Real-Time Anomaly Segmentation for Road Scenes

GitHub Repository

Youness Bouchari
Politecnico di Torino
s323624@studenti.polito.it

Luca Bergamini
Politecnico di Torino
s332167@studenti.polito.it

Luca Rabellino
Politecnico di Torino
s317688@studenti.polito.it

1 Abstract

Anomaly segmentation in road scenes is a key task for safe autonomous driving, but many high-performing models are too slow for real-time use. In this work, we start from baseline segmentation models and explore ways to speed up inference through pruning and quantization techniques. Our goal is to reduce computational load while keeping accuracy as high as possible.

2 Introduction

Anomaly segmentation in road scenes is a critical task for ensuring the safety and robustness of autonomous driving systems. Traditional semantic segmentation models are often trained in closed-world settings, where the assumption is that all test-time classes are known and present during training. However, in real-world road environments, unexpected or out-of-distribution (OoD) objects—such as road debris, construction materials, or animals—can appear. Accurately detecting and segmenting these anomalies in real time is essential for effective decision-making and hazard avoidance in autonomous vehicles.

In this project, we started by evaluating existing lightweight baseline models for anomaly segmentation, including ENet, ERFNet, and BiSeNet, using the Cityscapes dataset for training and the Fishyscapes and Road Anomaly datasets for evaluation. As an extension, we focused on improving the real-time feasibility of these models by applying model compression techniques—specifically pruning and quantization. These techniques aim to reduce model size and inference latency while maintaining acceptable performance in anomaly segmentation tasks.

Our results show that carefully applied pruning and quantization can lead to significant improvements in computational efficiency, making real-time deployment of anomaly segmentation models more practical without a drastic loss in accuracy.

3 Background

In recent years, semantic segmentation has seen substantial progress, particularly in the context of real-time applications such as autonomous driving. However, the task becomes significantly more challenging when the system must not only classify known categories but also detect unexpected or out-of-distribution (OoD) objects — a scenario common in open-world environments. This is where anomaly segmentation becomes essential.

To achieve reliable performance in real-time, especially on embedded systems with limited compute and memory, researchers have developed a range of lightweight semantic segmentation networks. These models form the foundation for our work.

One of the most influential early models for efficient segmentation is ENet [1], designed specifically for real-time applications. Its architecture adopts an asymmetric encoder-decoder structure and aggressively downsamples the input early in the network using a combination of strided convolutions and max pooling. This initial block reduces computational cost right from the start. ENet's core building unit is a bottleneck module that leverages factorized convolutions and residual connections to keep both parameter count and inference time low, while maintaining performance.

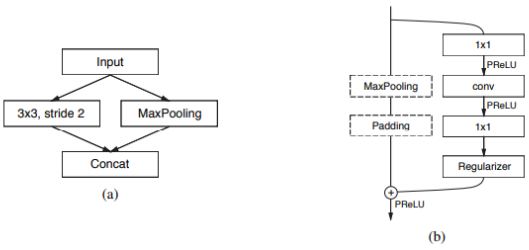

Figure 1: ENet architecture. **(a)** Initial block with strided convolutions and max pooling for early downsampling. **(b)** Lightweight bottleneck modules with factorized convolutions and residual connections for efficient feature extraction.

Another widely adopted architecture is BiSeNet [2], which introduced the idea of separating the processing of spatial and contextual information. The model uses two distinct branches: a spatial path that preserves high-resolution features and a context path that captures semantic information with a larger receptive field. These are fused later using a feature fusion module that balances detail and context effectively. BiSeNet V2 [3] refines this design with a more lightweight and modular structure, introducing a guided aggregation module to further improve speed and accuracy for mobile and edge use cases.

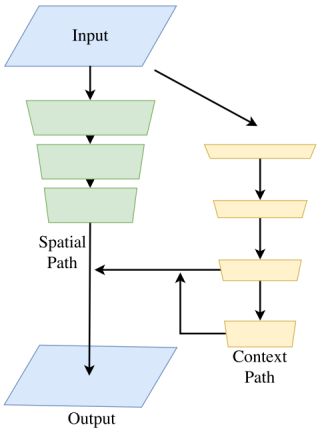

Figure 2: BiSeNet architecture

ERFNet [4] (Efficient Residual Factorized ConvNet) combines the strengths of residual learning with efficient depthwise separable convolutions. Its architecture is built from non-bottleneck-1D blocks arranged in a deep residual framework. This design allows ERFNet to capture semantic features effectively while maintaining real-time performance. Its balance of accuracy and efficiency makes it particularly suitable for real-world road scene segmentation and anomaly detection.

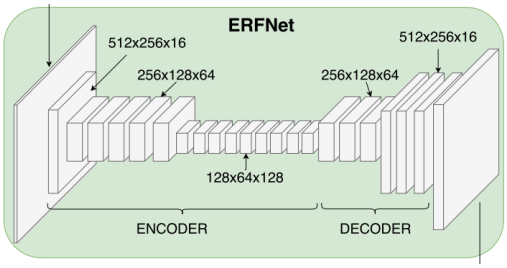

Figure 3: Erf-Net architecture

Although ICNet [5] also contributed to early real-time models by using a multi-resolution fusion strategy, in our work we focus more on ENet, BiSeNet, and ERFNet due to their modular design and suitability for compression techniques like pruning and quantization.

While these models are not explicitly designed to handle anomalies, they serve as powerful backbones for segmentation tasks. To detect anomalies effectively, the SegmentMeIfYouCan [6] and Fishyscapes [7] benchmarks were introduced. These datasets contain real-world road scenes with unexpected objects (e.g., animals, pedestrians in unusual poses, road debris) and serve as critical tools for evaluating the robustness of anomaly segmentation systems. They also define common evaluation metrics such as the Area under Precision-Recall Curve (AuPRC) and False Positive Rate at 95% Recall (FPR95).

Despite advances in segmentation and anomaly detection, deploying these models in real-time systems remains a challenge due to computational constraints. This motivates the use of model compression techniques — namely, pruning and quantization — to reduce inference latency and model size while preserving anomaly detection performance.

4 Materials and Methods

4.1 Datasets

Our experiments utilize several datasets tailored for semantic segmentation and anomaly detection in urban driving environments.

Cityscapes serves as our primary training dataset. It comprises 5,000 high-resolution images collected from 50 different cities, featuring diverse urban scenes with fine-grained, pixel-level annotations for 19 semantic classes. The dataset is divided into 2,975 training, 500 validation, and 1,525 test images, and includes a void class to represent background or rare objects. Its comprehensive labeling and scene diversity make it a strong foundation for training robust segmentation models.

Fishyscapes is employed for evaluating anomaly and out-of-distribution (OOD) detection capabilities. It introduces two evaluation tracks: **Lost and Found**, which contains scenes with small, unusual objects derived from the Lost and Found dataset, and **FS Static**, which augments Cityscapes validation images with anomalous objects not present during training. These tracks challenge models to detect both real and synthetic anomalies within familiar environments.

From SegmentMeIfYouCan, a benchmark specifically designed for anomaly segmentation, sub-datasets are extracted: **RoadAnomaly21**, which evaluates general anomaly segmentation in street scenes, and **RoadObstacle21**, which focuses on realistic road obstacles that represent immediate hazards for autonomous vehicles. These tracks emphasize the model's ability to detect and segment critical out-of-distribution elements within complex urban settings.

RoadAnomaly targets anomaly detection in real-world scenarios by providing 60 images with pixel-level annotations. The anomalous objects appear in unpredictable locations within the scene, posing a significant challenge for segmentation models in identifying safety-critical, rare hazards.

4.2 Baseline Methods and Metrics

Three well-known methods for anomaly inference were applied to assess the performance of a pre-trained ERFNet model on the test datasets. Maximum Softmax Probability **(MSP)** uses the highest softmax probability as a confidence score, where lower confidence values indicate possible anomalies. **Max-Logit** considers the maximum logit value before softmax normalization, providing an alternative confidence measure that can enhance detection reliability. **Max-Entropy** quantifies prediction uncertainty through the entropy of the softmax output, with higher entropy reflecting greater ambiguity and potential anomaly.

Model performance was evaluated using three key metrics:

- **Mean Intersection over Union (mIoU)**: In the context of semantic segmentation for road scenes, mIoU measures how well the model segments known object classes (e.g., roads, cars, pedestrians) by comparing predicted masks with ground truth annotations. It is computed as the average of the Intersection over Union for each class:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c}$$

where TP_c, FP_c, and FN_c denote the true positives, false positives, and false negatives for class c, and C is the total number of semantic classes. A higher mIoU indicates more accurate segmentation of standard road scene elements.

- **Area under the Precision-Recall Curve (AuPRC)**: For anomaly segmentation in driving scenarios, AuPRC evaluates the model's ability to detect unexpected or out-of-distribution objects (e.g., debris, animals, or foreign items not seen during training). It captures the balance between precision (correct identification of anomalies) and recall (detection coverage of all true anomalies):

$$\text{AuPRC} = \int_0^1 \text{Precision}(r)\, dr$$

where r represents the recall. A higher AuPRC signifies better anomaly localization with fewer false alarms.

- **False Positive Rate at 95% True Positive Rate (FPR95)**: This metric is crucial in safety-critical road applications where failing to detect an anomaly can be dangerous. FPR95 quantifies the proportion of normal (non-anomalous) pixels incorrectly flagged as anomalies when the model correctly identifies 95% of actual anomalous pixels:

$$\text{FPR}_{95} = \frac{FP_{95}}{FP_{95} + TN_{95}}$$

where FP_{95} and TN_{95} are the number of false positives and true negatives at 95% true positive rate. A lower FPR95 indicates a model that reliably distinguishes between normal and abnormal objects with minimal false detections.

Addiotionally since our target application involves real-time road scene understanding, we also report the model's **average inference time** per frame. This metric reflects the model's computational efficiency and responsiveness, which are critical for deployment in time-sensitive environments.

4.3 Temperature scaling

To improve the confidence calibration of anomaly predictions, temperature scaling was applied as a post-processing step on the model's logits. This technique adjusts the sharpness of the softmax output by dividing logits by a temperature parameter t, allowing better alignment between predicted probabilities and true uncertainties; since overconfident predictions can hinder reliable anomaly detection, tuning t helps mitigate this issue by smoothing confidence scores. We evaluated temperature scaling with the MSP method, testing multiple temperature values during inference to identify the optimal calibration.

4.4 Void Classifier

An alternative anomaly inference strategy was implemented by leveraging the void class present in the Cityscapes dataset. This approach is based on the principle that anomalies may correspond to classes absent from standard training labels and can thus be effectively identified through the void category. To preserve the semantic knowledge acquired during initial training and efficiently specialize the model for detecting the void class, we initialized from pre-trained networks (ENet, BiSeNet, and ERFNet) and continued training the model parameters. During training, we employed a weighted cross-entropy loss where the weight for the void class was computed based on its relative rarity in the training data.

$$w_{\text{void}} = \frac{1}{\ln(c + p_{\text{void}})} \tag{1}$$

where pvoid denotes the normalized frequency of the void class in the dataset, and c is a small constant to avoid division by zero.
This weighting strategy emphasizes the importance of correctly classifying the underrepresented void pixels, preventing the model from neglecting this class due to its low frequency. During inference, anomaly detection was performed by isolating the void class output, leveraging its specialized capacity to identify out-of-distribution or undefined regions.

4.5 Pruning and Quantization

To investigate potential reductions in model size and latency, we applied L1 unstructured pruning to the convolutional layers of the trained models. This pruning technique removes individual weights based on their absolute magnitude, zeroing out the less important connections without restricting to specific structures like entire filters or channels. After pruning, the reparameterization was removed to permanently zero out the pruned weights, effectively reducing the model's complexity. We tested different pruning fractions to observe how this trade-off impacts the segmentation performance and efficiency.However, since unstructured pruning maintains the original tensor shapes and most deep learning runtimes are optimized

for dense operations, the actual inference time remained largely unchanged. As a result, the effective reduction in computational cost was estimated in terms of FLOPs (floating point operations), which more accurately reflects the decrease in active computations. We tested different pruning fractions to observe how this trade-off impacts segmentation performance and model efficiency and estimated the potential inference time reduction that an optimized sparse runtime would bring.

In parallel, we explored post-training quantization to compress the models and accelerate inference. FX Graph Mode was specifically chosen to preserve the structure and behavior of our models without requiring architectural rewrites that could affect their semantic behaviour. Additionally, its compatibility with models structured in a modular and traceable way allowed us to apply it effectively to BiSeNet and ERFNet. ENet was not quantized due to its custom unpooling layers and index-passing operations, which are not supported by FX graph modules. Moreover, thanks to its shallow decoder design, ENet already prioritizes efficiency over fine detail reconstruction, partially achieving the benefits sought with quantization by trading some accuracy for speed.

5 Results and Discussion

5.1 Experimental Setup

All experiments were conducted using Google Colab as the development environment. Fine-tuning, and pruning and quantization were performed on CPU and TPU v4 runtimes, depending on availability.This setup allowed for efficient experimentation without requiring local hardware or high-end GPUs.

For the experiments, pre-trained weights were used to initialize BiSeNet and ENet models, both of which were trained on the Cityscapes dataset. The BiSeNet model was obtained from the official GitHub repository *https://github.com/CoinCheung/BiSeNet*, and the ENet weights were taken from the original implementation *https://github.com/davidtvs/PyTorch-ENet*.

5.2 Uncertainty-Based Anomaly Detection

Table 1 presents the performance of different uncertainty-based methods for out-of-distribution (OoD) detection across multiple datasets. The MaxLogit method consistently outperforms MSP and Max Entropy in terms of AuPRC across all datasets, particularly achieving the highest scores on FS Static (8.30) and Road Anomaly (15.58), while also maintaining competitive FPR95 values. Although Max Entropy sometimes provides slightly better FPR95 (e.g., on SMIYC RO-21 and FS L&F), MaxLogit strikes a better balance between precision and false positive rate, making it the most effective uncertainty-based method.

5.3 Temperature Scaling of MSP

Table 2 investigates the impact of temperature scaling on the MSP method. Adjusting the temperature parameter improves AuPRC performance on most datasets. For example, setting $t = 0.75$ increases the AuPRC from 2.71 to 2.57 on SMIYC RO-21 and from 1.75 to 1.49 on FS L&F. The best temperature scaling (labelled as "MSP (best t)") consistently improves performance over the baseline MSP, with a notable increase in FS Static (from 7.47 to 7.69 AuPRC) and Road Anomaly (from 12.42 to 12.66 AuPRC). This demonstrates that temperature scaling can enhance MSP's detection capabilities with only minor effect on FPR95.

5.4 Void Classification: Network Comparison

Table 3 compares the void classification performance of three semantic segmentation networks—ENet, ERFNet, and BiSeNet—using MaxLogit uncertainty estimation. This evaluation focuses specifically on the models' ability to identify out-of-distribution pixels via prediction of the void class, rather than overall semantic segmentation accuracy. BiSeNet shows the strongest void classification performance, achieving the highest AuPRC on datasets such as Road Anomaly (11.26) and maintaining relatively low FPR95 values. This indicates that BiSeNet is better at separating OoD content from known classes. While ENet is significantly faster and more lightweight, it lags behind in void classification performance, highlighting a trade-off between computational efficiency and robustness to unknown inputs.

5.5 Unstructured L1 Pruning: Performance vs. Speed Trade-off

Tables 4, 5, and 6 explore the effect of L1 pruning on ENet, ERFNet, and BiSeNet. Pruning leads to increased inference speed and reduced latency, with ENet achieving a 49.91× speed-up at 50% pruning (Table 4). However, aggressive pruning comes at the cost of reduced mIoU and degraded OoD performance. For ENet, pruning at 0.5 results in a drop in mIoU from 38.06% to 13.48%. Similarly, in BiSeNet (Table 6), mIoU declines from 60.74% to 36.41% at 0.5 pruning, although OoD performance remains relatively strong, suggesting that BiSeNet is more resilient to pruning compared to ENet or ERFNet.

5.6 Quantized Models: Efficiency vs. Performance

As we mentioned earlier, ENet could not be quantized using the FX Graph Mode quantization framework due to its use of internal layers that return tuple outputs, which are incompatible with the tracing-based quantization process. Consequently, we only evaluate quantized versions of ERFNet and BiSeNet.

Table 7 compares the performance and efficiency of these two models after quantization. ERFNet achieves the highest speed-up (94.40%), but this comes with a significant drop in mIoU (down to 41.40%). In contrast, BiSeNet maintains much better segmentation accuracy (60.73%) while still providing substantial efficiency gains. In terms of out-of-distribution detection, quantized BiSeNet also outperforms quantized ERFNet—for example, achieving a higher AuPRC on FS Static (5.02 vs. 1.08).

These results confirm that BiSeNet strikes a more favorable balance between speed and performance, making it more suitable for deployment in real-time semantic segmentation systems that require robust OoD detection.

5.7 Key Observations

- **Best uncertainty method:** MaxLogit, due to its superior AuPRC across diverse datasets.
- **Best temperature scaling:** Applying temperature scaling to MSP improves detection metrics without degrading FPR95.
- **Best network for void classification:** BiSeNet outperforms others in distinguishing OoD pixels, despite being slower than ENet.
- **Best pruning resilience:** BiSeNet maintains higher mIoU and better OoD scores under pruning compared to ENet and ERFNet.
- **Best quantized option:** Quantized BiSeNet balances speed and performance better than ERFNet.

Overall, the results suggest that combining temperature-scaled MaxLogit uncertainty estimation with a lightly pruned or quantized BiSeNet provides a robust and efficient solution for semantic segmentation and anomaly detection in real-world scenarios.

6 Conclusions and Future Works

In this work, we explored real-time anomaly segmentation for road scenes by leveraging model compression techniques—specifically, L1 unstructured pruning and post-training quantization. Our experiments showed that these methods can significantly reduce model size and computational requirements while maintaining competitive performance in both semantic segmentation and anomaly detection.

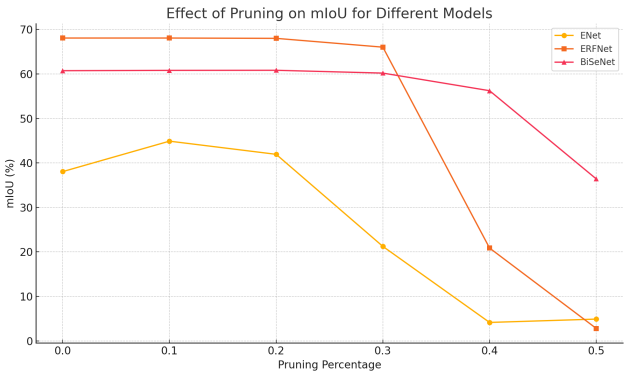

Figure 4: pruning effect on mIoU

Pruning proved effective in reducing the number of active weights, which led to notable improvements in estimated FLOPs and theoretical inference speed. However, due to the unstructured nature of pruning and the limitations of current hardware and software runtimes—especially on CPU and TPU—no substantial reduction in actual inference time was observed. This highlights the need for sparsity-aware inference engines or hardware accelerators that can exploit sparsity effectively.
Quantization using FX Graph Mode demonstrated strong potential, particularly for BiSeNet, which retained high segmentation accuracy while achieving substantial speed-ups. In contrast, ENet could not be quantized due to architectural constraints, although its inherently efficient design already provides a good trade-off between speed and accuracy.
Future work may consider hardware-aware training or deployment on platforms that natively support sparse computation (e.g., NVIDIA Ampere or FPGAs). However, we argue that structured pruning is not a promising direction for our setting. While it may offer better compatibility with existing inference frameworks, it requires restructuring the network architecture—such as removing entire filters or layers—which would compromise models already optimized for low latency. This would likely degrade performance and demand substantial reengineering effort, counteracting the very efficiency benefits we aim to preserve. Instead, further investigation into quantization-aware training or mixed-precision inference would be more practical paths forward for enhancing real-time anomaly segmentation.

References

[1] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," 2016.

[2] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," 2018.

[3] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with

guided aggregation for real-time semantic segmentation," 2020.

[4] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2018.

[5] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," 2018.

[6] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann, "Segmentmeifyoucan: A benchmark for anomaly segmentation," 2021.

[7] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "The fishyscapes benchmark: Measuring blind spots in semantic segmentation," *International Journal of Computer Vision*, vol. 129, p. 3119–3135, Sept. 2021.

Method	**mIoU %**	**SMIYC RA-21**		**SMIYC RO-21**		**FS L&F**		**FS Static**		**Road Anomaly**	
		AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
MSP	72.20	29.10	62.55	2.71	65.22	1.75	50.59	7.47	41.84	12.42	82.58
MaxLogit	72.20	38.32	59.34	4.63	48.44	3.30	45.49	9.50	40.30	15.58	73.25
Max Entropy	72.20	30.97	62.66	3.04	65.91	2.58	50.16	8.84	41.55	12.67	82.75

Table 1: Evaluation metrics (AuPRC and FPR@95) for different uncertainty methods across several anomaly detection datasets.

Method	mIoU %	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
		AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
MSP	72.20	29.10	62.55	2.71	65.22	1.75	50.59	7.47	41.84	12.42	82.58
MSP(t = 0.5)	72.20	27.06	62.73	2.42	63.23	1.28	66.74	6.60	43.48	12.19	82.02
MSP(t = 0.75)	72.20	28.16	62.49	2.57	64.13	1.49	51.76	6.99	42.50	12.32	82.31
MSP(t = 1.1)	72.20	29.40	62.65	2.76	65.87	1.86	50.18	7.69	41.62	12.46	82.73
MSP (best t)	72.20	29.40	62.65	2.76	65.87	1.86	50.18	7.69	41.62	12.46	82.73

Table 2: Evaluation metrics (AuPRC and FPR@95) for different temperature-scaled MSP methods across multiple anomaly detection datasets.

Network	mIoU %	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
		AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
ENet	44.84	14.18	85.48	1.49	51.65	3.71	75.09	5.17	43.00	8.32	79.48
ERF-Net	68.08	24.94	82.88	2.01	44.23	4.40	46.32	17.25	34.85	11.26	84.46
BiSeNet	60.74	22.74	69.62	4.66	22.25	10.05	31.62	5.39	49.77	13.80	82.75

Table 3: Performance of different semantic segmentation networks on anomaly detection benchmarks. Metrics include AuPRC and FPR@95.

Prun %	Inf. Time [s]	Speed-up %	mIoU %	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
				AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
0	0.0068	–	38.06	14.18	85.49	1.49	51.67	3.71	75.11	5.17	43.00	8.32	79.48
0.1	0.0059	13.45	44.90	14.20	84.45	1.28	43.28	3.83	73.67	5.21	42.46	8.42	80.53
0.2	0.0053	22.57	41.94	13.31	90.08	0.55	88.03	3.57	62.03	5.26	47.73	7.62	89.16
0.3	0.0047	31.67	21.24	12.43	93.77	0.40	98.90	2.04	70.53	4.24	59.37	7.35	94.72
0.4	0.0040	40.80	4.15	15.43	85.85	0.55	94.27	1.15	94.13	2.25	89.78	8.30	94.90
0.5	0.0034	49.91	4.90	13.48	89.97	0.57	91.64	0.32	94.69	2.74	84.52	7.30	97.52

Table 4: Effect of Unstructured L1 Pruning on Estimated Inference Time, mIoU, and OoD performance across multiple datasets using ENet. Inference time is measured in seconds per image (estimated).

Prun %	Inf. Time [s]	Speed-up %	mIoU %	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
				AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
0	0.0420	–	68.08	24.94	82.88	2.01	44.23	4.40	46.32	17.25	34.85	11.26	84.46
0.1	0.0380	9.52	68.08	25.04	82.78	2.04	44.31	4.39	46.92	17.31	34.38	11.29	84.18
0.2	0.0340	19.05	68.01	24.42	82.40	1.95	44.10	4.39	45.52	16.27	34.18	11.54	84.06
0.3	0.0300	28.57	66.03	23.41	80.87	2.34	42.12	3.83	41.93	19.18	31.56	9.33	86.12
0.4	0.0260	38.10	20.87	22.27	81.52	1.18	65.98	1.07	53.19	21.82	35.80	7.61	87.58
0.5	0.0220	47.61	2.80	16.35	89.56	0.74	90.25	0.54	73.96	3.12	86.35	4.12	91.33

Table 5: Effect of Unstructured L1 Pruning on Inference Time, mIoU, and OoD performance across multiple datasets using ErfNet.

Prun %	Inf. Time [s]	Speed-up %	mIoU %	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
				AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
0	0.0477	–	60.74	22.74	69.62	4.66	22.25	10.05	31.62	5.39	49.77	13.80	82.75
0.1	0.0429	9.99	60.81	22.65	69.61	4.70	21.53	9.96	31.86	5.48	49.15	13.71	82.95
0.2	0.0382	11.09	60.83	21.34	68.86	4.23	23.70	9.69	33.60	4.66	51.53	13.34	83.06
0.3	0.0334	22.18	60.20	20.94	71.88	4.34	21.09	8.96	37.01	4.69	50.97	12.85	86.06
0.4	0.0286	33.31	56.24	19.32	83.57	3.07	27.49	3.73	42.28	5.33	54.65	11.46	88.74
0.5	0.0239	44.38	36.41	12.72	86.23	1.08	79.97	1.36	70.65	2.61	88.08	10.08	89.37

Table 6: Effect of Unstructured L1 Pruning on Inference Time, mIoU, and OoD performance across multiple datasets using BiseNet.

Network	Speed-up %	mIoU %	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
			AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
ErfNet	94.40	41.40	21.57	86.26	0.77	80.97	1.55	68.07	14.08	35.92	10.83	86.33
BiSeNet	85.98	60.73	31.67	73.39	8.08	31.38	8.51	42.62	5.02	54.70	15.94	78.31

Table 7: Comparison of speed-up, segmentation accuracy (mIoU), and OoD detection metrics (AuPRC and FPR@95% TPR) across multiple datasets for quantized ErfNet and BiSeNet.