Anomaly Segmentation: Pixel-based ERFNet compared to Mask-based EoMT. Implementing different post-hoc methods for anomaly scoring.

Luigi Grasso Kareem Fakhreddine
Haroun Khaled
Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
s334166@studenti.polito.it, s333147@studenti.polito.it, s334079@studenti.polito.it

Abstract

*In the context of autonomous driving systems, one of the most safety-critical requirements would be a reliable detection of Out-of-Distribution (**OoD**) objects when deployed in open world environments. Standard semantic segmentation models often exhibit the phenomenon of overconfidence when dealing with anomalous inputs, resulting in model failure. This work is aimed at presenting a thorough evaluation and enhancement of anomaly segmentation frameworks. In this work, we deal with two baselines for evaluation: pixel-based baselines, namely **ERFNet**, and the state-of-the-art (**SOTA**) mask classification baselines, specifically the Encoder-only Mask Transformer (**EoMT**) with a **DINOv2** backbone. We evaluate the efficacy of the following post-hoc methods: Maximum Softmax Probability (**MSP**), Maximum Logit (**MaxL**), Maximum Entropy (**MaxE**), and "Rejected-by-All" method (**RbA**) for mask classification architectures. Furthermore, we address the issue of model overconfidence by implementing an additional layer to the MSP post-hoc function known as temperature scaling. Our Experiments are implemented on SegmentMeIfYouCan (**SMIYC**) RoadAnomaly21 (**RA21**) and RoadObsticle21 (**RO21**) datasets, Fishyscapes (**FS**) Lost and Found (**FS LF**) and static (**FS static**) dataset, as well as RoadAnomaly (RA). Our experiments are evaluated based on two metrics: the Area Under the Precision–Recall Curve (**AuPRC**) and False Positive Rate for a 95% True Positive Rate (**FPR95**). An additional metric is considered, mean Intersection over union (**mIoU**).*

1. Introduction

Many deep neural networks have achieved remarkable success in the semantic segmentation of urban scenes. However, detecting "unknown objects" or OoDs (anomalous objects that were not present in the training distribution) remains a persistent challenge to these networks. The ERFNet model, for example, with a pixel-based classifier often assigns high probability scores to these anomalies due to the "overconfidence" phenomenon caused by the Cross-entropy optimization of the original ERFNet, where the model increases logit magnitude to minimize loss.

This work explores the transition from pixel-based classification to mask classification architecture by making use of the EoMT model. Preliminary experimentation shows a superior performance of mask architectures compared to pixel-based ones. However, mask architectures remain susceptible to overconfidence as well. In order to address this issue, the following pipeline was proposed:

1. **Architecture Transition:** Preliminary comparison between pixel-based ERFNet and mask-based EoMT.
2. **Post-hoc investigation:** Testing different post-hoc methods applied to both architectures and examining their performance.
3. **Temperature Scaling:** Investigate the effect of temperature-scaling on improving the MSP performance on EoMT.

We demonstrate a slight improvement brought by the addition of the temperature scaling method, which signifies the important role that normalizing outputs plays.

We achieve a robust separation between In-Distribution (ID) and OoD data, significantly outperforming standard baselines on the Road Anomaly and SMIYC benchmarks.

2. Related Work

- **Anomaly Segmentation:** The task of segmenting OoD anomalous objects while maintaining ID accuracy. This task is benchmarked by specialized datasets such as SMIYC [RA21 and RO21], Fishyscapes [Lost and Found and Static], and RoadAnomaly [RA]. Early works on this task involved approaches that rely on pixel-level uncertainty estimation.

- **Mask Architecture:** New approaches have been recently presented, such as Mask2Former and EoMT, which shifted the paradigm from pixel-based methods to predicting binary masks and class labels via transformer decoders. The adaptation of this new approach opens the way for new anomaly detection methods such as "Rejected-by-All" (RbA), which groups and identifies regions that are not identified to belong to any class query.

- **ERFNet:** A deep neural network architecture designed for **real-time semantic segmentation**. It realies on a "Non-Bottleneck-1D" module where convolutions are decomposed to 1D factorized convolutions (Instead of 2D convolution, e.g, 3 x 3, decomposed 1D convolutions, e.g, 3 x 1 followed by 1 x3). The architecture is mainly constituted of an Encoder and a Decoder. The encoder is composed of 16 layers with three downsampling stages to gather context. The decoder is composed of seven layers (layers 17 to 23, included) which upsample the features to match the input resolution. Our starting code implements the ERFNet architecture while wrapping it with an anomaly detection post-hoc method which extracts logits to calculate anomaly scores.

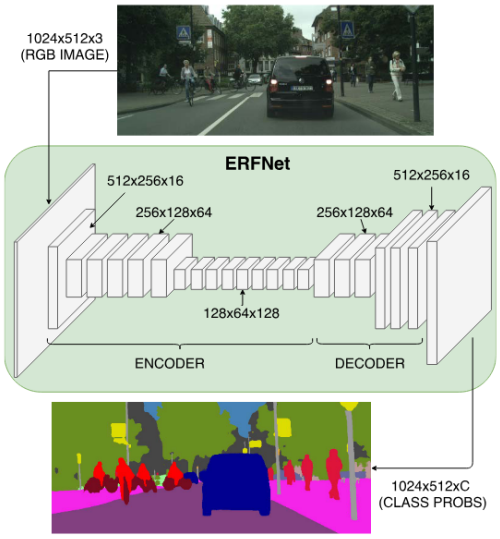

Figure 1. Diagram that depicts the proposed segmentation system (ERFNet) for an example input image and its corresponding output (C = 19 classes). The depicted volumes correspond to the feature maps produced by each layer. All spatial resolution values are with regard to the example input (1024×512), but the network can operate with arbitrary image sizes.[8]

- **Confidence Calibration:** Neural networks are usually uncalibrated. To mitigate this, temperature scaling is employed as a standard post-hoc method that softens the SoftMax distribution. However, to better mitigate the issue of confidence calibration, it is suggested to investigate the effect of modifying the loss logic in the EoMT architecture by replacing standard Cross-Entropy with Logit Normalization, which normalizes the logits during training, hence addressing the issue at its roots.

3. Method

3.1. Architectures and Baselines

We employ two distinct architectures:

- **Pixel-Based:** ERFNet [8], a lightweight efficient network used as a baseline.
- **Mask-Based:** EoMT (Environment-aware Mask Transformer) initialized with a DINOv2 backbone [7].

EoMT2: An architecture designed for mask classification that repurposes a plain Vision Transformer (ViT) for segmentation tasks. The architecture is constituted of a ViT backbone, specifically, DINOv2, which is divided into two stages[5]: L1 blocks extract features from image patch tokens, after extraction, learnable object queries are concatenated to the tokens, which in turn are processed by the L2 blocks to predict mask and class labels. Our starting code makes use of this original EoMT architecture while wrapping it with mask-specific anomaly detection methods. The EoMT architecture allows the introduction of a new post-hoc method, RbA (Rejected by All), to identify OoD objects.

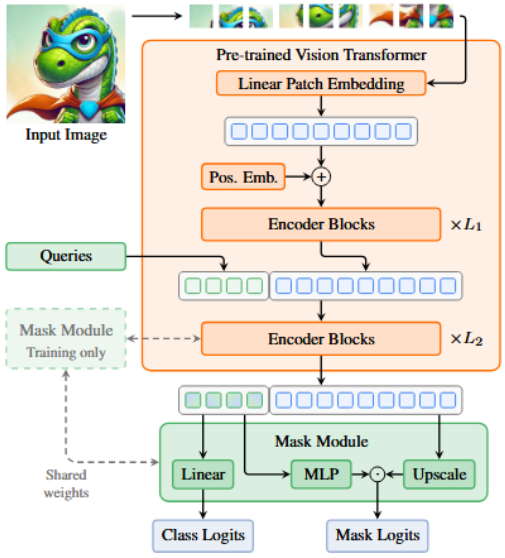

Figure 2. EoMT architecture. Learnable queries are concatenated to the patch tokens after the first L1 ViT encoder blocks. These concatenated tokens are then jointly processed by the last L2 blocks and used to predict class and mask logits.[5]

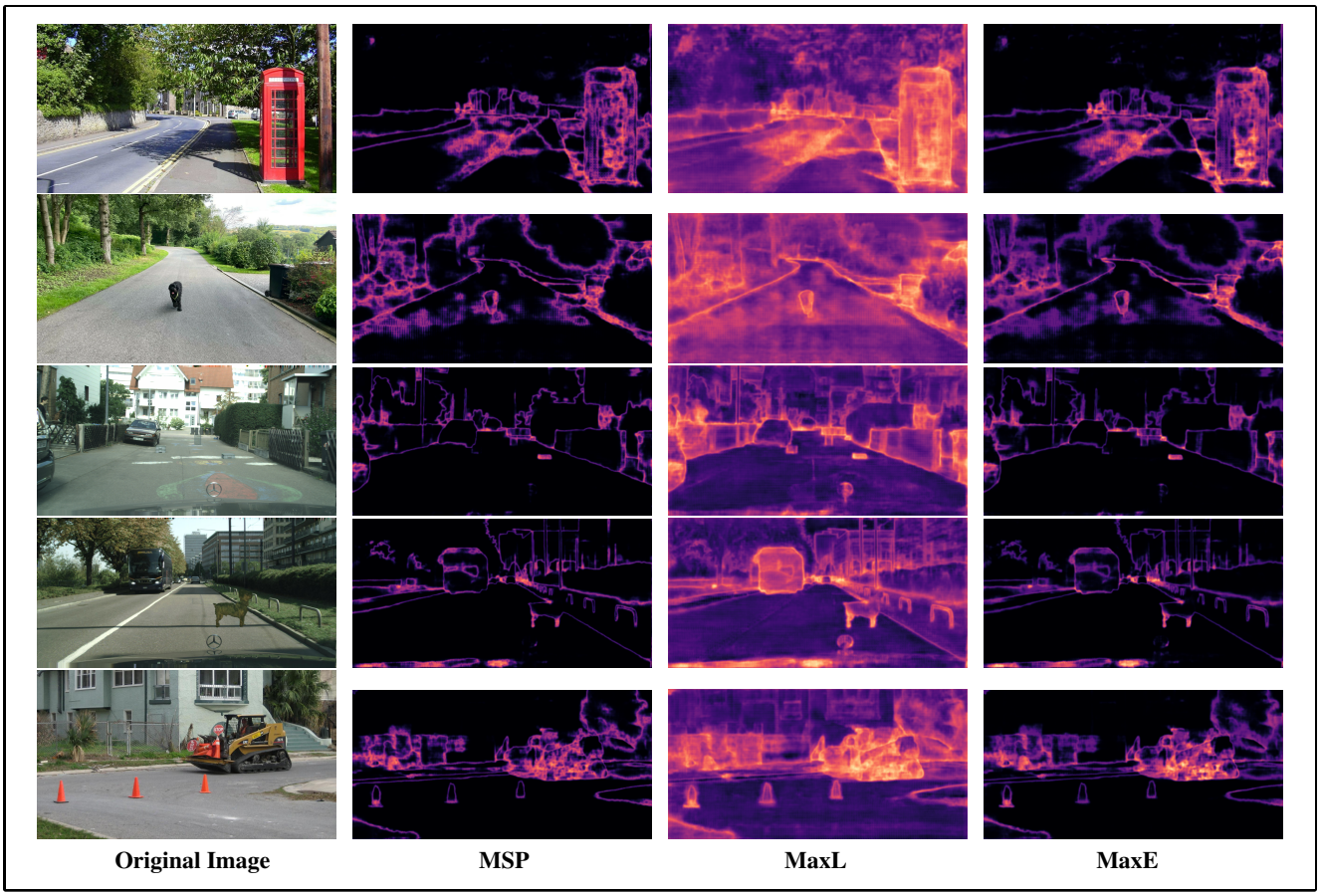

Figure 3. Qualitative comparison for ERFNet. (1) SMIYC RoadAnomaly21 - (2) SMIYC RoadObsticle21 - (3) Fishyscapes Lost & Found - (4) Fishyscapes Static - (5) RoadAnomaly. The columns show: **Original Image**, **MSP**, **MaxLogit**, and **MaxEntropy**. Brighter regions indicate higher anomaly scores.

3.2. Post-Hoc Detection Scoring

We evaluate several scoring functions $S(x)$ to distinguish ID from OoD pixels:

- **MSP:**
$$S(x) = 1 - \max_c \text{Softmax}(f(x))_c \tag{1}$$
- **MaxLogit:**
$$S(x) = -\max_c f_c(x) \tag{2}$$
- **MaxEntropy:**
$$S(x) = -\sum_c p_c(x) \log p_c(x), \tag{3}$$
$$\text{where } p_c(x) = \text{Softmax}(f(x))_c$$
- **RbA (Rejected by All):** Specific to mask architectures, this method aggregates the inverse probabilities of mask queries to identify regions rejected by all semantic classes [6].

3.3. Calibration via Temperature Scaling

To address overconfidence in the pre-trained EoMT model, we apply Temperature Scaling [3]. The logits z are divided by a scalar T:

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{4}$$

We perform a grid search for $T \in \{0.5, 0.75, 1.1, \dots\}$ to find the "Best T" that maximizes the Area Under the Precision-Recall Curve (AuPRC) on the validation set.

4. Numerical Experiments

4.1. Experimental Setup

- **Training Data:** Cityscapes Dataset (5,000 finely annotated urban images)[2].
- **Test Data (OoD):** SMIYC Road Obstacle 21 (RO-21), Road Anomaly (RA-21), RoadAnomaly (RA), and Fishyscapes Lost and Found and Static.[1].

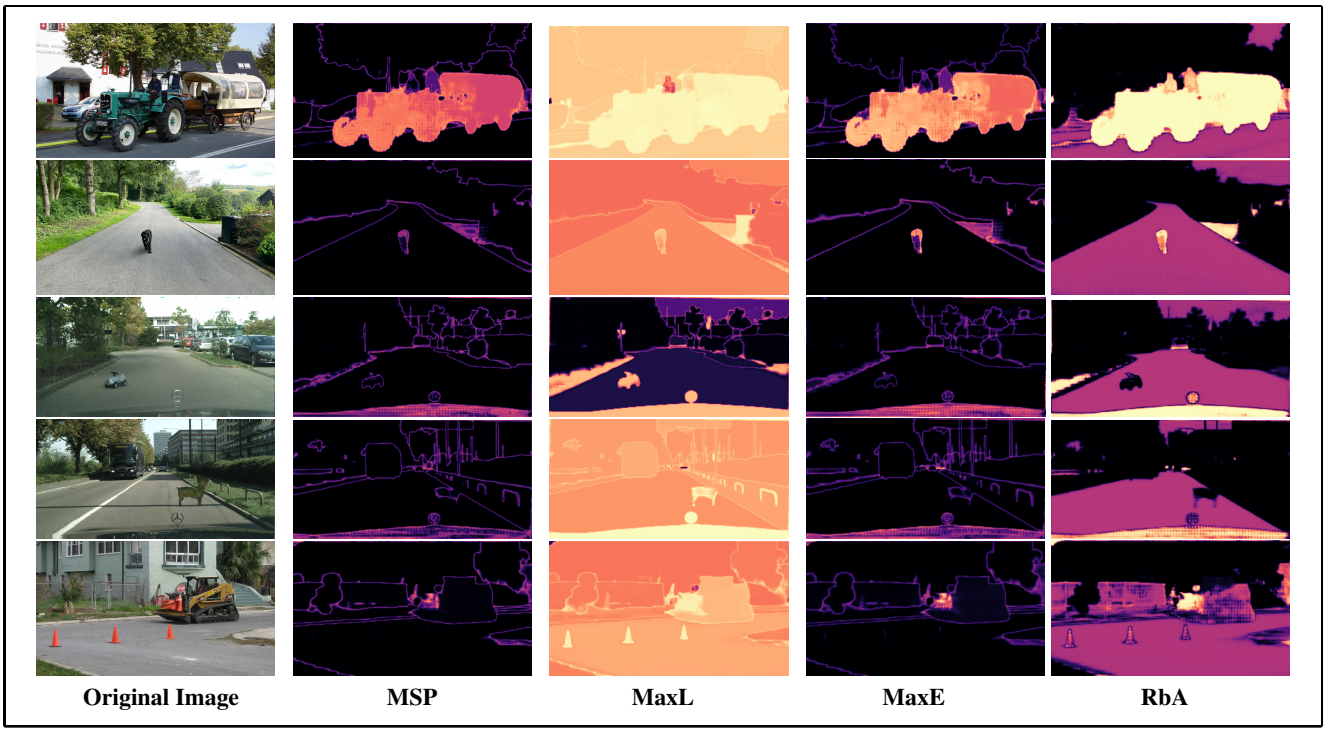

Figure 4. Qualitative comparison for EoMT. (1) SMIYC RoadAnomaly21 - (2) SMIYC RoadObsticle21 - (3) Fishyscapes Lost & Found - (4) Fishyscapes Static - (5) RoadAnomaly. The columns show: **Original Image**, **MSP**, **MaxLogit**, **MaxEntropy**, and **RbA**. Brighter regions indicate higher anomaly scores.

- **Metrics:** Mean Intersection over Union (mIoU) for ID performance; AuPRC and False Positive Rate at 95% Recall (FPR95) for anomaly detection.
- **Hardware:** Experiments were conducted locally on a single GPU: NVIDIA GeForce RTX 3050 Ti.

4.2. Procedure

1. **Baseline Evaluation:** Run inference with ERFNet and Pre-trained EoMT using MSP, MaxLogit, and MaxEntropy[4].
2. **Mask-Specific Evaluation:** Apply RbA to EoMT[6].
3. **Temperature Optimization:** Save logits from EoMT and optimize T offline.

5. Results

5.1. Baseline Performance

Table 1 compares a conventional pixel-based architecture (ERFNet) to a mask-classification model (EoMT) under multiple post-hoc anomaly scoring functions (MSP, MaxLogit, MaxEntropy), and the mask-specific RbA score for EoMT. Overall, the backbone choice has a strong impact on OoD segmentation performance: using the same MSP scoring, EoMT yields large gains over ERFNet on the SMIYC benchmarks (e.g., SMIYC RA-21 AuPRC 29.08%→61.39% with FPR95 62.55%→33.28%, and SMIYC RO-21 AuPRC 2.71%→70.80% with FPR95 65.25%→9.10%). These results suggest that the mask-based formulation separates ID road pixels from OoD obstacles more effectively than dense pixel-wise classification, while also improving ID segmentation accuracy (mIoU 46.38%→61.81% under MSP).

Within the same model, the choice of post-hoc score produces different trade-offs. For EoMT, MaxEntropy provides a strong and comparatively stable detection signal across datasets (e.g., RO-21 AuPRC 77.17% with FPR95 12.20%), while MaxLogit is less reliable in our setting and degrades both ID segmentation and OoD detection performance (e.g., mIoU drops to 24.50% and RO-21 FPR95 increases to 85.85%). RbA leverages the query-based structure of mask-classification models by highlighting regions rejected by all ID class queries; while it provides complementary behavior and can qualitatively emphasize anomalies, it is not uniformly best across all datasets in Table 1 (e.g., on RA-21, AuPRC 54.62% vs. 61.39% for MSP and 63.47% for MaxEntropy).

5.2. Impact of Temperature Scaling

Table 2 evaluates temperature scaling applied to EoMT logits for MSP-based anomaly scoring. We treat $T = 1.0$

as the uncalibrated baseline and evaluate multiple temperatures. Temperature scaling can improve separation between ID and OoD by reducing overconfident predictions: on SMIYC RA-21, $T = 1.1$ increases AuPRC from 61.39% to 62.51% and reduces FPR95 from 33.28% to 30.61%, as Figures 5 and 6 show. It can also slightly improve ID segmentation in this case (mIoU 61.81%→62.67%). However, the benefit is dataset-dependent; for example, on FS Static a lower temperature ($T = 0.5$) substantially reduces FPR95 (91.92%→69.77%), while $T = 1.1$ yields only a modest improvement (91.92%→89.60%). In our experiments, the best-performing temperature for Road Anomaly under our validation protocol was approximately $T = 1.1$, consistent with the intuition that softening the softmax distribution ($T > 1$) can mitigate overconfidence [3].

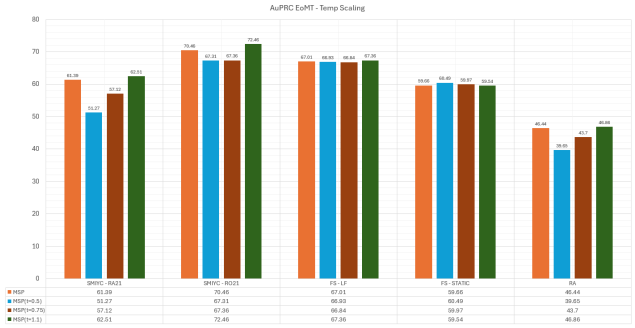

Figure 5. Effect of temperature scaling on EoMT with MSP anomaly scoring. AuPRC across datasets as T varies.

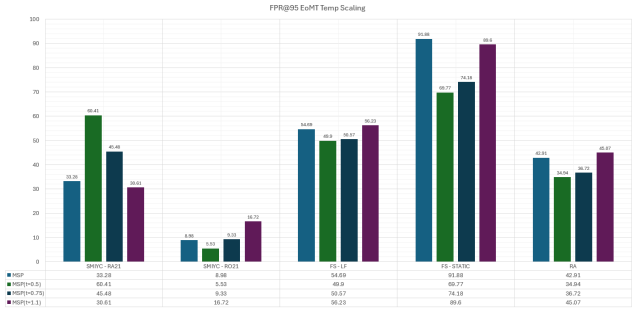

Figure 6. Effect of temperature scaling on EoMT with MSP anomaly scoring. FPR95 across datasets as T varies.

6. Conclusion

In this project, we built and evaluated an anomaly segmentation pipeline for road scenes in open-world settings, where semantic segmentation models must reliably highlight out-of-distribution (OoD) obstacles while preserving in-distribution (ID) accuracy. Across the evaluated benchmarks (SMIYC RA-21/RO-21, Fishyscapes Lost & Found/Static, and RoadAnomaly), the mask-classification paradigm (EoMT) consistently provides a stronger basis for OoD separation than the pixel-based ERFNet baseline when paired with standard post-hoc confidence scores. We further analyzed multiple post-hoc scoring functions (MSP, MaxLogit, MaxEntropy) and observed that the score choice can significantly change the precision–recall vs. false-positive trade-off, highlighting the importance of selecting a detection score that matches the target OoD regime. For mask-based models, the Rejected-by-All (RbA) signal offers an additional, architecture-specific cue by exploiting query-level rejection; while it is not uniformly best across all datasets, it provides complementary behavior and can qualitatively emphasize regions that are not well captured by purely confidence-based scores. Finally, we investigated confidence calibration via temperature scaling and found that it can yield modest improvements by softening overconfident predictions; however, the optimal temperature is dataset-dependent, suggesting that calibration should be tuned with care to the intended deployment conditions. Future work should therefore focus on more general methods to better mitigate the overconfidence issue. As mentioned, targeting the training phase of the EoMT model by introducing an anomaly-oriented loss logic (e.g, Logit Normalization instead of Cross-Entropy) can potentially provide a more general and robust solution to the issue of overconfidence.

			SMIYC RA-21		SMIYC RO-21		FS LF		FS Static		Road Anomaly	
Model	**Method**	**mIoU↑**	**AuPRC↑**	**FPR95↓**	**AuPRC↑**	**FPR95↓**	**AuPRC↑**	**FPR95↓**	**AuPRC↑**	**FPR95↓**	**AuPRC↑**	**FPR95↓**
ERFNet	**MSP**	**46.38**	**29.08**	**62.55**	**2.71**	**65.25**	**40.28**	**86.42**	**32.90**	**82.95**	**12.42**	**82.58**
	MaxL	43.97	32.88	70.6	5.26	44.3	44.45	83.94	37.94	84	16.16	79.58
	MaxE	47.02	30.61	62.74	3.03	62.56	42.87	86.73	35.82	83.03	12.84	82.61
EoMT	**MSP**	**61.81**	**61.39**	**33.28**	**70.8**	**9.1**	**67.03**	**54.73**	**59.67**	**91.92**	**46.44**	**42.91**
	MaxL	24.5	55.08	39.73	23.64	85.85	70.58	81.14	80.75	67.88	47.24	62.22
	MaxE	54.02	63.47	32.75	77.17	12.2	71.43	55.3	63.13	83.00	51.04	40.03
	RbA	50.87	54.62	59.53	66.87	99.60	58.94	65.13	39.78	61.41	30.69	71.48

Table 1. Quantitative comparison of anomaly detection performance between ERFNet (CNN-based) and EoMT (Transformer-based). Results are reported for MSP, MaxLogit, MaxEntropy, and RbA across the SMIYC RoadAnomaly21, SMIYC RoadObsticle21, Fishyscapes Lost & Found, Fishyscapes Static, and RoadAnomaly datasets. Evaluation metrics include mean Intersection over Union (mIoU%), Area Under the Precision–Recall Curve (AuPRC%), and False Positive Rate at 95% True Positive Rate (FPR95%).

			SMIYC RA-21		SMIYC RO-21		FS LF		FS Static		Road Anomaly	
Model	**Method**	**mIoU↑**	**AuPRC↑**	**FPR95↓**	**AuPRC↑**	**FPR95↓**	**AuPRC↑**	**FPR95↓**	**AuPRC↑**	**FPR95↓**	**AuPRC↑**	**FPR95↓**
EoMT	MSP	**61.81**	**61.39**	**33.28**	**70.8**	**9.1**	**67.03**	**54.73**	**59.67**	**91.92**	**46.44**	**42.91**
	$T = 0.5$	59.12	51.27	60.41	67.31	5.53	66.93	49.90	60.49	69.77	39.65	34.94
	$T = 0.75$	59.98	57.12	45.48	67.36	9.33	66.84	50.57	59.97	74.18	43.70	36.72
	$\mathbf{T = 1.1}$	**62.67**	**62.51**	**30.61**	**72.46**	**16.72**	**67.36**	**56.23**	**59.54**	**89.60**	**46.86**	**45.07**

Table 2. Impact of temperature scaling on EoMT anomaly detection performance. Comparison between the baseline MSP method and MSP-T with $T_{\text{best}} = 1.1$.

References

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

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. 3, 5

[4] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 4

[5] Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your vit is secretly an image segmentation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25303–25313, 2025. 2

[6] Nazir Nayal, Misra Yavuz, Joao F Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 711–722, 2023. 3, 4

[7] 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

[8] 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*, 19(1):263–272, 2017. 2