Real-Time Anomaly Segmentation for Road Scenes

Keyvan Delfarah
Politecnico di Torino
s300372@studenti.polito.it

Nima Shamanadi
Politecnico di Torino
s299869@studenti.polito.it

Armin Hooman
Politecnico di Torino
s302082@studenti.polito.it

Abstract

Deep neural networks (DNNs) for semantic segmentation of images are typically trained on a closed set of predefined in-distribution object classes, which contradicts the real world with limitless object types. Hence, detecting out-of-distribution examples is of paramount importance and has applications in detecting such as detecting novel biological phenomena and self-driving cars. This study addresses the challenge of anomaly segmentation in real-world scenarios and aims at building tiny anomaly segmentation models to segment anomalies that could be deployed in real-time leveraging the Cityscapes dataset for training. We survey baseline methods such as Maximum softmax probability (MSP), MaxLogit, and Max Entropy using a pre-trained ERF-Net model. Next, temperature scaling is implemented to consider the impact of T parameter variation on OoD object detection performance. While, all the aforementioned models performances are measured by AuPRC and FPR95, the ability of the model for in-distribution object detection and image classification measure by mean intersection over union. The results indicate that anomaly detection is extremely degraded in case of domain shift.

1. Introduction

In recent years, spectacular advances in computer vision task semantic segmentation have been achieved by deep learning [15]. In real-world applications, deep convolutional neural networks (CNNs) are expected to be confronted with input that differs significantly from the data used to train the model. [3]. Although cutting-edge deep neural networks (DNNs) perform exceptionally well on trained datasets, they usually only offer predictions for a limited number of semantic classes [4]. As such, they cannot assign an item to any existing categories [17]. There are an infinite number of possible items in open-world environments [3]. A significant quantity of annotated data is needed to define extra classes, and doing so could result in performance declines [5].

Adding a none-of-the-known output for items that do not fit into any of the specified classes is counted as a logical approach. [17] . In another way, one handles OoD (Out-of-Distribution) objects by imposing an alternate model output for such samples and employing a set of object classes that suffice for most cases. From the perspective of functional safety, neural networks must be able to consistently detect objects outside of their intended domain to trigger a fallback strategy. [3]

Only a small portion of everyday images might be out-of-distribution (OoD) items. It is crucial for practical use to pinpoint these OoD objects. Hence, significant research interest is aimed at identifying abnormal regions within images, indicating the presence of foreign objects in the scene [2], [3]. Applications that have to deal with the diversity of the actual world, like perception in automated driving and dependability in the presence of unfamiliar things, are essential to their success. [4].

This research aims to build small anomaly segmentation models that can be used in real-time, especially in areas like computer vision, constant learning, and autonomous driving. The goal is to create models that work with memory limitations to serve edge applications that use intelligent cameras with low computing power on board. The ERF-Net model is used in this project; it is available on GitHub at Link and was trained on the Cityscapes dataset. The pre-trained model is also used to perform tests on different datasets to make different anomalous inferences. These tests include Fishyscapes Lost & Found, Fishyscapes Online, Road Anomaly, Road Anomaly21, and Road Obstacle21. Fishyscapes Online is an open-world setup where Cityscapes photographs are overlaid with components that are commonly visited online.

We start this project by evaluating relevant work, techniques, and benchmarks in the following sections. All dataset tests are then run using three distinct approaches (MSP, MaxLogit, and Max Entropy), and the outcomes are contrasted. In conclusion, we investigate using the temperature scaling approach to identify the dataset tests that yield the best anomaly segmentation outcomes.

2. Related Work

Significant progress has been made in the area of semantic segmentation in the last several years. The main focus of this section is on important explanation and approaches that are especially relevant to our study.

2.1. Semantic Segmentation

Extract meaningful information from images or input from a video or recording frame. It is the way to perform the extraction by checking pixels by pixel using a classification approach [16]. In another words, the ability to divide an unknown scene into separate elements and objects—beach, ocean, sun, dog, and swimmer, for example—is known as semantic segmentation. Furthermore, segmentation is superior to object recognition since segmentation does not require recognition. Interestingly, humans are capable of segmenting images without knowing what items are in them beforehand [8].

2.2. Real-time Semantic Segmentation

Real-time semantic segmentation algorithms demand rapid generation of high-quality predictions. Several models serve as foundational frameworks in the domain of real-time segmentation, including SegNet, E-Net, and ICNet [16]. Despite their ability to achieve real-time inference speeds, our study employs the ERF-Net model, pertained on the cityscape dataset, for our specific investigation.

2.3. State-of-the-art models

Semantic segmentation cutting-edge models primarily use fully-convolutional deep networks that are trained via pixel-wise supervision. These models often use an encoder-decoder architecture, whereby feature maps are first reduced in spatial resolution and then up-sampled using methods such fixed bilinear interpolation, learnt transposed convolution, or unpooling [2] . The ERF-Net presents a deep model architecture that is intended to do real-time activities while providing accurate semantic segmentation. This architecture is based on a new layer that uses factorized convolutions and residual connections to provide efficiency without compromising exceptional accuracy (Bergasa, et al. 2018).

2.4. Anomaly Segmentation

Methods designed to estimate the uncertainty of a model for a given input may mistakenly attribute high uncertainty to non-anomalous inputs, despite variability and possible confounders such as high input noise. Nonetheless, anomaly detection continues to be a widely used benchmark technique for estimating uncertainty, based on the idea that unusual inputs ought to naturally provide greater uncertainty than any training data [6].

2.5. Pixel-wise OoD Detection

The techniques used in Pixel-wise Out-of-Distribution (OoD) Detection include freezing or retraining the segmentation model. Traditionally, the segmentation model was frozen and OoD pixels were classified using distance metrics like Mahalanobis or measures of posterior distribution like entropy. In order to improve OoD detection accuracy, recent improvements have included incorporating other networks; nevertheless, these techniques may generate confirmation bias. On the other hand, OoD identification can be enhanced by retraining the segmentation model with OoD images from outlier datasets; however, this would decrease the accuracy of closed-set segmentation. Creating techniques that reliably identify abnormalities while reducing their influence on the performance of inlier segmentation is the difficult part [13].

3. Underlying Architecture

The ERFNet (Efficient Residual Factorized ConvNet) architecture, introduced in the paper "ERFNet: Efficient Residual Factorized ConvNet for Real-time Semantic Segmentation" by Romera et al. (2017), is tailored for efficient semantic segmentation of images. Key architectural features include factorized convolutions, residual connections, a bottleneck design, an encoder-decoder architecture, and efficient pooling operations.

Factorized convolutions, combining 1x1 and 3x3 convolutions, are utilized to reduce computational complexity while retaining spatial information. This design choice enables efficient parameter reduction without sacrificing performance. Moreover, residual connections are incorporated to facilitate gradient flow during training, thereby enhancing learning efficiency.

The bottleneck design, reminiscent of ResNet, comprises sequences of 1x1 convolutions for channel reduction, a central 3x3 convolution layer, and subsequent 1x1 convolutions for channel restoration. This design not only minimizes computational burden but also preserves expressive power. [12]

The encoder-decoder architecture is employed, with the encoder extracting features from input images and the decoder upscaling these features to generate segmentation maps. Skip connections between corresponding encoder and decoder layers are employed to fuse high-resolution information during upsampling.

Efficient pooling operations, including average pooling, are utilized to reduce spatial dimensions while mitigating information loss. This choice further contributes to the overall efficiency of the network.

ERFNet is characterized as fully convolutional, allowing it to process images of variable sizes and produce corresponding segmentation maps. This property renders it suitable for real-time applications where computational efficiency and flexibility are paramount considerations. Fig. 1

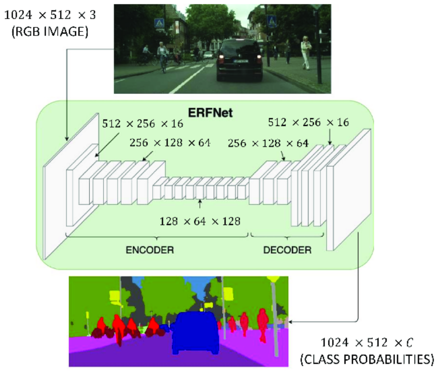


Figure 1. The diagram illustrates the proposed segmentation system (ERFNet) , showcasing an example input image alongside its corresponding output (with C = 19 classes). The depicted volumes represent the feature maps generated by each layer. Although all spatial resolution values are based on the example input (1024×512), the network is capable of processing images of arbitrary sizes. (Credit: Romera et al., 2018 [14])

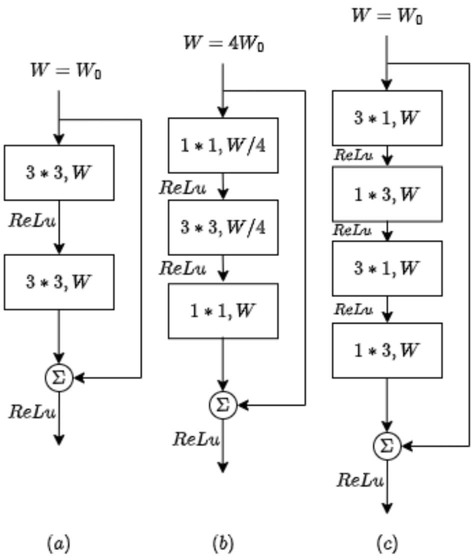


Figure 2. Depiction of the two residual layers proposed in [9], (a) Non-bottleneck. (b) Bottleneck. (c) Non-bottleneck-1D (Non-bottleneck and Bottleneck)

4. OoD Detection Methodologies

In this section, we first provide a brief introduction to the methods evaluated in our benchmark, which comprise our initial leader board. Following this, we provide additional technical details on the introduced methods.

Given an input image, the **Maximum Softmax Probability (MSP)**, refers to the highest probability assigned by the softmax function to any class label for a given input sample. **Max Logit**, refers to the highest logit value among all the class logits for a given input sample, where logits are the raw, unnormalized outputs of a neural network before applying the softmax function. Speaking of randomness, Softmax entropy is a measure of uncertainty in the predicted probability distribution over classes. **Maximized Softmax Entropy**, involves training a model to produce predictions with a high level of uncertainty, that is to acknowledge uncertainty in the model's predictions.

4.1. Maximum Logit

Let $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$ represent the input image and C denote the number of pre-defined classes, where H and W denote the image height and width, respectively. The logit output $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ can be obtained from the segmentation network before the softmax layer. Subsequently, the max logit $\mathbf{L} \in \mathbb{R}^{H\times W}$ and prediction $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times W}$ at each location h, w are defined as [12]:

$$\mathbf{L}_{h,w} = \max_c \mathbf{F}_{c,h,w} \tag{1}$$

$$\hat{\mathbf{Y}}_{h,w} = \text{argmax}_c \mathbf{F}_{c,h,w} \tag{2}$$

4.2. Maximized Softmax Probability (MSP)

Let $f(x) \in (0,1)^q$ denote the softmax probabilities of the input image $x \in X$ trained with a DNN model $f : X \to (0,1)^q$, and let $q = |C| \in \mathbb{N}$ denote the number of classes. Furthermore, let $f_z(x) \in (0,1)^{|C|}$ denote the pixel-wise softmax probability at pixel location $z \in Z$ as . The metric yields the OoD score for each pixel $z \in Z$:

$$1 - \max_{j\in C} f_j^z(x) = 1 - f^z_{\hat{c}(z)}(x), \quad x \in X \tag{3}$$

4.3. Maximized Softmax Entropy

Referring to Sec. 4.2, holding the same mathematical definition, for a given softmax probability, $f(x) \in (0,1)^q$, the softmax entropy can be determined as follows:

$$E(f(x)) = -\sum_{j\in C} f_j(x) \log(f_j(x)) \tag{4}$$

We denote an "in-distribution" example as $(x, y(x)) \sim D_{\text{in}}$, where $y(x) \in C$ represents the ground truth class label of input x, and an "out-distribution" example as $x_o \sim D_{\text{out}}$, where no ground truth label is given. Our objective is to minimize the overall loss function. We define L as the objective function:

$$\begin{aligned} L :=& (1-\lambda)\mathbb{E}_{(x,y)\sim D_{\text{in}}}[\ell_{\text{in}}(f(x), y(x))] \\ &+ \lambda\mathbb{E}_{x_0\sim D_{\text{out}}}[\ell_{\text{out}}(f(x_0))], \quad \lambda \in [0,1] \end{aligned} \tag{5}$$

where

$$\ell_{\text{in}}(f(x), y(x)) := -\sum_{j \in C} \mathbf{1}_{j=y(x)} \log(f_j(x)) \tag{6}$$

$$\ell_{\text{out}}(f(x_0)) := -\sum_{j \in C} \frac{1}{q} \log(f_j(x_0)) \tag{7}$$

For in-distribution samples, we apply the commonly used empirical cross-entropy loss, which is the negative log-likelihood of the target class. The indicator function $\mathbf{1}_{j=y(x)}$ takes values in $\{0, 1\}$, equal to one if $j = y(x)$ and zero otherwise. In the context of semantic segmentation, one aims to minimizes the averaged pixel-wise classification loss over the image, cf. Eq. (5) [3]. For a given pixel location z, it is considered to be out of distribution if the normalized entropy, $\bar{E}(f^z(x))$ is greater than a specific threshold $t \in [0, 1]$, i.e. z is predicted to be OoD if:

$$\hat{\mathcal{Z}}_{\text{out}}(x) := \{z_o \in \mathcal{Z} : \bar{E}(f^{z_o}(x)) \geq t\} \tag{8}$$

5. Model Evaluation Metrics

Assuming a binary classifier model providing scores for an image $s(x) \in \mathbb{R}^Z$, in which x is an image in the $\mathcal{X}$ image set, and Z signifies the set of image pixel locations. The model diagnoses between the anomaly and non-anomaly classes. The precision and recall are defined accordingly:

$$\text{precision}(\delta) = \frac{|\mathcal{Y}_{c1} \cap \hat{\mathcal{Y}}_{c1}(\delta)|}{|\hat{\mathcal{Y}}_{c1}(\delta)|}, \quad \text{recall}(\delta) = \frac{|\mathcal{Y}_{c1} \cap \hat{\mathcal{Y}}_{c1}(\delta)|}{|\mathcal{Y}_{c1}|}$$

Where $\mathcal{Y} \subseteq \{\text{“anomaly”}, \text{“not anomaly”}\}^{N \times \mathbb{Z}}$ and $\hat{\mathcal{Y}}(\delta)$ are respectively the pixel level ground truth and predicted labels and N indicates number of images. It should be underlined that the prediction label is obviously dependent on the threshold δ value. In the equations of this section, c_1 and c_2 denote the anomaly and non-anomaly class, respectively.

5.1. FPR95

FPR95, or False Positive Rate at 95% True Positive Rate, is a specialized performance metric utilized in anomaly detection and segmentation. It quantifies the proportion of normal instances wrongly classified as anomalies (FPR) when the model correctly identifies 95% of the anomalies (TPR for anomaly class) [4]. Thus, the anomaly class is calculated as :

$$FPR_{95} = \frac{|\hat{\mathcal{Y}}_{c1}(\delta') \cap \mathcal{Y}_{c2}|}{|\mathcal{Y}_{c2}|} \quad \text{s.t.} \quad \frac{|\mathcal{Y}_{c1} \cap \hat{\mathcal{Y}}_{c1}(\delta')|}{|\mathcal{Y}_{c1}|} = 0.95,$$

5.2. AuPRC

The Precision-Recall Curve is a graph that depicts the trade-off between precision and recall for different threshold values. The AuPRC itself is the integral of the precision-recall curve. The curve involves the the precision-recall tuples for different threshold values. Thus, it can be concluded that oppose to the FPR95 which is a thresholdependent performance metric , AuPRC is a threshold independent evaluation metric method which makes it a more robust when the threshold value is uncertain.

6. Experimental Setup

6.1. Datasets

6.1.1 Trained Model

Cityscapes: The underlying model for the task of semantic segmentation is a DNN pre-trained model on the Cityscapes dataset. The Cityscapes dataset is a contemporary collection of urban scenes, renowned for its broad range of scenarios and intricate set of 19 labeled classes, making it a prominent choice for semantic segmentation benchmarks. It comprises a train set consisting of 2975 images, a validation set containing 500 images, and a test set comprising 1525 images. [14] The accuracy of the results for the trained model are examined using the commonly adopted Intersection-over-Union (IoU) metric:

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

where, TP, FP, and FN represent the number of true positives, false positives, and false negatives at the pixel level, respectively. Mean Intersection over Union (mIoU) consists of comparing the overlap between the predicted and ground truth segmentation mask, in particular, is calculated by dividing the area of overlap between the predicted and ground truth mask by the area of union between the two partitions [3]. Then it is computed for each class, and in the end, averaged for all of them.Fig. 3 and Tab. 1, represent IoU per each class of the trained model along with the mean IoU.

6.1.2 Evaluation Dataset

RoadAnomaly21: The road anomaly track evaluates general anomaly segmentation in full street scenes, where various road surfaces can be observed with a dataset of 100 images annotated at the pixel level. [4]
RoadObstacle21: The road obstacle track is dedicated to enhancing safety in automated driving. The evaluation data emphasizes objects that consistently manifest on the road ahead, representing realistic and potentially hazardous obstacles crucial for detection.. [4]
Fishyscapes LostAndFound: The Fishyscapes LostAndFound (FS L&F) validation dataset comprises 100 images from the original LostAndFound data, featuring refined labels. These labels allow anomalous objects to appear not only on the road but also anywhere in the image, aligning with our benchmark's anomaly track. [4]

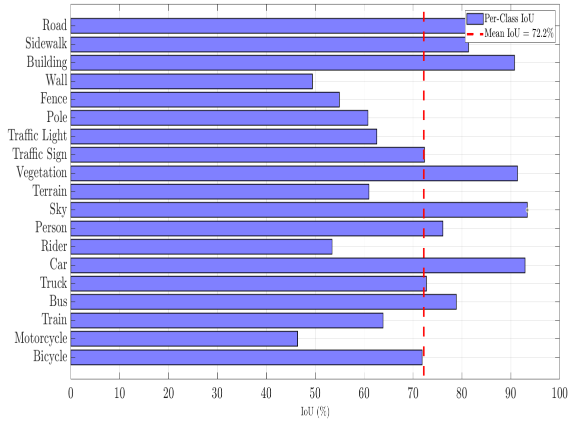


Figure 3. Per-class Intersection over Union(IoU) & Mean IoU(mIoU) of the trained model classes

Class	IoU (%)
Road	97.62
Sidewalk	81.37
Building	90.77
Wall	49.43
Fence	54.93
Pole	60.81
Traffic Light	62.60
Traffic Sign	72.32
Vegetation	91.35
Terrain	60.97
Sky	93.38
Person	76.11
Rider	53.45
Car	92.91
Truck	72.78
Bus	78.87
Train	63.86
Motorcycle	46.41
Bicycle	71.89
Mean IoU	**72.20**

Table 1. Per-Class IoU

RoadAnomaly: The dataset comprises images depicting uncommon hazards encountered by vehicles on the road, such as animals, rocks, traffic cones, and other obstacles. Its primary aim is to evaluate the effectiveness of autonomous driving perception algorithms in handling infrequent yet safety-critical scenarios.

Each image in the dataset is accompanied by per-pixel labels, facilitating precise annotation of the depicted hazards. The labeling process was conducted using our LabelGrab tool, ensuring accuracy and consistency in the dataset annotations. [1]

7. Model Evaluation

7.1. Baseline

The results of the considered models implemented on the testing images namely SMIYC RA-21, SMIYC RO-21, FS L&F, FS Static are provided and elaborated in this section. The results are summarized in Tabs. 2 to 4. Overall, the methods identified in the model evaluation section, are originally developed for image classification including Mas Soft Max Probability (MSP), Max logit, and Max Entropy. At first, Authors in [11], proposed the simple baseline of using the MSP of the classifier on an input to gauge whether the input is OoD. According to [10], MSP was basically developed for image classification and its performance in terms of AuPRC and FPR95 is not well generalized. [10] mentioned that MSP was originally developed for small-scale problems and does not scale well to the challenging conditions. They also underlined that MSP's reliance on softmax probabilities makes it unsuitable for multi-label data, which is a common scenario in real-world applications. Reaching the same conclusion of the aforementioned studies and as can be observed from Tabs. 2 to 4, the AuPRC for MSP is 14.7, 0.68%, 0.27%, 2.04%, and 9.71% for SMIYC RA-21, SMIYC RO-21, FS L&F, FS Static, and Road Anomaly which obviously indicates the poor performance of the module, considering the fact that even a random binary classifier with no skill expectedly reaches 50% in terms of AuPRC. Regarding the FPR95, the lower value of this parameter signifies the better performance of the model, where a random binary classifier with no skill expectedly reaches FPR95 of 95%. Meanwhile, MSP leads to approximately 95% FPR95 for all the evaluated dataset. FPR95 results of MSP emphasize the poor performance of the MSP module.

[11] introduced the anomaly detector tool based on maximum logit, which is named Maxlogit module in this study. Maxlogit outperforms MSP in large-scale multi-class, multi-label, segmentation tasks. The MaxLogit detector places lower scores on in-distribution image regions, including object outlines, while also doing a better job of highlighting anomalous objects compared to MSP. However, AuPRC and FPR95 do not show notable improvement in comparison with MSP. More specifically, in FS L&F and FS Static, and SMIYC RA-21, and Road Anomaly, Maxlogit even showed slightly less accurate results. Regarding the other datasets, SMIYC RO-21 is the only dataset in which Maxlogit outperforms the MSP with 1.15%. The FPR95 related results indicate not less than 95% (random binary classifier) unless SMIYC RO-21 and

Road anomaly datasets with 86.8% and 96.4%, respectively. Thus, despite the claimed better performance of MaxLogit with respect to its predecessors in segmentation tasks, Maxlogit AuPRC and FPR95 are not sufficient.

Method	SMIYC RA-21		SMIYC RO-21	
	AuPRC	FPR95	AuPRC	FPR95
MSP	14.744	95.028	0.682	94.959
MaxLogit	13.193	97.015	1.153	86.816
Max Entropy	14.646	94.974	0.685	96.746

Table 2. Performance Comparison of RoadAnomaly21 and Road-Obstacle21 validation dataset

Method	FS L&F		FS Static	
	AuPRC	FPR95	AuPRC	FPR95
MSP	0.275	95.207	2.043	95.109
MaxLogit	0.214	96.441	1.645	96.461
Max Entropy	0.260	96.526	2.000	95.131

Table 3. Performance Comparison of Fishyscapes LostAndFound and Fishyscapes static validation dataset

Method	Road Anomaly	
	AuPRC	FPR95
MSP	9.715	95.095
MaxLogit	8.708	93.764
Max Entropy	9.601	93.780

Table 4. Performance Comparison of Road Anomaly validation dataset

Likewise the previously induce methods max entropy tunes previously trained to the task of anomaly segmentation. In this method a loss function specifically defined for the OoD data (L_{out}) is added to the conventional cross entropy loss function to establish a combined loss function to simultaneously take into account D_{in} and D_{out}. Moreover, this method has the potential to take OoD data during the training process in addition to the regular in-distribution cityscape dataset. This enables the method to effectively predict the OoD object by determining a normalize entropy to each pixel and comparing it with the pre-defined threshold. The method assigns OoD label to the pixels with higher than the pre-defined threshold [3]. It is noteworthy that the authors in [3] utilized OoD data of COCO dataset during training the data which might be the main reason for better OoD detection results in [4]. Meanwhile, as our model is trained only by cityscape dataset, the entropy base method does not lead to improved results in terms of AuPRC and FPR95, Tabs. 2 to 4. More in detail, the max entropy results are similar to that of MSP except for FPR95 in SMIYC RO-21 and Road Anomaly with 96.70% and 93.78%, respectively.

7.2. Model Calibration

7.2.1 Overview

In real-world decision-making systems, classification networks are required to be not only accurate but also capable of indicating their uncertainty. When a detection network fails to confidently predict the presence or absence of immediate obstacles, the car must rely more on the outputs of other sensors for braking. Therefore, a network should provide a calibrated confidence measure alongside its prediction. This means that the probability associated with the predicted class label should accurately reflect the likelihood of its correctness. Regarding the importance of calibration, various calibration approaches exist, all of which are post-processing steps aimed at producing calibrated probabilities. Each method necessitates a hold-out validation set, which in practice can be the same set used for hyper parameter tuning. [7]

7.2.2 Temperature Scaling

Temperature scaling is a post-processing technique to make neural networks calibrated. This method effectively calibrates modern neural networks by adjusting the softmax temperature to align predictions with actual outcomes. Implementing the temperature scaling the output of neural network will be more confident. Temperature scaling utilizes a single scalar parameter T $>$ 0 for all the in-distribution classes. This scaling parameter softens the softmax probability acquired through MSP method (f(x)).

$$q_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}} \tag{10}$$

When T tends to infinite the probability distribution of the in-distribution classes will be uniform which is equivalent to evaluate randomly. On the other hand, when T limits to 0, one of the classes would be assigned as with the definite prediction (i.e. $\hat{q}_i$ = 1). [7] This enhances model reliability without sacrificing accuracy, crucial for applications needing dependable probability estimates and boosting confidence in decision-making based on neural network predictions. In this study, the aim is to explore temperature scaling for model calibration and assesses the AuPRC and FPR95 metrics on all validation datasets by using MSP as a method to evaluate anomaly detection results. The final evaluation, presented in Tabs. 2 to 4, is conducted without incorporating the temperature parameter. Specifically, the selected parameters for this evaluation are 0.5, 0.75, and 1.1 respectively.

Opposed to the expected performance of temperature scaling, post processing method does not have noticeable implications on the OoD detection results, Tabs. 4 to 6 indicate this statement. In summary, neither of the assessed parameters indicate a considerable variation with respect to the baseline (MSP, t=1).

Method	SMIYC RA-21		SMIYC RO-21	
	AuPRC	FPR95	AuPRC	FPR95
MSP	14.744	95.028	0.682	94.959
MSP (t=0.5)	14.775	95.014	0.679	94.980
MSP (t=0.75)	14.759	95.021	0.680	94.970
MSP (t=1.1)	14.738	95.031	0.683	94.956

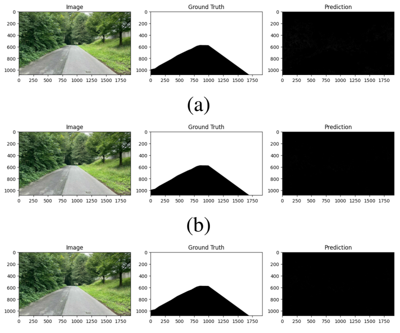
Table 5. Performance Comparison of RoadAnomaly21 and RoadObstacle21 validation dataset using Temperature scaling

Method	FS L&F		FS Static	
	AuPRC	FPR95	AuPRC	FPR95
MSP	2.043	95.109	2.043	95.109
MSP (t=0.5)	0.279	95.105	2.061	95.061
MSP (t=0.75)	0.277	95.156	2.052	95.087
MSP (t=1.1)	0.2746	95.2268	2.040	95.117

Table 6. Performance Comparison of Fishyscapes LostAndFound and FS Static validation dataset using Temperature scaling

Method	Road Anomaly	
	AuPRC	FPR95
MSP	9.715	95.095
MSP (t=0.5)	9.780	95.048
MSP (t=0.75)	9.747	95.072
MSP (t=1.1)	9.702	95.104

Table 7. Performance Comparison of Road Anomaly using Temperature scaling

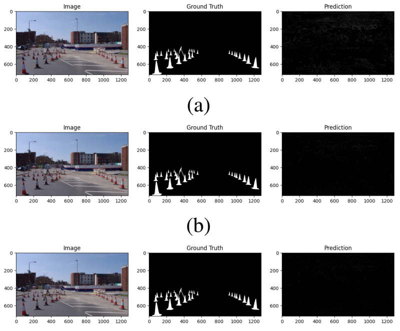


(a)

(b)

(c)

Figure 4. Comparison of (a) MaxLogit, (b) MaxEntropy, and (c) MSP for RoadObstacle21



(a)

(b)

(c)

Figure 5. Comparison of (a) MaxLogit, (b) MaxEntropy, and (c) MSP for RoadAnomaly

8. Conclusion

The application of deep convolutional neural networks (CNNs) has significantly increased the precision and throughout of semantic segmentation tasks; nonetheless, issues still arise, particularly with regard to addressing out-of-distribution (OoD) objects and ensuring robust performance in real-world scenarios. In this work, we have investigated three proposed methods namely MSP, MAX Logit, and Max Entropy for the task of OoD object detection, and concluded that Max Entropy outperforms Max Logit and MSP in the scientific literature experimental results. Despite the acceptable mIoU results of the ERF-Net model that signifies sufficient in-distribution image classification accuracy, models indicated poor performance for OoD object detection. The rational behind this performance can be elaborated accordingly: 1-domain shift between the training and validation datasets. 2- Lack of OoD labelled dataset during training process. 3- The majority of the state-of-the-art models (such as MSP) merely perform well on the small-scale datasets. Regarding the temperature scaling, variations of the T value had minor effects on the AuPR and FPR95, unexpectedly. For the future work, do to the fact that training dataset lacking OoD objects are not able to exhibit satisfying performance, application of OoD data through the training step might be a potentially solution to this issue which can be further investigated.

References

[1] Road anomaly dataset. https://www.epfl.ch/labs/cvlab/data/road-anomaly/. Retrieved on [insert retrieval date]. 5

[2] 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*, 35:3119–3135, 2021. 1, 2

[3] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5108–5117, 2021. 1, 4, 6

[4] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Segmentmeifyoucan: A benchmark for anomaly segmentation. *arXiv*, pages 1–35, 2021. 1, 4, 6

[5] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Lecture Notes in Computer Science*, pages 71–84. Springer, Berlin, 2010. 1

[6] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 16913–16922, 2021. 2

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. 6

[8] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, pages 87–93, 2018. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[10] 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, 2022. 5

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018. 5

[12] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation, 2021. 2, 3

[13] Yuyuan Liu et al. Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1151–1161. IEEE/CVF, 2023. 2

[14] Eduardo Romera, José M. Álvarez, 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, 2018. 3, 4

[15] Jingdong Wang et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 3349–3364, 2021. 1

[16] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, pages 3051–3068, 2021. 2

[17] Xiang Zhang and Yann LeCun. Universum prescription: Regularization using unlabeled data. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1