30.32	93.76	0.35	16.46	13.99	59.42	43.11	71.67	15.67	81.68
	$T = 1.0$ (def.)	68.90	30.32	93.78	0.35	16.47	13.99	59.42	43.11	71.77	15.61	81.68
	$T = 1.1$	68.92	30.32	93.78	0.35	16.47	13.99	59.42	43.11	71.81	**15.60**	81.68
Class Logits	$T = 0.5$	67.65	61.48	**94.12**	0.52	14.56	18.50	56.66	41.18	70.51	16.58	**81.79**
	$T = 0.75$	67.67	45.45	94.06	0.44	15.48	16.54	58.08	**40.94**	71.59	16.00	81.76
	$T = 1.0$ (def.)	68.90	30.32	93.78	0.35	16.47	13.99	59.42	43.11	**71.77**	15.61	81.68
	$T = 1.1$	**69.57**	**27.53**	93.59	**0.35**	**16.77**	**13.49**	**59.85**	47.41	71.75	15.73	81.66

Table 3. Comparison of Temperature Scaling applied to Pixel Logits vs. Class Logits (MSP scoring). Best results per column are shown in **bold**.

improves from 18.36% to 78.44%. On SMIYC RA21, the best OE configuration (Road + Perspective with RbA) achieves 86.18% AuPRC compared to 69.13% for the baseline. These gains confirm the effectiveness of outlier supervision for mask classification architectures.

As expected, **RbA becomes the best-performing scoring function after OE training**, since the NLS loss directly optimizes the RbA formulation. In contrast, without OE the standard scores (MSP, MaxLogit, MaxEntropy) outperform RbA, which suffers from catastrophic failure on SMIYC RA21 (FPR95 = 96.04%).

All OE configurations reduce mIoU by 2–4 percentage points (81.68% → 78–80%), reflecting the typical trade-off between anomaly detection and closed-set performance observed in prior work [12].

Comparison Between OE Strategies

Comparing the four OE strategies using the RbA scoring function:

Road + Perspective achieves the best results on most anomaly benchmarks: highest AuPRC on SMIYC RA21 (86.18%), Fishyscapes Static (97.86%), and SMIYC RO21 (97.44%), as well as the lowest FPR95 on SMIYC RO21 (0.04%), and Fishyscapes Static (0.19%). The combination of both constraints may produce more realistic training samples: objects are placed on drivable surfaces where real anomalies pose the greatest risk, and their size varies according to apparent depth, reflecting the natural appearance of obstacles at different distances. However, it shows higher FPR95 on RoadAnomaly (21.85%) compared to other strategies.

Base OE achieves the second-highest AuPRC on SMIYC RA21 (85.21%) and competitive performance across all benchmarks, suggesting that random placement provides good generalization.

Road Only achieves the best result on Fishyscapes Lost & Found (78.44% AuPRC) and on the validation set RoadAnomaly (94.46% AuPRC). This may be attributed to the characteristics of these datasets: according to the Fishyscapes benchmark [3], FS Lost & Found contains small real-world obstacles primarily located on the road surface in housing areas and parking lots. By restricting outlier placement to road and sidewalk regions during training, the Road Only strategy better matches this distribution.

Perspective Scaling alone shows the weakest results among OE strategies, though it still improves over the baseline on most datasets. However, on SMIYC RO21 it achieves lower performance than the baseline without OE (88.69% vs. 93.85% AuPRC).

Effect of Unfreezing Encoder Blocks

The best results for Base OE and Road Only were obtained with 3 unfrozen encoder blocks, while Perspective and Road + Perspective performed best with 2 unfrozen blocks. Unfreezing additional encoder layers improved performance across all configurations compared to fine-tuning

Method	Score	SMIYC RA21		SMIYC RO21		FS L&F		FS Static		RoadAnomaly*		mIoU
		AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	AuPRC↑	FPR95↓	
Baseline (no OE)	MSP	68.90	30.32	93.78	0.35	16.47	13.99	59.42	43.11	71.77	15.61	**81.68**
	MaxLogit	68.23	31.58	93.85	0.35	16.46	13.78	59.56	46.56	71.08	15.20	
	MaxEntropy	69.13	30.53	93.84	0.33	18.36	13.81	58.33	43.42	73.89	14.88	
	RbA	63.43	96.04	93.53	0.44	16.62	9.78	61.75	73.66	70.53	15.26	
Base OE (3 unfrozen)	MSP	72.49	81.62	96.86	0.13	60.54	46.10	92.54	3.37	88.99	9.65	79.78
	MaxLogit	77.90	29.95	97.28	0.12	62.41	13.62	93.14	2.97	89.58	9.23	
	MaxEntropy	73.51	76.50	97.04	0.11	58.83	42.72	92.97	2.79	88.69	9.29	
	RbA	85.21	32.30	96.57	0.04	64.65	27.63	97.02	0.69	91.80	4.69	
Road Only (3 unfrozen)	MSP	66.92	98.54	94.23	0.49	69.09	34.45	80.64	12.53	92.67	6.88	78.82
	MaxLogit	68.03	98.53	94.64	0.34	70.94	11.70	81.76	10.32	93.02	6.56	
	MaxEntropy	67.66	98.54	94.31	0.45	71.78	22.81	82.11	11.77	92.85	6.66	
	RbA	77.57	28.40	97.34	0.07	**78.44**	8.88	90.80	1.82	**94.46**	**3.19**	
Perspective (2 unfrozen)	MSP	68.07	93.31	87.30	1.02	53.34	7.87	91.62	2.13	87.32	15.20	78.54
	MaxLogit	71.33	73.86	87.64	0.99	54.59	**7.24**	92.15	2.02	87.85	14.59	
	MaxEntropy	68.63	92.78	87.48	1.01	56.12	7.33	92.47	1.91	87.58	14.35	
	RbA	73.71	48.62	88.69	0.74	59.43	7.45	95.18	0.60	89.61	6.66	
Road + Persp. (2 unfrozen)	MSP	75.31	36.12	96.27	0.20	63.31	10.20	93.01	1.36	86.80	18.33	78.06
	MaxLogit	79.82	**14.14**	96.57	0.17	66.43	9.21	93.98	1.26	87.70	17.64	
	MaxEntropy	76.53	31.27	96.31	0.19	65.52	9.89	94.40	1.26	87.90	17.78	
	RbA	**86.18**	16.29	**97.44**	**0.04**	71.23	15.55	**97.86**	**0.19**	88.99	21.85	

Table 4. Outlier Exposure results with all scoring functions. Metrics (except for mIoU) were computed with a 2× spatial downsampling to manage memory constraints. Results on RoadAnomaly* are biased due to validation-based checkpoint selection. Best result per dataset in **bold**.

only the prediction head and the queries.

6. Conclusion

In this study, we improved how unexpected objects (anomalies) are detected in road scenes using the mask classification paradigm. We showed that the EOMT architecture provides a stronger baseline than traditional pixel-level classifiers such as ERFNet.

Our main contribution came from creating more realistic training data. With context-aware Outlier Exposure (OE), instead of randomly pasting objects into images, we placed them on drivable areas and scaled them according to perspective. These choices help the model to generalize better to real-world anomalies. Experiments on the SMIYC, RoadAnomaly, and Fishyscapes datasets confirm that using realistic training examples improves outlier detection.

Although we noticed a small decrease in standard semantic segmentation accuracy (mIoU), the significant boost in safety makes this approach essential for self-driving cars. In future work, we plan to explore more advanced generative methods to make synthetic anomalies even closer to real ones.

References

[1] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Mathieu Salzmann, Pascal Fua, and Matthias Rottmann. SegmentMeIfYouCan: A benchmark for anomaly segmentation. *arXiv preprint arXiv:2104.14812*, 2021. 1, 3, 5, 6

[2] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis, 2019. 1

[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 (IJCV)*, 129:3119–3135, 2021. 1, 6, 7

[4] 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 (CVPR)*, pages 3213–3223, 2016. 1, 4, 5

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017. 1, 3

[6] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 4

[7] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 19(1): 263–272, 2018. 1

[8] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 2, 5

[10] Yuxiang Bai, Kun Mei, Tianjiao Yuan, Yixing Wang, Jieru Zhang, and Bingbing Ni. Your ViT is secretly an image segmentation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[11] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[12] Nazir Nayal, Misra Yavuz, Joao F Henriques, and Fatma Guney. RbA: Segmenting unknown regions rejected by all. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 4, 7

[13] 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 *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 2, 4, 5

[14] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning (ICML)*, 2022. 3