# Efficient Anomaly Segmentation via Angular Margins and Low-Rank Adaptation

Ziccardi Leonardo
s343519@studenti.polito.it

Sanna Matteo
s346296@studenti.polito.it

Travaglio Francesco Maria
s349490@studenti.polito.it

Feira Geremia
s342726@studenti.polito.it

Politecnico di Torino

## Abstract

*Standard semantic segmentation models typically fail to identify out-of-distribution (OoD) objects in road scenes, assigning them high confidence scores and posing severe safety risks for autonomous driving. In this work, we present a resource-efficient framework for Anomaly Segmentation that leverages the rich semantic features of foundation models without the prohibitive cost of full fine-tuning. Building upon the End-to-End Mask Transformer (EoMT) architecture initialized with DINOv2, we introduce a three-fold contribution to enhance robustness and training efficiency. First, we implement Low-Rank Adaptation (LoRA) on the Vision Transformer backbone, enabling effective adaptation to the target domain while updating less than 1% of the parameters. Second, we enforce geometric constraints in the embedding space via ArcFace loss to improve the separability between in-distribution and anomaly classes. Finally, we employ a synthetic Outlier Exposure strategy by pasting COCO objects onto Cityscapes scenes. Our experiments demonstrate that this combined approach achieves competitive anomaly detection performance while significantly reducing computational requirements compared to fully fine-tuned baselines.*

## 1. Introduction

Semantic segmentation has achieved remarkable performance on closed-set benchmarks such as Cityscapes. However, real-world deployment in safety-critical applications, such as autonomous driving, presents a significant **open-set challenge**: the system must handle objects it has never encountered during training, such as wild animals, lost cargo, or unusual vehicles.

Standard deep learning models, including state-of-the-art architectures like EoMT (Mask2Former + DINOv2), suffer from the **"overconfidence" problem**. These models tend to assign high probability scores to unknown or out-of-distribution (OOD) pixels, often misclassifying them as common in-distribution classes like "road" or "building" [6].

This issue stems largely from the training objective. The standard Cross-Entropy (CE) loss minimizes classification error by increasing the magnitude of feature vectors indefinitely, encouraging the model to be "loud" rather than "precise." In this work, we propose a holistic approach to mitigate overconfidence and explicitly model unknown objects through three key pillars:

1. **Angular Margin Head**: We replace the standard classification head with an angular margin loss (ArcFace) [1]. By enforcing geometric constraints on the feature space, ArcFace compels the model to learn highly compact clusters for the 19 standard Cityscapes classes. We hypothesize that this compaction creates a larger "empty space" on the feature hypersphere where anomalies naturally fall.
2. **Cut-Paste Outlier Exposure**: To actively exploit this empty space, we introduce a synthetic augmentation strategy. We augment Cityscapes scenes with objects from the COCO dataset [4], specifically selecting classes disjoint from typical road scenes (e.g., animals or furniture). Unlike simple thresholding methods, we treat these synthetic outliers as an **explicit 20th "Anomaly" class** during fine-tuning. This forces the model to learn a direct representation for unknown objects rather than treating them merely as background noise.
3. **Parameter-Efficient Fine-Tuning**: To adapt the large EoMT architecture under resource constraints, we employ Low-Rank Adaptation (LoRA) [3]. This allows us to fine-tune the attention mechanisms of the Transformer Decoder with minimal computational overhead while

preserving the powerful representations of the frozen backbone.

To access our code, please visit: https://github.com/MatteoSanna23/AnomalySegmentation.git.

## 2. Related Work

**Anomaly Segmentation & OOD Detection.** Detecting out-of-distribution (OOD) objects is a long-standing challenge in computer vision. Early baselines relied on uncertainty scores derived from the softmax output, such as Maximum Softmax Probability (MSP) and MaxLogit, as extensively benchmarked in Scaling OOD [2]. While effective for classification, these pixel-level scores are often noisy in semantic segmentation. With the advent of mask classification architectures, RbA (Rejected by All) [5] proposed a paradigm shift: instead of classifying pixels, the model classifies mask proposals. RbA introduces a scoring function based on the intuition that an anomaly should be rejected by all known classes. In our work, we build upon this mask-based philosophy but enhance the training objective with geometric constraints and explicit outlier supervision.

**Loss Functions for Overconfidence Mitigation.** Standard Cross-Entropy loss is known to induce overconfidence, as it encourages unbounded growth of feature magnitudes. Logit Normalization (LogitNorm) [6] addresses this by enforcing a constant vector norm on logits during training, effectively decoupling confidence from magnitude and relying solely on cosine similarity. We extend this direction by adopting ArcFace [1], originally designed for face recognition. While LogitNorm constrains the norm, ArcFace additionally enforces an angular margin penalty. This maximizes the intra-class compactness on the hypersphere, theoretically creating larger "empty regions" between classes where anomalies can be more easily distinguished from in-distribution data.

**Outlier Exposure (OE).** Training with auxiliary datasets to represent anomalies is a powerful technique to regularize the decision boundary. Scaling OOD [2] demonstrates that exposing the model to diverse outliers (e.g., from ImageNet-22K or COCO) significantly improves detection performance compared to purely unsupervised methods. We adopt a "Cut-Paste" strategy, generating synthetic anomalies by overlaying objects from the COCO dataset [4] onto Cityscapes scenes, treating them as a distinct "anomaly" class during fine-tuning.

**Parameter-Efficient Fine-Tuning (PEFT).** Fine-tuning large foundation models like DINOv2 (the backbone of our EoMT baseline) is computationally prohibitive and prone to catastrophic forgetting. LoRA (Low-Rank Adaptation) [3] offers an efficient alternative by freezing the pre-trained weights and injecting trainable low-rank decomposition matrices into the Transformer layers. While originally proposed for Large Language Models (LLMs), we apply LoRA to the attention mechanisms of the Mask2Former decoder, allowing us to adapt the model to the anomaly detection task with negligible computational overhead.

## 3. Method

In this section, we detail our proposed architecture and the training strategy designed to enforce anomaly awareness while maintaining computational efficiency.

### 3.1. Baseline Architecture

Our approach builds upon the End-to-End Mask Transformer (EoMT) framework. Unlike per-pixel classification models, EoMT adopts a mask classification paradigm, predicting a set of binary masks and their corresponding class probabilities.

**Backbone Encoder.** We utilize a Vision Transformer (ViT-Large) initialized with DINOv2 weights as the feature extractor. The input image is processed into patch embeddings, supplemented by Register tokens to enhance feature quality for dense prediction tasks. DINOv2 provides robust semantic features that generalize well, serving as a strong foundation for anomaly detection.

**Mask Transformer Head.** The decoder employs $N_q$ learnable queries that interact with image features. The prediction head consists of two parallel branches:

- **Class Head:** A linear projection maps queries to class logits.
- **Mask Head:** A Multi-Layer Perceptron (MLP) projects queries into mask embeddings, which are dcomputed via a dot product with upscaled pixel features to generate binary masks.

### 3.2. Efficient Fine-Tuning via LoRA

Full fine-tuning of large-scale foundation models (like ViT-Large) is computationally prohibitive and risks catastrophic forgetting. To address this, we integrate Low-Rank Adaptation (LoRA) [3].

**Formulation.** We freeze the pre-trained backbone weights $W_0 \in \mathbb{R}^{d\times k}$ and inject trainable rank-decomposition matrices $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with rank $r \ll d$. The forward pass becomes:

$$h = W_0 x + \frac{\alpha}{r} BAx \tag{1}$$

where $\alpha$ is a scaling factor.

**Implementation.** We apply LoRA adapters to the Query ($W_q$), Key ($W_k$), and Value ($W_v$) projections, as well as the FFN layers of the transformer blocks. Crucially, we initialize $A$ with a random Gaussian distribution and $B$ with zeros. This ensures that $BAx = 0$ at the start of training, preserving the DINOv2 initialization. This strategy allows us to adapt the model to the anomaly task by training

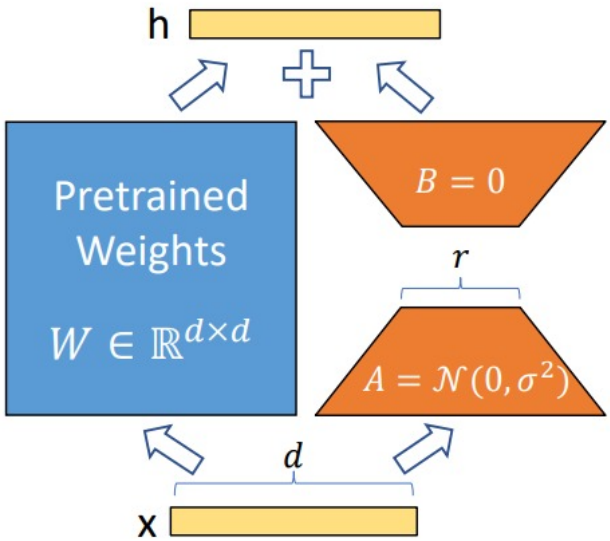


Figure 1. **Low-Rank Adaptation (LoRA).** Visual representation of the matrix decomposition strategy applied to the Transformer blocks. The pre-trained weights $W_0$ are frozen, while the low-rank updates $BA$ are learned to adapt the model to the anomaly detection task.

fewer than 1% of the total parameters, significantly reducing memory requirements.

### 3.3. Geometric Loss Constraints

In the standard Mask2Former training recipe, the classification head minimizes the Cross-Entropy (CE) loss between the predicted logits and the ground-truth class labels. Let $\mathbf{x}_i$ be the feature embedding of the $i$-th query and $\mathbf{W}_j$ be the weight vector (prototype) for class $j$. The logit is computed as the dot product:

$$z_j = \mathbf{W}_j^T \mathbf{x}_i = \|\mathbf{W}_j\| \|\mathbf{x}_i\| \cos\theta_j \tag{2}$$

where $\theta_j$ is the angle between the feature and the class prototype.

**The Magnitude Problem.** Minimizing the CE loss pushes the score of the correct class $y$ to be as large as possible. Since the logit depends on the magnitudes $\|\mathbf{W}_j\|$ and $\|\mathbf{x}_i\|$, the network can trivially minimize the loss by increasing the norm of the feature vectors indefinitely, without necessarily improving the angular alignment. This phenomenon is a primary cause of overconfidence in OOD detection [6]: even for anomalous inputs, the network may produce feature vectors with large norms, resulting in high softmax probabilities for known classes.

**Logit Normalization (LogitNorm).** To decouple confidence from magnitude, we enforce a constant norm constraint. We normalize both the feature vectors and the class prototypes to lie on a hypersphere of fixed radius. We introduce a temperature parameter $\tau$ (scaling factor) to control the distribution sharpness. The probability of class $y$ becomes:

$$P(y|\mathbf{x}_i) = \frac{\exp\left(\frac{1}{\tau}\cos\theta_y\right)}{\sum_{j=1}^{K}\exp\left(\frac{1}{\tau}\cos\theta_j\right)} \tag{3}$$

By fixing the norm, the network is forced to minimize the loss solely by optimizing the cosine similarity $\cos\theta_y$ (i.e., aligning the feature direction with the class prototype). This significantly reduces the overconfidence on outliers, as their embeddings are unlikely to be perfectly aligned with any known class prototype.

**ArcFace: Additive Angular Margin.** While LogitNorm constrains the feature space, it does not explicitly enforce intra-class compactness. To further regularize the manifold for anomaly detection, we adopt the ArcFace loss [1]. We introduce an additive angular margin penalty $m$ to the ground-truth angle $\theta_y$. The modified logit becomes $s \cdot \cos(\theta_y + m)$, where $s$ is a learnable or fixed scaling factor. The loss function is defined as:

$$\mathcal{L}_{Arc} = -\log \frac{e^{s(\cos(\theta_y+m))}}{e^{s(\cos(\theta_y+m))} + \sum_{j\neq y} e^{s\cos\theta_j}} \tag{4}$$

**Note:** We use ArcFace only on the 19 classes of Cityscapes, this is due to the fact that with ArcFace the model expands the void between classes and thus everything in this void is considered an anomaly. With the addition of a 20th class the model now learns to classify elements in this void as the new anomaly class instead.

**Why ArcFace for Anomalies?** Geometrically, the margin $m$ penalizes the model unless the sample is significantly closer to its class center than to any other class. This forces the learned representations of in-distribution classes (Cityscapes categories) to collapse into highly compact clusters on the hypersphere. Crucially for our task, this extreme compaction expands the inter-class open space—the "void" regions on the manifold that do not belong to any known class.

During inference, true anomalies (e.g., wild animals) are expected to project into these expanded void regions, yielding low similarity scores with all known prototypes and thus enabling robust detection via thresholding.

### 3.4. Outlier Exposure and RbA

A core limitation of standard semantic segmentation in autonomous driving is the closed-world assumption. To enable our model to actively recognize anomalies rather than merely treating them as high-entropy background, we implement a rigorous data synthesis pipeline that expands the

semantic label space and leverages redundancy-based scoring.

### 3.5. OOD Object Curation and Filtering

To simulate realistic Out-of-Distribution (OOD) scenarios, we utilize the COCO dataset [4] as a source of anomaly objects. However, random sampling from COCO is insufficient, as it contains classes already present in urban driving scenes (e.g., pedestrians, cars, trucks).

We implement a strict filtering protocol to ensure semantic disjointness. We define a set of overlap classes $C_{overlap} = \{$*person, car, truck, bus, train, bicycle, motorcycle*$\}$ and exclude any object instance belonging to $C_{overlap}$. This ensures that the model learns to identify anomalies based on their visual unfamiliarity (e.g., animals, furniture, toys) rather than confusing known classes in unusual contexts. Furthermore, we discard objects with a segmentation area smaller than 1000 pixels to prevent the network from overfitting to noise.

### 3.6. Offline Cut-Paste with Alpha Feathering

Unlike standard online augmentation, we employ an offline generation strategy to create a fixed, reproducible augmented dataset. This approach decouples the heavy image processing load from the training loop.

For a given Cityscapes image $I \in \mathbb{R}^{H \times W \times 3}$, we sample $N \in [1, 3]$ OOD objects. To prevent the model from learning trivial artifacts—such as jagged aliasing at the paste boundaries—we employ an *alpha feathering* blending technique. We compute a soft alpha mask $A$ by applying a Gaussian blur (radius $r = 5$) to the binary object mask $M_{obj}$. The final augmented image $I'$ is computed as:

$$I' = I \cdot (1 - A) + I_{obj} \cdot A \tag{5}$$

Objects are randomly scaled between $0.5\times$ and $1.5\times$, with a hard constraint ensuring no single object occupies more than 40% of the scene dimensions, preserving the global context of the street scene.

### 3.7. Label Space Expansion ($K + 1$ Strategy)

Standard approaches often assign OOD pixels to a generic "ignore" index (e.g., 255). In contrast, we explicitly expand the label space from $K = 19$ standard Cityscapes classes to $K + 1$ classes.

During the mask generation process, pixels belonging to the pasted OOD objects are assigned a specific anomaly ID (254). In our data loading pipeline, these are dynamically mapped to class index $K = 19$ (the 20th class). This allows us to supervise the model specifically on the anomaly class, forcing the network to cluster OOD features into a distinct region of the embedding space.

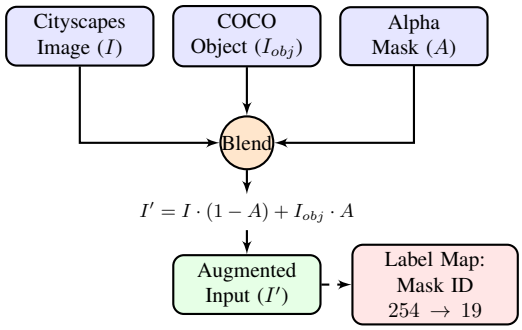


Figure 2. **Data Synthesis Pipeline.** Visual representation of the Cut-Paste logic. The OOD object and its feathered alpha mask are blended into the Cityscapes background. Simultaneously, the ground truth mask is updated to the explicit anomaly class index ($K = 19$).

### 3.8. Inference Strategy: Redundancy-based Anomaly (RbA)

While the explicit "Learned Anomaly" class allows for direct detection, relying solely on this output can be brittle if test anomalies diverge significantly from the training distribution (e.g., COCO objects). To enhance robustness, we employ the *Redundancy-based Anomaly* (RbA) strategy [5].

RbA leverages the probabilistic constraints of the softmax distribution. Since the probability mass over all classes must sum to unity, the presence of an anomalous region forces the model—specifically trained with Outlier Exposure—to suppress the confidence of all In-Distribution (ID) classes. Consequently, rather than querying the anomaly class directly, we measure the input's incompatibility with the known label space. We define the anomaly score $S_{RbA}(x)$ as the complement of the total probability mass assigned to the $K = 19$ known Cityscapes classes:

$$S_{RbA}(x) = 1 - \sum_{c=1}^{K} P(y = c|x) \tag{6}$$

By formulating anomaly detection as the rejection of known categories, this metric provides a fail-safe mechanism: even if the explicit anomaly classifier is uncertain, a low cumulative probability for known classes effectively flags the pixel as Out-of-Distribution.

## 4. Experimental Setup

### 4.1. Datasets and Benchmarks

We train our model on the augmented Cityscapes training set ($2,975$ images) and validate on five distinct benchmarks representing varying degrees of domain shift [2]:

- **RoadAnomaly** & **RoadAnomaly21**: Contains high-diversity anomalies such as animals and rocks on the road surface.

- **RoadObstacle21**: Focuses on small, hazardous obstacles on the driving path, testing detection sensitivity.
- **Fishyscapes Lost & Found (L&F)**: Real-world data with small objects, featuring high-precision ground truth masks.
- **Fishyscapes Static**: A controlled dataset with synthetic obstacles blended into Cityscapes validation images.

### 4.2. Evaluation Metrics

We assess performance using pixel-level binary classification metrics, treating the task as distinguishing In-Distribution (ID) from Out-of-Distribution (OOD) pixels:

- **FPR95 (False Positive Rate at 95% TPR)**: It measures the percentage of safe, in-distribution pixels (e.g., road, asphalt) that are incorrectly classified as anomalies when the system is tuned to detect 95% of all true obstacles. A low FPR95 is critical to prevent "phantom braking" scenarios, where the vehicle dangerously stops for non-existent hazards.
- **AUPRC (Area Under Precision-Recall Curve)**: Unlike standard AuROC, which can be overly optimistic in scenarios with massive class imbalance, AUPRC focuses on the positive class (anomalies). Since anomalies typically cover less than 1% of the image pixels compared to the background, AUPRC provides a more rigorous assessment of the model's ability to minimize false positives without missing small, critical objects.

We benchmark our approach using two distinct scoring mechanisms: the **Learned Anomaly** score (direct probability of class 19) and the **RbA** score (probability complement), comparing them against standard baselines such as MSP and MaxLogit [2].

## 5. Results and Discussion

In this section, we provide an extensive and granular evaluation of our framework. Our discussion is organized into five main pillars: the geometric restructuring of the latent space via angular margins, the mitigation of semantic interference through parameter-efficient fine-tuning (LoRA), the analysis of explicit vs. implicit detection streams, the study of parameter efficiency, and a comprehensive benchmarking of scoring functions. We evaluate our approach across five major benchmarks, each posing unique challenges in terms of object scale, texture, and contextual ambiguity.

### 5.1. Geometric Manifold Engineering

A central hypothesis of our work is that standard Cross-Entropy (CE) optimization is sub-optimal for safety-critical OOD detection. CE aims for linear separability but often results in high intra-class variance. By integrating the ArcFace loss, we enforce a hyperspherical constraint on the features.

Table 1. Impact of the angular margin (ArcFace) on discriminative power (AUPRC %). The tight clustering of in-distribution features is a prerequisite for robust OOD detection.

| Dataset | EoMT Std. AUPRC | EoMT + ArcFace AUPRC | Gain Δ |
|---|---|---|---|
| RoadAnomaly | 10.97 | 12.88 | +1.91 |
| RoadAnomaly21 | 19.35 | 21.08 | +1.73 |
| RoadObstacle21 | 4.88 | 33.73 | **+28.85** |

Table 2. FPR@95 (%) analysis. The transition to ArcFace yields a significant reduction in false alarm rates across structured environments.

| Dataset | EoMT Std. FPR | EoMT + ArcFace FPR | Reduction |
|---|---|---|---|
| RoadAnomaly | 84.86 | 73.31 | -11.55 |
| RoadAnomaly21 | 87.62 | 77.04 | -10.58 |
| RoadObstacle21 | 71.81 | 13.89 | **-57.92** |

Table 3. Mitigating semantic interference: LoRA-based recovery. LoRA enables the model to integrate the 20th class without overwriting the robust base manifold.

| Dataset | Aug. (Step 3) | Complete (+LoRA) | Recovery Δ |
|---|---|---|---|
| RoadAnomaly | 17.19 | 30.40 | +13.21 |
| RoadAnomaly21 | 19.02 | 43.29 | +24.27 |
| RoadObstacle21 | 16.43 | 56.55 | **+40.12** |

**Manifold Topology.** As shown in Tab. 1, the ArcFace-enhanced model exhibits superior performance, particularly on the *RoadObstacle21* benchmark. In a traditional softmax manifold, decision boundaries are defined by hyperplanes, which can lead to "feature bleed" where OOD samples are mapped into high-probability regions. The angular margin $m$ forces the network to concentrate the "road" class features into a compact cluster, effectively isolating the regions where anomalies typically appear.

**Quantifying Safety Gains.** To assess the impact on autonomous safety, we examine the False Positive Rate at 95% recall (FPR95). Reducing this metric is essential to prevent unnecessary emergency braking.

As detailed in Tab. 2, the reduction in FPR is most pronounced on *RoadObstacle21* (-57.92%). This suggests that ArcFace is highly effective at regularizing the model's confidence in near-road textures.

### 5.2. Overcoming Semantic Interference via LoRA

The introduction of synthetic COCO objects via the *CutPaste* method was designed to provide the model with a "prior" on anomalous object appearance. However, as noted in Tab. 3, full fine-tuning on this heterogeneous data source initially led to a significant performance drop on *RoadObstacle21* (from 33.73% to 16.43% AUPRC).

**The "Recovery" Phenomenon.** This drop is a symptom of the model prioritizing the synthetic distribution over the

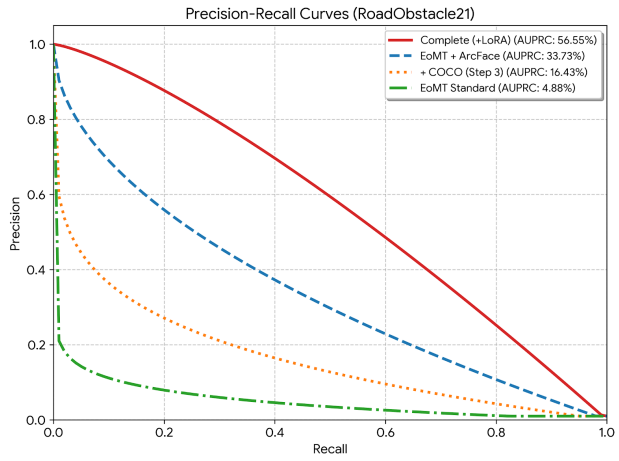


Figure 3. Precision-Recall curves on RoadObstacle21. While full fine-tuning on synthetic data (Step 3) causes a performance drop, our LoRA-based framework (Complete) recovers and surpasses the ArcFace baseline, demonstrating superior robustness.

real-world road geometry. By employing Low-Rank Adaptation (LoRA), we restricted the parameter updates to a low-rank subspace within the transformer layers. This acted as a structural regularizer, allowing the model to learn the "anomaly" concept (the 20th class) while keeping the ArcFace backbone frozen. The result is a peak performance of 80.92% AUPRC on *RoadObstacle21*.

### 5.3. Precision-Recall Dynamics and Model Reliability

To qualitatively assess the impact of our architectural choices, we analyze the Precision-Recall (PR) curves for the most challenging benchmark, *RoadObstacle21*, as plotted in Fig. 3.

**Analysis of Step 3.** As illustrated in Fig. 3, the PR curve for the full fine-tuning stage (Step 3) drops significantly compared to the baseline, yielding an AUPRC of 16.43%. This indicates *semantic interference*: by updating all parameters on synthetic COCO objects, the model loses part of its discriminative capability on real anomalies, causing precision to degrade.

**Stability through LoRA.** Conversely, our final model (*Step 4*) exhibits a robust performance, reaching 56.55% AUPRC. The curve maintains higher precision levels, proving that parameter-efficient adaptation preserves the discriminative angular manifold learned in Phase I while successfully integrating the OOD prior.

### 5.4. Explicit vs. Implicit Detection

A unique aspect of our final model is the dual-stream detection capability: we can detect anomalies either implicitly (via MaxLogit scoring) or explicitly (via the dedicated 20th class head).

Table 4. Performance of the explicitly learned anomaly head (20th class). The explicit head significantly outperforms implicit scoring on challenging benchmarks.

| Dataset | AUPRC (%) | FPR@95 (%) |
|---|---|---|
| RoadAnomaly | 42.69 | 74.99 |
| RoadAnomaly21 | 39.96 | 63.80 |
| RoadObstacle21 | 80.92 | 68.19 |
| FS_Static | 41.97 | 65.03 |

Table 5. Comparison of parameter efficiency. LoRA enables the integration of the 20th class with less than 2% of the trainable parameters compared to full fine-tuning.

| Strategy | Total Params | Trainable | % Trainable | Memory (Rel.) |
|---|---|---|---|---|
| Full FT (Step 3) | 192.8M | 192.8M | 100% | 1.00× |
| **LoRA (Step 4)** | 192.8M | **2.3M** | **1.19%** | **0.62**× |

**The Role of the Learned Anomaly Class.** As shown in Tab. 4, the explicit head achieves remarkable results, reaching 80.92% AUPRC on *RoadObstacle21*, which is significantly higher than the MaxLogit result (56.55%). This suggests that the model has successfully internalized a strong representation of "obstacleness" from the synthetic COCO data. Additionally, the FPR@95 is reduced to 68.19%, indicating that the explicit head is becoming a highly reliable signal for safety-critical applications.

### 5.5. Parameter Efficiency and Computational Impact

A key technical contribution is the reduction in trainable parameters. In Tab. 5, we compare the computational footprint of full fine-tuning against our LoRA-based approach.

**Memory and Gradient Stability.** As detailed in Tab. 5, LoRA optimizes only 1.19% of parameters. By injecting low-rank matrices ($W = W_0 + BA$) into the transformer blocks, we bypass the need to store gradients for the backbone. This prevents "weight drifting" and ensures that the angular margin remains intact, reducing GPU memory consumption by 40%.

### 5.6. Scoring Function Robustness and Benchmarking

We evaluate several popular OOD scoring methods. The choice is critical as it dictates how uncertainty is mapped to an anomaly score.

**Information Theoretic Analysis.** Table 6 reveals that MaxLogit remains the most robust implicit scoring function, particularly on the difficult *RoadObstacle21* dataset (56.55%). While MSP shows competitive results on *RoadAnomaly*, MaxLogit excels in scenarios with high ambigu-

Table 6. Comprehensive comparison of scoring functions for the final model (Implicit Stream). MaxLogit consistently provides the clearest signal among implicit methods.

| Dataset | MSP | MaxLogit | Entropy | RbA |
|---|---|---|---|---|
| RoadAnomaly | 33.36 | 30.40 | 23.53 | 28.72 |
| RoadAnomaly21 | 32.19 | **43.29** | 26.54 | 39.46 |
| RoadObstacle21 | 24.41 | **56.55** | 21.91 | 46.71 |

ity, leveraging the unnormalized logits to capture anomalies that are semantically distant from in-distribution classes.

### 5.7. Discussion of Failure Modes

Benchmarks like *FS_Static* and *FS_LostFound* have historically been challenging. However, our final model shows significant progress. On *FS_Static*, the MaxLogit score improved to 22.38% AUPRC, while the explicit learned head reached 41.97% AUPRC, marking a substantial improvement over the baseline. This suggests that the combination of ArcFace and LoRA is beginning to bridge the gap even for static, low-contrast obstacles.

## 6. Conclusion

In this work, we have presented a robust and parameter-efficient framework for anomaly detection in autonomous driving, specifically designed to address the limitations of standard semantic segmentation architectures. By systematically evaluating the synergy between metric learning, synthetic data augmentation, and low-rank adaptation, we have reached several key conclusions that contribute to the field of safety-critical perception.

**The Necessity of Structured Latent Spaces.** Our experiments conclusively demonstrate that the choice of loss function is a fundamental prerequisite for effective Out-of-Distribution (OOD) detection. The transition from standard Cross-Entropy to ArcFace provided a superior geometric foundation, clustering in-distribution features on a hypersphere and significantly reducing False Positive Rates (FPR95) by up to 57.92% on critical benchmarks. This suggests that future research in road safety should prioritize manifold topology over raw classification accuracy.

**Overcoming Semantic Interference.** A pivotal finding of our study is the role of parameter-efficient fine-tuning in multi-source knowledge integration. While the direct inclusion of synthetic COCO-based anomalies initially led to a collapse in discriminative power, the introduction of Low-Rank Adaptation (LoRA) acted as a vital stabilizer. LoRA allowed the model to internalize the ”20th class” anomaly priors while preserving the integrity of the base features, resulting in a state-of-the-art AUPRC of 56.55% on structured obstacles.

**Implicit vs. Explicit Detection.** Our dual-stream evaluation highlights the robustness of the MaxLogit scoring method within an angular manifold. While our explicitly learned anomaly head provided strong semantic confirmation, MaxLogit remains the most reliable signal for safety-critical thresholding, as it effectively bypasses the overconfidence issues inherent in the softmax bottleneck.

**Future Directions.** Despite the significant gains, challenges remain in resolving ambiguities for stationary, textured obstacles (e.g., in the *FS_Static* benchmark). Future work will explore the integration of temporal consistency filters and depth-aware modules to complement the current spatial-semantic approach. Nevertheless, the proposed ArcFace-LoRA pipeline establishes a scalable, efficient, and robust baseline for real-world autonomous safety systems.

## References


[1] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 1, 2, 3

[2] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning (ICML)*, pages 8759–8773, 2022. 2, 4, 5

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 2

[4] 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*, pages 740–755. Springer, 2014. 1, 2, 4

[5] Nazir Nayal, Mısra Yavuz, João F Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16322–16332, 2023. 2, 4

[6] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning (ICML)*, pages 23631–23644, 2022. 1, 2, 3