Task-Aware Anomaly Segmentation with Mask-Based Architectures

Adriano Giuliani Giacomo Lopez Davide Reverberi Vincenzo Sammito
Politecnico di Torino
Turin, Italy
{s345789, s336297, s345766, s346202}@studenti.polito.it

Abstract

Anomaly segmentation is a crucial task for deploying semantic segmentation models in open-world scenarios such as autonomous driving, where unexpected objects may appear at inference time. Existing approaches largely rely on post-hoc confidence-based methods applied to models trained solely for closed-set semantic segmentation, which limits their effectiveness in detecting out-of-distribution regions. In this work, we systematically evaluate pixel-based and mask-based architectures for anomaly segmentation on road-scene benchmarks using common post-hoc methods. Our analysis confirms that mask-based architectures inherently provide superior representations for anomaly detection compared to pixel-based baselines. Building on these findings, we propose a task-aware extension that incorporates a dedicated Feature-Enriched Anomaly Head supervised by a synthetic data generation pipeline. Experimental results demonstrate that while statistical uncertainty alone is insufficient, integrating visual features allows our model to significantly outperform baselines on obstacle detection benchmarks. However, we also highlight the generalization challenges of synthetic training when facing diverse semantic anomalies, offering critical insights into the trade-offs between artifact detection and semantic understanding. The code is publicly available at: https://github.com/Giacomo-FMJ/MaskArchitectureAnomaly

1. Introduction

Deep neural networks for semantic segmentation have achieved remarkable performance in closed-world settings, where all object categories are known at training time. However, when deployed in real-world environments such as autonomous driving, these models inevitably encounter unknown or out-of-distribution (OoD) objects. In such scenarios, conventional semantic segmentation models tend to produce overconfident and incorrect predictions, posing severe safety risks.

Anomaly segmentation addresses this problem by identifying image regions that do not belong to the training distribution, without requiring explicit semantic labels for unknown objects. Most existing approaches tackle this task using post-hoc techniques, such as confidence thresholding or entropy-based scoring, applied to pretrained semantic segmentation models. While effective to some extent, these methods operate on representations that were not explicitly optimized for anomaly detection.

Recent advances in mask-based architectures, such as MaskFormer and its successors, offer a new perspective on anomaly segmentation. By predicting object-centric masks rather than per-pixel class labels, these models provide richer intermediate representations that can be exploited for anomaly detection. In particular, the EoMT architecture leverages pretrained vision transformers to produce expressive embeddings that are well-suited for open-world tasks.

In this work, we first provide a comprehensive evaluation of pixel-based and mask-based segmentation models using standard post-hoc anomaly detection techniques. We then move beyond post-hoc analysis and propose a task-aware architectural extension for anomaly segmentation. Specifically, our work extends the core EoMT architecture through the integration of a specialized anomaly head, that enhances internal uncertainty and visual context, complementary to the semantic head. Unlike traditional methods that rely on post-hoc metrics, our approach allows for end-to-end anomaly scoring during inference. Furthermore, we developed a synthetic data generation pipeline to train the anomaly head, effectively injecting OoD artifacts into the standard training image samples.

2. Related Work

Anomaly segmentation has emerged as a critical extension of semantic segmentation for open-world scenarios, particularly in safety-critical applications such as autonomous driving. Several benchmarks have been introduced to formalize this task and provide standardized evaluation protocols. Fishyscapes [1] focuses on detecting anomalous objects in urban driving scenes by evaluating models trained on closed-set semantic segmentation, while Seg-

mentMeIfYouCan [3] extends this setting by introducing multiple datasets and anomaly types, highlighting the limitations of conventional segmentation models when exposed to out-of-distribution inputs.

Early approaches to anomaly segmentation predominantly rely on post-hoc confidence-based methods applied to pretrained semantic segmentation models. Hendrycks and Gimpel [7] propose Maximum Softmax Probability (MSP) as a simple baseline for detecting out-of-distribution samples, which has since been widely adopted and extended to dense prediction tasks. Subsequent work explores alternative uncertainty measures, such as max-logit and entropy-based scoring, as well as large-scale evaluations of out-of-distribution detection methods in realistic settings [9]. While effective to some extent, these methods operate on representations that were not explicitly optimized for anomaly detection.

Recent works have shown that incorporating spatial structure can improve anomaly segmentation performance. In particular, Region-based Anomaly (RbA) scoring [2] leverages region-level consistency to better identify unknown areas by rejecting regions that are not confidently assigned to any known class. Such methods benefit from architectures that provide object- or region-level predictions rather than purely pixel-wise outputs.

Mask-based architectures represent a significant shift in semantic segmentation by reformulating the task as a set prediction problem. MaskFormer [4] demonstrates that predicting a set of masks and associated class labels can outperform traditional pixel-based approaches, while Mask2Former [5] further improves this paradigm through masked attention mechanisms. These architectures unify semantic, instance, and panoptic segmentation and naturally provide region-level representations that are well suited for anomaly segmentation.

Building on these ideas, recent work has explored the use of pretrained vision transformers for segmentation. EoMT [11] shows that features learned through self-supervised pretraining, such as DINOv2, can be effectively adapted for mask-based segmentation tasks. The rich and object-centric representations produced by such models make them particularly attractive for open-world perception problems, including anomaly segmentation, where unknown objects may emerge as distinct masks or outlier embeddings.

3. Method

In this section, we describe the methodology adopted for anomaly segmentation in road scenes. We first formalize the anomaly segmentation problem and define the output space considered in our work. We then briefly describe the baseline architectures and post-hoc anomaly scoring methods used for evaluation, before introducing our proposed task-aware architectural extension.

3.1. Problem Formulation

Let $x \in \mathbb{R}^{H \times W \times 3}$ denote an input RGB image depicting a road scene, and let $\mathcal{C} = \{1, \ldots, K\}$ be the set of semantic classes observed during training. In standard semantic segmentation, a model is trained to assign each pixel p in x a label $y_p \in \mathcal{C}$. However, in open-world settings, the input image may contain regions that do not belong to any class in $\mathcal{C}$.

The goal of anomaly segmentation is to identify such out-of-distribution regions at pixel level. Formally, anomaly segmentation can be cast as a binary classification problem, where each pixel is assigned either an in-distribution label or an anomaly label. In practice, anomaly segmentation is often evaluated by computing an anomaly score map $A \in \mathbb{R}^{H \times W}$, where higher values indicate a higher likelihood of a pixel being anomalous.

In this work, we consider both post-hoc and task-aware formulations of anomaly segmentation. In the post-hoc setting, anomaly scores are derived from the outputs of a semantic segmentation model trained on $\mathcal{C}$, without modifying the model architecture. In contrast, in the task-aware setting, the model is explicitly designed to produce predictions that account for anomalous regions through an integrated architectural extension. Specifically, we explore an augmented output space that maps a combination of stochastic uncertainty signals and visual features into a dense anomaly score. This formulation allows the model to directly predict anomalous regions at the pixel level during inference, without relying on heuristic post-hoc thresholding of confidence scores.

3.2. Baseline Architectures and Post-hoc Anomaly Scoring

We evaluate both pixel-based and mask-based segmentation architectures as baselines for anomaly segmentation. As a pixel-based model, we use ERFNet, a lightweight convolutional neural network designed for real-time semantic segmentation. ERFNet produces dense per-pixel class logits over the closed set of semantic classes $\mathcal{C}$.

As a mask-based model, we adopt the EoMT architecture, which reformulates semantic segmentation as a set prediction problem. Given an input image, EoMT predicts a set of mask embeddings along with corresponding class logits, which are subsequently projected back to pixel space. This object-centric formulation provides region-level representations that can be exploited for anomaly segmentation.

For both architectures, anomaly segmentation is first addressed in a post-hoc manner by deriving anomaly scores from the model outputs without modifying the training procedure. Specifically, we consider the following post-hoc scoring methods. Maximum Softmax Probability (MSP)

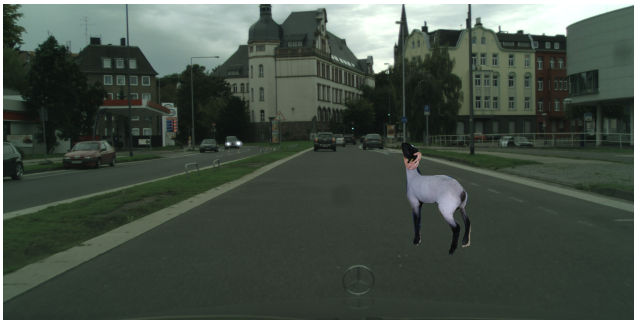

Figure 1. Example of a synthetic training image used for the proposed extension. An anomalous object is inserted into a real road scene to simulate out-of-distribution regions.

computes the anomaly score of each pixel as the inverse of the maximum predicted class probability. Max Logit uses the maximum pre-softmax activation as a confidence measure, while Max Entropy measures prediction uncertainty via the entropy of the softmax distribution. For mask-based predictions, we additionally apply Region-based Anomaly (RbA) scoring, which aggregates anomaly scores at the region level to enforce spatial consistency.

These post-hoc methods serve as strong baselines and allow for a direct comparison between pixel-based and mask-based architectures. However, since anomaly scores are derived from representations optimized solely for closed-set semantic segmentation, these approaches are inherently limited in their ability to capture out-of-distribution regions. With the aim of overcoming these issues, the next section discusses the creation of a synthetic dataset, which is a fundamental starting point for achieving our proposed extension.

3.3. Syntethic Dataset

To support the experimental evaluation, a synthetic anomaly segmentation dataset was constructed by combining real urban driving scenes with out-of-distribution object instances. Images from the Cityscapes dataset were used as in-distribution backgrounds, while object instances extracted from the COCO dataset were inserted to simulate anomalous events that are unlikely to occur in normal driving conditions. The dataset generation process is fully configurable and deterministic, allowing controlled variation in the number, scale, and placement of anomalies across different dataset splits.

The synthesis pipeline enforces semantic and geometric constraints derived from Cityscapes annotations, ensuring that anomalies are placed in meaningful regions of the scene (e.g., roads or sidewalks). Both pixel-level anomaly maps and instance-level anomaly masks are generated, enabling evaluation of methods based on dense predictions as well as object-level reasoning. Although the visual realism of the generated anomalies is limited by the quality of COCO instance segmentations and the absence of physical scene modeling, the dataset serves as a practical and reproducible testbed for studying anomaly detection behavior under controlled conditions. The dataset is publicly available online and can be accessed at [6].

3.4. Proposed Extension

As highlighted in the baseline analysis, the identification of anomalies is currently treated as an external evaluation process performed exclusively at inference time, rather than being intrinsic to the model architecture. Our proposal integrates this step directly into the network, aiming to create a self-contained model capable of predicting anomaly scores explicitly, without relying on post-processing of class probabilities. The introduction of the anomaly head is therefore designed to render the model self-sufficient for anomaly detection tasks.

However, designing such an internal mechanism raises a fundamental question regarding the source of the anomaly information within the network. Consequently, the rationale behind this extension is to investigate the intrinsic nature of the anomaly signal. Our goal is to first determine if the model's internal uncertainty, expressed through a **Statistical Uncertainty-based Head**, is sufficient to differentiate between known and unknown semantic states. We then evolve this concept into a **Feature-Enriched Anomaly Head**, which integrates visual evidence from the ViT backbone to contextualize these statistical signals.

Preliminary Experiments. Our initial approach attempted to detect anomalies by treating the object queries, directly emerged from the ViT backbone [11], in parallel with the pretrained class head. We hypothesized that the current trained model could be flexibly used as-is also to recognize the anomalous. We further extended this by combining the queries with MSP and Entropy to inject predictive information. However, these experiments failed, as the model was not allowing the direct segmentation of OOD without unfreezing other parts of the network, risking noticeable performances drops for the primary task. These limitations, motivated a shift in our paradigm: moving away from the analysis of discrete query embeddings toward an architectural framework that first decodes the model's internal uncertainty and subsequently anchors these diagnostic signals to the backbone's latent visual representations.

3.4.1. Statistical Uncertainty-based Anomaly Head

EoMT [11] adapts a standard ViT for segmentation by jointly processing object queries and visual tokens to generate mask embeddings and class distributions. Although trained for closed-set accuracy, the model exhibits distinct uncertainty patterns when encountering Out-of-Distribution (OoD) inputs. Our proposed extension leverages these

frozen output signals as diagnostic features for explicitly detecting anomalies.

The Pixel Anomaly Head processes a per-pixel descriptor $\mathbf{s}_{i,j} \in \mathbb{R}^4$ derived from the dense semantic map $\mathbf{M}$, which fuses mask probabilities with class distributions. The descriptor comprises four uncertainty metrics:

- **Max Probability:** The peak confidence score across classes ($\max(\mathbf{M}_{i,j})$), representing the model's highest certainty for the pixel.
- **Predictive Entropy:** Computed as $-\sum p \log(p)$, quantifying the ambiguity in class assignment.
- **Energy Proxy:** The sum of semantic probabilities, indicating the total activation strength or evidence for known classes.
- **MSP Baseline:** The complement of the maximum probability ($1 - \max\_prob$), serving as a direct measure of non-confidence.

The anomaly head is implemented as a pixel-wise Multi-Layer Perceptron (MLP) f_ϕ. It maps the 4-dimensional statistical descriptor to a single scalar output:

$$\mathcal{A}i, j = f\phi(\mathbf{s}i, j) \in \mathbb{R}^1 \tag{1}$$

where $\mathcal{A}i, j$ denotes the unnormalized anomaly logit. By approximating a non-linear decision boundary within this uncertainty space, the model learns to distinguish the statistical profile of anomalous regions from inherent in-distribution ambiguity.

3.4.2. Enriching with Visual Features

The statistical approach refines raw uncertainty metrics into an anomaly score, but operates without direct access to the visual input. This "semantic blindness" limits the model's ability to distinguish between noise-induced entropy and genuine anomalies. To resolve this, we enrich the statistical profile with visual evidence extracted from the frozen backbone.Following the EoMT processing logic, we extract patch tokens from the `ViT-Base-DINOv2` encoder, discarding CLS and register prefixes. The sequence is reshaped into a feature grid $\mathbf{F}_{grid} \in \mathbb{R}^{C \times H_g \times W_g}$. To fuse this high-dimensional representation ($C = 768$) efficiently with the statistical descriptors, we apply a bottleneck projection:

1. **Compression:** A 1×1 convolution projects the visual features from $\mathbb{R}^{768} \to \mathbb{R}^{32}$, preserving salient descriptors while reducing dimensionality.
2. **Alignment:** The compressed features are bilinearly interpolated from the grid resolution to match the spatial resolution of the statistical map $\mathbf{s}_{i,j}$.
3. **Fusion:** The resulting 32-channel visual descriptor is concatenated pixel-wise with the 4-channel statistical descriptor.

The resulting 36-dimensional vector serves as input to the MLP, allowing the model to ground its uncertainty estimates in the local visual context.

4. Experiments

In this section, we present the experimental evaluation of the proposed anomaly segmentation framework. We first establish a performance baseline using standard post-hoc scoring methods applied to pretrained semantic segmentation models. We then evaluate our proposed architectural extensions, analyzing the impact of training the specialized anomaly heads on the synthetic dataset described in Section 3.3.

4.1. Datasets

Evaluation Datasets. We evaluate anomaly segmentation performance on multiple benchmark datasets commonly used in the literature [3] [1]. In particular, we consider the SegmentMeIfYouCan benchmark, including the Road Anomaly and Road Obstacle splits (SMIYC RA-21 and SMIYC RO-21), as well as the Fishyscapes Lost and Found (FS L&F) and Fishyscapes Static (FS Static) datasets. These datasets contain real-world urban driving scenes with anomalous objects that are not part of the training distribution of the segmentation models.

Fine-tuning Dataset. To train the proposed extended architecture, and specifically to optimize the anomaly head, we leverage the synthetic dataset described in Section 3.3. The adoption of a synthetic source is required by the significant lack of real-world datasets containing explicit anomaly annotations suitable for training. In the current literature, datasets with such annotations (as those previously cited) are reserved exclusively for testing and benchmarking purposes, thus requiring the generation of synthetic samples to supervise the learning process. [6]

4.2. Evaluation Metrics

We follow standard evaluation protocols for anomaly segmentation, relying on two primary metrics: the Area Under the Precision-Recall Curve (AU-PRC) and the False Positive Rate at 95% True Positive Rate (FPR@95). These metrics robustly quantify the ability to distinguish anomalous pixels from in-distribution ones. While AU-PRC and FPR@95 focus on detection performance, mIoU provides insight into the overall segmentation quality of the underlying model.

4.3. Baseline Models and Post-hoc Methods

We compare a pixel-based baseline, ERFNet, against the mask-based EoMT architecture. Both models were pretrained on the Cityscapes dataset in a standard closed-set setting, observing no anomalous samples during training. To establish a reference for performance, we evaluate anomaly detection using the post-hoc scoring methods defined in Section 3.2: MSP, Max Logit, and Max

Model	Method	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
		AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95
ERFNet	MSP	29.095	62.549	2.709	65.223	1.749	50.594	7.472	41.836	12.423	82.575
	MaxLogit	**38.319**	**59.337**	**4.626**	**48.443**	**3.301**	**45.495**	**9.498**	**40.300**	**15.581**	**73.247**
	MaxEntropy	30.969	62.658	3.044	65.912	2.583	50.163	8.836	41.545	12.668	82.748

Table 1. Post-hoc anomaly segmentation results using a pretrained ERFNet model. The model achieves a mean IoU of 72.20% on the Cityscapes validation set. Higher AU-PRC and lower FPR@95 indicate better performance.

Model	Method	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
		AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95
EoMT	MSP	74.39	33.72	90.20	**0.55**	25.32	14.91	29.93	34.18	75.67	19.35
	MaxLogit	73.99	34.31	90.14	0.58	25.35	14.71	30.12	37.65	75.15	19.15
	MaxEntropy	**77.71**	**33.59**	**90.11**	0.67	**29.02**	14.73	28.93	**34.08**	**78.11**	**18.84**
	RbA	69.99	80.06	87.25	99.95	25.88	**8.74**	**32.79**	84.42	74.62	20.43

Table 2. Post-hoc anomaly segmentation results using a pretrained EoMT model evaluated at a fixed temperature ($T = 1$). The model achieves a mean IoU of 81.68% on the Cityscapes validation set. Higher AU-PRC and lower FPR@95 indicate better performance. RbA is applicable only to mask-based predictions.

Entropy. For EoMT, we additionally evaluate the mask-specific Region-based Anomaly (RbA) score to assess the benefit of object-level aggregation.

4.4. ERFNet Baseline Results

Table 1 presents the baseline results for ERFNet. Across all benchmarks, Max Logit consistently outperforms MSP and entropy-based scoring, achieving superior AU-PRC and FPR@95 values. This performance advantage likely stems from the unnormalized nature of logits, which preserves the magnitude of activations that softmax normalization typically compresses, allowing for better separation between in-distribution and anomalous pixels.

4.5. EoMT Baseline Results

Table 2 presents the results for the pretrained EoMT model ($T = 1$), which achieves an 81.68% mIoU on Cityscapes, significantly surpassing the pixel-based baseline. In contrast to ERFNet, Entropy-based scoring generally yields the highest AU-PRC, particularly on the Road Anomaly and Road Obstacle benchmarks. Max Logit and MSP perform comparably, suggesting that the advantages of unnormalized logits are less pronounced for this mask-based architecture.Region-based Anomaly (RbA) scoring exhibits high variance across datasets. While it effectively reduces false positives on Fishyscapes Lost & Found, it severely degrades performance on Road Obstacle. This suggests that post-hoc aggregation is sensitive to specific anomaly characteristics and lacks consistent robustness across diverse scenarios.

4.6. Temperature Scaling Analysis

We analyze the impact of temperature scaling on MSP-based anomaly segmentation for the pretrained EoMT model. Given per-pixel logits $\mathbf{z}$, temperature scaling rescales them as $\mathbf{z}' = \mathbf{z}/T$ before applying the softmax, and MSP anomaly scores are computed as $A(p) = 1 - \max_k \mathrm{softmax}(\mathbf{z}'_p)_k$. This transformation modifies the confidence calibration of the model without changing the predicted class ranking, and can therefore affect confidence-based anomaly detection.

We evaluate MSP at four representative temperature values $T \in \{0.5, 0.75, 1.0, 1.1\}$ and additionally report MSP at the best temperature, denoted as MSP (best T). The best temperature is selected from a broader sweep as the value maximizing the average AU-PRC across the evaluated datasets.[1]

As shown in Table 3, temperature scaling produces only marginal variations in performance. In particular, AU-PRC changes are small across all datasets, with the largest gains observed on Road Anomaly when using the best temperature. FPR@95 remains almost unchanged across temperatures, indicating that rescaling logits primarily affects confidence calibration rather than improving the separability between anomalous and in-distribution pixels. Overall, this analysis suggests that calibration alone is insufficient to substantially improve post-hoc anomaly segmentation, motivating task-aware approaches that explicitly model anomalies in the prediction space.

4.7. Anomaly Head Training

Standard segmentation backbones, optimized for closed-world classification, often fail to produce calibrated confidence scores for unseen objects. To bridge this gap, we introduce a dedicated Anomaly Head that transforms the backbone's internal representations into an explicit discrim-

[1]Selecting T to maximize AU-PRC is used here as a simple model-agnostic criterion to summarize the effect of calibration.

Method	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
	AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95	AU-PRC	FPR@95
MSP (T=0.5)	74.18	33.72	90.20	0.55	25.31	14.92	29.90	34.18	75.49	19.41
MSP (T=0.75)	74.29	33.72	90.20	0.55	25.31	14.91	29.92	34.18	75.60	19.37
MSP (T=1.0)	74.39	33.72	90.20	0.55	25.32	14.91	29.93	34.18	75.67	19.35
MSP (T=1.1)	74.39	33.72	90.20	0.55	25.32	14.91	29.93	34.18	75.70	19.35
MSP (best T)	**74.83**	33.73	90.20	0.55	**25.45**	14.91	29.97	34.2	**76.28**	**19.31**

Table 3. Effect of temperature scaling on MSP-based anomaly segmentation for the pretrained EoMT model. Results are reported for selected temperature values. The best temperature is $T = 8.0$, selected from a broader sweep as the value maximizing the average AU-PRC across datasets.

inative score for Out-of-Distribution (OoD) regions. By keeping the EoMT backbone and primary segmentation heads frozen, we ensure the model retains its original semantic proficiency while the trainable MLP learns to decode the uncertainty signals of the frozen components. The following sections detail the training protocol for the statistical and feature-enriched configurations.

4.8. Statistical Uncertainty-based Anomaly Head

In this first experimental phase, we evaluated the model's ability to learn the concept of anomaly based exclusively on the uncertainty metrics extracted from the frozen EoMT backbone, as described in section 3.4.1. The fine-tuning focused solely on the parameters of the *PixelAnomalyHead* (Stats-Only configuration), while keeping the encoder and the original semantic projections freezed.

Training and Loss Function. The fine-tuning process was formulated as a pixel-wise binary classification task. During each iteration, the anomaly logits produced by the MLP were compared against a binary ground truth mask, where anomalous pixels are labeled as the positive class. To address the severe class imbalance inherent in anomaly detection (where Out-of-Distribution (OoD) pixels in the majority of real cases are only a small fraction of the image) we employed a *Binary Cross Entropy (BCE) with Logits* loss function. Specifically, we applied a positive class weight of $w = 10.0$ to penalize misclassifications of anomalous regions more heavily. We adopted his strategy based on the hypothesis that anomalies are the minority inside the image.Training was conducted using learning rate of 1×10^{-4} and 5 epochs, with seed equal to 0.

4.8.1. Stats-Only Anomaly Head Results

The results reported in Table 4 highlight a critical behavior that requires a detailed analysis. Although the AuPRC remains at competitive levels (e.g., 77.81 on RoadAnomaly) compared with the baseline, suggesting that the anomaly head successfully learns to rank anomalous pixels higher than In-Distribution ones, the model suffers from a significant calibration collapse in terms of False Positive Rate.

The catastrophic FPR95 of 95.52% on **SMIYC RA-21**

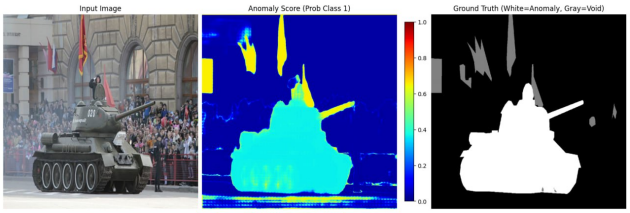

Figure 2. Evidence of training-set bias causing calibration collapse on rare anomalies (Uncertanty score-based head, SMIYC RA-21)

can be attributed to both the necessity of some better hyperparameter tuning and to the specific composition of the training samples. One example is clearly illustrated in Figure 2, where a prominent anomaly (the tank) is assigned an unexpectedly low anomaly score. These results suggest that the limitations are intrinsically tied to the dataset's specific samples and the "semantic silence" of the frozen backbone, which the MLP cannot resolve through uncertainty metrics alone.

Despite these localized failures, the results on other datasets, while slightly below the baseline, confirm that the supervised mapping of uncertainty signals is a viable direction. This performance gap highlights the need for a more descriptive input space, leading us to experiment a more sophisticated configuration: the Feature Enriched Anomaly Head.

4.9. Feature enriched anomaly head

The training protocol largely mirrored the statistical-only configuration, with specific adjustments to accommodate the increased dimensionality of the input. We found that a reduced learning rate of 1×10^{-6} and a short training duration of just 3 epochs yielded the optimal balance between convergence and generalization. This aggressive regularization was crucial to prevent the high-capacity visual head from overfitting to the specific artifacts of the synthetic training dataset. Since the architectural innovations are detailed in Section 3.4.2, we focus here on the experimental outcomes reported in Table 4.

Method	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
	AP ↑	FPR ↓	AP ↑	FPR ↓	AP ↑	FPR ↓	AP ↑	FPR ↓	AP ↑	FPR ↓
Baselines: EoMT Frozen Head (Statistical)										
EoMT (MSP T=1.0)	74.39	33.72	90.20	0.55	25.32	14.91	29.93	34.18	75.67	19.35
EoMT (MaxLogit)	73.99	34.31	90.14	0.58	25.35	14.71	30.12	37.65	75.15	19.15
EoMT (Entropy)	**77.71**	**33.59**	90.11	0.67	29.02	14.73	28.93	34.08	**78.11**	18.84
EoMT (RbA)	69.99	80.06	87.25	99.95	25.88	8.74	32.79	84.42	74.62	20.43
Ours: Trainable Anomaly Head										
Stats-Only Head (V1)	72.87	95.52	88.65	37.91	27.89	12.21	54.83	37.96	77.81	18.63
Feature-Enriched Head (V2)	74.22	37.90	**95.72**	**0.14**	**75.37**	**1.54**	**96.66**	**0.30**	72.15	**16.47**

Table 4. Comprehensive comparison of anomaly segmentation methods. AP denotes Area Under Precision-Recall Curve (%), and FPR denotes False Positive Rate at 95% TPR. The **Stats-Only Head** (V1) is a lightweight MLP trained on frozen uncertainty metrics. The **Feature-Enriched Head** (V2) adds visual features from the DINOv2 backbone, yielding massive gains on obstacle benchmarks (FS L&F, FS Static, RO-21) while maintaining competitive performance on general anomalies.

Performance Overview The results reveal a stark dichotomy in model performance. On the obstacle-centric benchmarks, **FS Lost & Found**, **FS Static**, and **SMIYC RO-21**, the Feature-Enriched head achieves transformative improvements, reducing the False Positive Rate by orders of magnitude compared to the statistical baseline (e.g., from 37.91% to 0.14% on RO-21). This confirms that integrating DINOv2 visual features effectively resolves the 'semantic blindness' of the statistical approach, enabling the model to identify obstacles through distinct visual and textural cues. Also the **FS Lost & Found** contains mostly very small anomalies that were correctly identified, showing that posing the loss $w = 10.0$ was beneficial in this context.

Domain Shift Analysis Conversely, the method yielded lower ranking performance (AU-PRC) on the remaining datasets (**RoadAnomaly** and **SMIYC RA-21**). We attribute this gap to the synthetic domain shift inherent in the training process. Since the model was supervised using synthetically generated anomalies, it excels at detecting distinct "foreign objects" (as seen in Fishyscapes). However, it struggles to generalize to the diverse, semantic anomalies found in RoadAnomaly, where the anomalous object (e.g., a strange vehicle) often shares consistent lighting and texture with the environment.

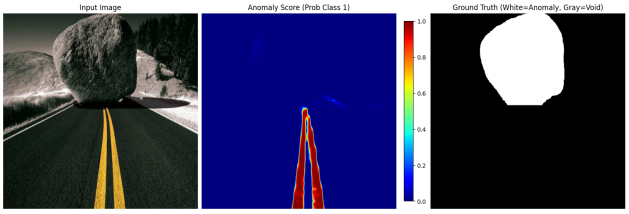

Figure 4. Evidence of a feature-bias from the from the training distribution (Feature-enriched head, SMIYC RA-21)

Failure Modes Qualitative analysis reveals a specific vulnerability to "visual mimicry." As illustrated in Figure 4, objects that texturally blend with the background can trick the feature-enriched head. In these scenarios, the strong visual prior (identifying the texture as "background-like") overrides the uncertainty signal, leading to false negatives that the purely statistical baseline would have correctly flagged.

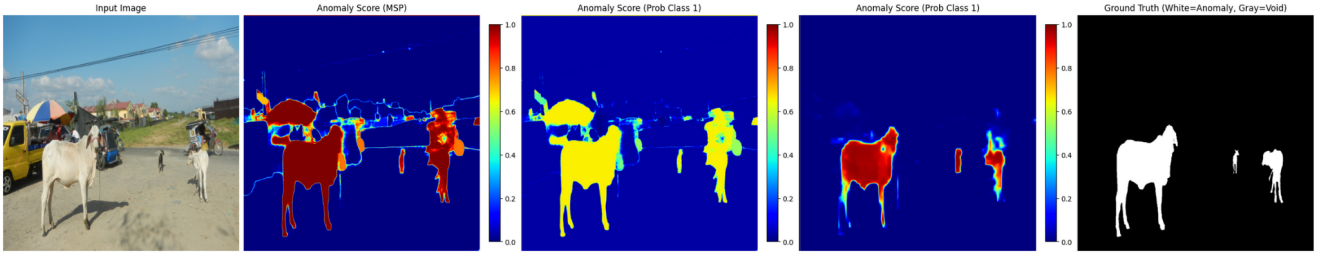

Figure 3. Qualitative comparison of results. From left to right: Input (SMIYC RA-21), baseline (MSP T=1.0), stats-only, stats+features, Ground truth

4.10. Critical Discussions

The experimental results from our OoD training regimen underline the persistent challenges intrinsic to open-world segmentation. A fundamental bottleneck identified in this study is the ontological ambiguity of the term "anomaly" and, the dissonance between the definition of anomalies during training (often restricted to synthetic artifacts or specific subsets) and their manifestation in testing benchmarks and real-world scenarios as well [3, 6]. Our findings suggest that a data-centric perspective, investing in high-fidelity datasets or effort in syntetic generation, could lead to improve generalization across diverse settings [8, 9].

Despite these data-level limitations, our architectural analysis confirms the robustness of the EoMT framework. We validated that frozen Vision Transformer backbones, particularly those leveraging self-supervised pre-training, inherently encode semantically rich features that are adaptable to anomaly detection without requiring elaborate task-specific modules [10, 11].

We also believe that both our synthetic dataset generation and our anomaly head architecture retain high potential given the obtained results 3 and that there is a noticeable margin to be gained with more exploration of the tuning parameters, especially for the dataset generation.

5. Conclusions

In this work, we presented a comprehensive study on anomaly segmentation for autonomous driving, transitioning from standard post-hoc analyses to a task-aware architectural extension. We demonstrated that mask-based architectures, specifically EoMT, provide a superior foundation for anomaly detection compared to traditional pixel-based models, offering richer object-centric representations. Our primary contribution is the proposal of a specialized Pixel Anomaly Head that operates in parallel with the semantic backbone. Two key hypotheses were validated experimentally. First, we showed that relying solely on statistical uncertainty retrieved from the model class confidence is insufficient for precise segmentation, as it leads to calibration collapse. Second, and most importantly, we demonstrated that grounding uncertainty with visual features is the key to robust detection. Our Feature-Enriched Anomaly Head architecture achieved remarkable improvements on obstacle detection benchmarks, effectively solving the "semantic blindness" of the previous tested methods. Ultimately, this research confirms that anomaly detection should not be treated only as a mere post-processing step. By embedding the "awareness of the unknown" directly into the network architecture, we pave the way for safer and more self-aware autonomous perception systems.

References

[1] 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*, 2021. 1, 4

[2] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Segmenting unknown regions rejected by all. In *ICCV*, 2021. 2

[3] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *NeurIPS*, 2021. 2, 4, 8

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

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

[6] Adriano Giuliani, Giacomo Lopez, Davide Reverberi, and Vincenzo Sammito. Synthetic anomaly dataset for road scene segmentation. https://www.kaggle.com/datasets/adrygiuliani/cityscape-cutpaste-coco-highlighted?resource=download, 2026. Accessed: January 2026. 3, 4, 8

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 2

[8] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 8

[9] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Scaling out-of-distribution detection for real-world settings. In *CVPR*, 2020. 2, 8

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8

[11] Chenfeng Xu, Wenhai Wang, Shuo Wang, Bo Li, et al. Your vit is secretly an image segmentation model. In *CVPR*, 2025. 2, 3, 8