Pixel-Wise Gradient Uncertainty for Convolutional Neural Networks Applied to Out-of-Distribution Segmentation

Kira Maag* and Tobias Riedlinger*

Technical University of Berlin, Germany

Abstract: In recent years, deep neural networks have defined the state-of-the-art in semantic segmentation where their predictions are constrained to a predefined set of semantic classes. They are to be deployed in applications such as automated driving, although their categorically confined expressive power runs contrary to such open world scenarios. Thus, the detection and segmentation of objects from outside their predefined semantic space, i.e., out-of-distribution (OoD) objects, is of highest interest. Since uncertainty estimation methods like softmax entropy or Bayesian models are sensitive to erroneous predictions, these methods are a natural baseline for OoD detection. Here, we present a method for obtaining uncertainty scores from pixel-wise loss gradients which can be computed efficiently during inference. Our approach is simple to implement for a large class of models, does not require any additional training or auxiliary data and can be readily used on pre-trained segmentation models. Our experiments show the ability of our method to identify wrong pixel classifications and to estimate prediction quality at negligible computational overhead. In particular, we observe superior performance in terms of OoD segmentation to comparable baselines on the SegmentMeIfYouCan benchmark, clearly outperforming other methods.

1 INTRODUCTION

Semantic segmentation decomposes the pixels of an input image into segments which are assigned to a fixed and predefined set of semantic classes. In recent years, deep neural networks (DNNs) have performed excellently in this task (Chen et al., 2018; Wang et al., 2021), providing comprehensive and precise information about the given scene. However, in safety-relevant applications like automated driving where semantic segmentation is used in open world scenarios, DNNs often fail to function properly on unseen objects for which the network has not been trained, see for example the bobby car in Figure 1 (top). These objects from outside the network's semantic space are called *out-of-distribution* (OoD) objects. It is of highest interest that the DNN identifies such objects and abstains from deciding on the semantic class for those pixels covered by the OoD object. Another case are OoD objects which might belong to a known class, however, appearing differently to substantial significance from other objects of the same class seen during training. Consequently, the respective predictions are prone to error. For these
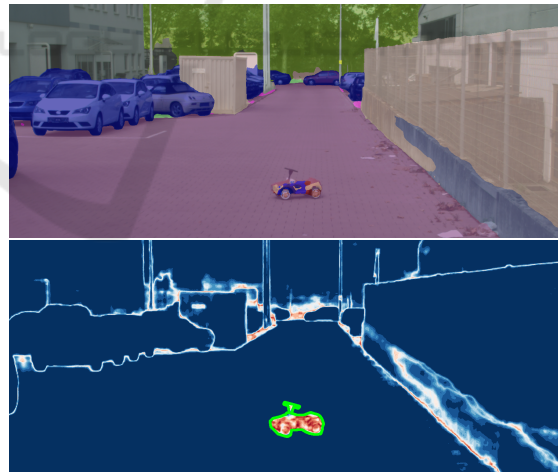
*Equal contribution

Figure 1: Top: Semantic segmentation by a deep neural network. Bottom: Gradient uncertainty heatmap obtained by our method.

objects, as for classical OoD objects, marking them as OoD is preferable to the likely case of misclassification which may happen with high confidence. This additional classification task should not substantially degrade the semantic segmentation performance itself outside the OoD region. The computer vision tasks

of identifying and segmenting those objects is called *OoD segmentation* (Chan et al., 2021a; Maag et al., 2022).

The recent contributions to the emerging field of OoD segmentation are mostly focused on OoD training, i.e., the incorporation of additional training data (not necessarily from the real world), sometimes obtained by large reconstruction models (Biase et al., 2021; Lis et al., 2019). Another line of research is the use of uncertainty quantification methods such as Bayesian models (Mukhoti and Gal, 2018) or maximum softmax probability (Hendrycks and Gimpel, 2016). Gradient-based uncertainties are considered for OoD detection in the classification task (Huang et al., 2021; Lee et al., 2022; Oberdiek et al., 2018) but up to now, have not been applied to OoD segmentation. In (Grathwohl et al., 2020), it is shown that gradient norms perform well in discriminating between in- and out-of-distribution. Moreover, gradient-based features are studied for object detection to estimate the prediction quality in (Riedlinger et al., 2023). In (Hornauer and Belagiannis, 2022), loss gradients w.r.t. feature activations in monocular depth estimation are investigated and show correlations of gradient magnitude with depth estimation accuracy.

In this work, we introduce a new method for uncertainty quantification in semantic segmentation based on gradient information. Magnitude features of gradients can be computed at inference time and provide information about the uncertainty propagated in the corresponding forward pass. The features represent pixel-wise uncertainty scores applicable to prediction quality estimation and OoD segmentation. An exemplary gradient uncertainty heatmap can be found in Figure 1 (bottom). Calculating gradient uncertainty scores does not require any re-training of the DNN or computationally expensive sampling. Instead, only one backpropagation step for the gradients with respect to the final convolutional network layer is performed per inference to produce gradient scores. Note, that more than one backpropagation step can be performed to compute deeper gradients which considers other parameters of the model architecture. An overview of our approach is shown in Figure 2.

An application to dense predictions such as semantic segmentation has escaped previous research, potentially due to the superficial computational overhead. Single backpropagation steps per pixel on high-resolution input images quickly become infeasible given that around 10^6 gradients have to be calculated. To overcome this issue, we present a new approach to exactly compute the pixel-wise gradient scores in a batched and parallel manner applicable to a large class of segmentation architectures. This is possible due to a convenient factorization of p-norms for appropriately factorizing tensors, such as the gradients for convolutional neural networks. We use the computed gradient scores to estimate the model uncertainty on pixel-level and also the prediction quality on segment-level. Moreover, the gradient uncertainty heatmaps are investigated for OoD segmentation where high scores indicate possible OoD objects. We demonstrate the efficiency of our method in explicit runtime measurements and show that the computational overhead introduced is marginal compared with the forward pass. This suggests that our method is applicable in extremely runtime-restricted applications of semantic segmentation.

For our method, we only assume a pre-trained semantic segmentation network ending on a convolutional layer which is a common case. In our experiments, we employ a state-of-the-art semantic segmentation network (Chen et al., 2018) trained on Cityscapes (Cordts et al., 2016) evaluating in-distribution uncertainty estimation. We demonstrate OoD detection performance on four well-known OoD segmentation datasets, namely LostAndFound (Pingera et al., 2016), Fishyscapes (Blum et al., 2019a), RoadAnomaly21 and RoadObstacle21 (Chan et al., 2021a). The source code of our method is made publicly available at https://github.com/tobiasriedlinger/uncertainty-gradients-seg. We summarize our contributions as follows:

- We introduce a new gradient-based method for uncertainty quantification in semantic segmentation. This approach is applicable to a wide range of common segmentation architectures.

- For the first time, we show an efficient way of computing gradient norms in semantic segmentation on the pixel-level in a parallel manner making our method far more efficient than sampling-based methods which is demonstrated in explicit time measurements.

- We demonstrate the effectiveness of our method to predictive error detection and OoD segmentation. For OoD segmentation, we achieve area under precision-recall curve values of up to 69.3% on the LostAndFound benchmark outperforming a variety of methods.

2 RELATED WORK

Uncertainty Quantification. Bayesian approaches (MacKay, 1992) are widely used to estimate model uncertainty. The well-known ap-

proximation, MC Dropout (Gal and Ghahramani, 2016), has proven to be computationally feasible for computer vision tasks and has also been applied to semantic segmentation (Lee et al., 2020). In addition, this method is considered to filter out predictions with low reliability (Wickstrøm et al., 2019). In (Blum et al., 2019a), pixel-wise uncertainty estimation methods are benchmarked based on Bayesian models or the network's softmax output. Uncertainty information is extracted on pixel-level by using the maximum softmax probability and MC Dropout in (Hoebel et al., 2020). Prediction quality evaluation approaches were introduced in (DeVries and Taylor, 2018; Huang et al., 2016) and work on single objects per image. These methods are based on additional CNNs acting as post-processing mechanism. The concepts of meta classification (false positive / FP detection) and meta regression (performance estimation) on segment-level were introduced in (Rottmann et al., 2020). This line of research has been extended by a temporal component (Maag et al., 2020) and transferred to object detection (Schubert et al., 2021; Riedlinger et al., 2023) as well as to instance segmentation (Maag et al., 2021; Maag, 2021).

While MC Dropout as a sampling approach is still computationally expensive to create pixel-wise uncertainties, our method computes only the gradients of the last layer during a single inference run and can be applied to a wide range of semantic segmentation networks without architectural changes. Compared with the work presented in (Rottmann et al., 2020), our gradient information can extend the features extracted from the segmentation network's softmax output to enhance the segment-wise quality estimation.

OoD Segmentation. Uncertainty quantification methods demonstrate high uncertainty for erroneous predictions, so they are often applied to OoD detection. For instance, this can be accomplished via maximum softmax (probability) (Hendrycks and Gimpel, 2016), MC Dropout (Mukhoti and Gal, 2018) or deep ensembles (Lakshminarayanan et al., 2017) the latter of which also capture model uncertainty by averaging predictions over multiple sets of parameters in a Bayesian manner. Another line of research is OoD detection training, relying on the exploitation of additional training data, not necessarily from the real world, but disjoint from the original training data (Blum et al., 2019b; Chan et al., 2021b; Grcic et al., 2022; Grcic et al., 2023; Liu et al., 2023; Nayal et al., 2023; Rai et al., 2023; Tian et al., 2022). In this regard, an external reconstruction model followed by a discrepancy network is

considered in (Biase et al., 2021; Lis et al., 2019; Lis et al., 2020; Vojir et al., 2021; Vojiř and Matas, 2023) and normalizing flows are leveraged in (Blum et al., 2019b; Grcic et al., 2021; Gudovskiy et al., 2023). In (Lee et al., 2018; Liang et al., 2018), adversarial perturbations are performed on the input images to improve the separation of in- and out-of-distribution samples.

Specialized training approaches for OoD detection are based on different kinds of re-training with additional data and often require generative models. Meanwhile, our method does not require OoD data, re-training or complex auxiliary models. Moreover, we do not run a full backward pass which is, however, required for the computation of adversarial samples. In fact we found that it is often sufficient to only compute the gradients of the last convolutional layer. Our method is more related to classical uncertainty quantification approaches like maximum softmax, MC Dropout and ensembles. Note that the latter two are based on sampling and thus, much more computationally expensive compared to single inference.

3 METHOD DESCRIPTION

In the following, we consider a neural network with parameters θ yielding classification probabilities $\hat{\pi}(x, \theta) = (\hat{\pi}_1, \ldots, \hat{\pi}_C)$ over C semantic categories when presented with an input x. During training on paired data (x, y), where $y \in \{1, \ldots, C\}$ is the semantic label given to x, such a model is commonly trained by minimizing some kind of loss function $\mathcal{L}$ between y and the predicted probability distribution $\hat{\pi}(x, \theta)$ using gradient descent on θ. The gradient step $\nabla_\theta \mathcal{L}(\hat{\pi}(x, \theta) \| y)$ then indicates the direction and strength of training feedback obtained by (x, y). Asymptotically (in the amount of data and training steps taken), the expectation $\mathbb{E}_{(X,Y) \sim P}[\nabla_\theta \mathcal{L}(\hat{\pi}(X, \theta) \| Y)]$ of the gradient w.r.t. the data generating distribution P will vanish since θ sits at a local minimum of $\mathcal{L}$. We assume, that strong models which achieve high test accuracy can be seen as an approximation of such a parameter configuration θ. Such a model has small gradients on in-distribution data which is close to samples (x, y) in the training data. Samples that differ from training data may receive larger feedback. Like-wise, it is plausible to obtain large training gradients from OoD samples that are far away from the effective support of P.

In order to quantify uncertainty about the prediction $\hat{\pi}(x, \theta)$, we replace the label y from above by some fixed auxiliary label for which we make two concrete choices in our method. On the one hand,

we replace y by the class prediction one-hot vector $y_k^{oh} = \delta_{k\hat{c}}$ with $\hat{c} = \arg\max_{k=1,\ldots,C} \hat{\pi}_k$ and the Kronecker symbol δ_{ij}. This quantity correlates strongly with training labels y on in-distribution data. On the other hand, we regard a constant uniform, all-warm label $y_k^{uni} = 1/C$ for $k = 1,\ldots,C$ as a replacement for y. To motivate the latter choice, we consider that classification models are oftentimes trained on the categorical cross entropy loss

$$\mathcal{L}(\hat{\pi}\|y) = -\sum_{k=1}^{C} y_k \log(\hat{\pi}_k). \qquad (1)$$

Since the gradient of this loss function is linear in the label y, a uniform choice y^{uni} will return the average gradient per class which is expected to be large on OoD data where all possible labels are similarly unlikely. The magnitude of $\nabla_\theta \mathcal{L}(\hat{\pi}\|y)$ serves as uncertainty / OoD score. In the following, we explain how to compute such scores for pixel-wise classification models.

3.1 Efficient Computation in Semantic Segmentation

We regard a generic segmentation architecture utilizing a final convolution as the pixel-wise classification mechanism. In the following, we also assume that the segmentation model is trained via the commonly-used pixel-wise cross entropy loss $\mathcal{L}_{ab}(\phi(\theta)\|y)$ (cf. eq. (1)) over each pixel (a,b) with feature map activations $\phi(\theta)$. However, similarly compact formulas hold for arbitrary loss functions. Pixel-wise probabilities are obtained by applying the softmax function Σ to each output pixel, i.e., $\hat{\pi}_{ab,k} = \Sigma^k(\phi_{ab}(\theta))$. With eq. (1), we find for the loss gradient

$$\nabla_\theta \mathcal{L}_{ab}(\Sigma(\phi(\theta))\|y) = \sum_{k=1}^{C} \Sigma^k(\phi_{ab})(1-y_k) \cdot \nabla_\theta \phi_{ab}^k(\theta). \qquad (2)$$

Here, θ is any set of parameters within the neural network. Here, ϕ is the convolution result of a pre-convolution feature map ψ (see Figure 2) against a filter bank K which assumes the role of θ. K has parameters $(K_e^h)_{fg}$ where e and h indicate in- and out-going channels respectively and f and g index the spatial filter position. The features ϕ are linear in both, K and ψ, and explicitly depend on bias parameters β in the form

$$\phi_{ab}^d = (K*\psi)_{ab}^d + \beta^d = \sum_{j=1}^{\kappa} \sum_{p,q=-s}^{s} (K_j^d)_{pq} \psi_{a+p,b+q}^j + \beta^d. \qquad (3)$$

We denote by κ the total number of in-going channels and s is the spatial extent to either side of the filter K_j^d which has total size $(2s+1) \times (2s+1)$. We make

different indices in the convolution explicit in order to compute the exact gradients w.r.t. the filter weights $(K_e^h)_{fg}$, where we find for the last layer gradients

$$\frac{\partial \phi_{ab}^d}{\partial (K_e^h)_{fg}} = \delta_{dh} \psi_{a+f,b+g}^e. \qquad (4)$$

Together with eq. (2), we obtain the closed form for computing the correct backpropagation gradients of the loss on pixel-level for our choices of auxiliary labels which we state in the following paragraph. The computation of the loss gradient can be traced in Figure 2.

If we take the predicted class $\hat{c}$ as a one-hot label per pixel as auxiliary labels, i.e., y^{oh}, we obtain for the last layer gradients

$$\frac{\partial \mathcal{L}_{ab}(\Sigma(\phi)\|y^{oh})}{\partial K_e^h} = \Sigma^h(\phi_{ab}) \cdot (1-\delta_{h\hat{c}}) \cdot \psi_{ab}^e \qquad (5)$$

which depends on quantities that are easily accessible during a forward pass through the network. Note, that the filter bank K for the special case of (1×1)-convolutions does not require spatial indices which is a common situation in segmentation networks, albeit, not necessary for our method to be applied. Similarly, we find for the uniform label $y_j^{uni} = 1/C$

$$\frac{\partial \mathcal{L}_{ab}(\Sigma(\phi)\|y^{uni})}{\partial K_e^h} = \frac{C-1}{C} \Sigma^h(\phi_{ab}) \psi_{ab}^e. \qquad (6)$$

These formulas reveal a practical factorization of the gradient which we will exploit computationally in the next section. Therefore, pixel-wise gradient norms are simple to implement and particularly efficient to compute for the last layer of the segmentation model.

3.2 Uncertainty Scores

We obtain pixel-wise scores, i.e., still depending on a and b, by computing the partial norm $\|\nabla_K \mathcal{L}_{ab}\|_p$ of this tensor over the indices e and h for some fixed value of p. This can again be done in a memory efficient way by the natural decomposition $\partial \mathcal{L}/\partial K_e^h = S^h \cdot \psi^e$. In addition to their use in error detection, these scores can be used in order to detect OoD objects in the input, i.e., instances of categories not present in the training distribution of the segmentation network. We identify OoD regions with pixels that have a gradient score higher than some threshold and find connected components like the one shown in Figure 1 (bottom). We also consider values $0 < p < 1$ in addition to positive integer values. Note, that such choices do not strictly define the geometry of a vector space norm, however, $\|\cdot\|_p$ may still serve as a notion of magnitude and generates a partial ordering. The tensorized multiplications required in eqs. (5) and (6)
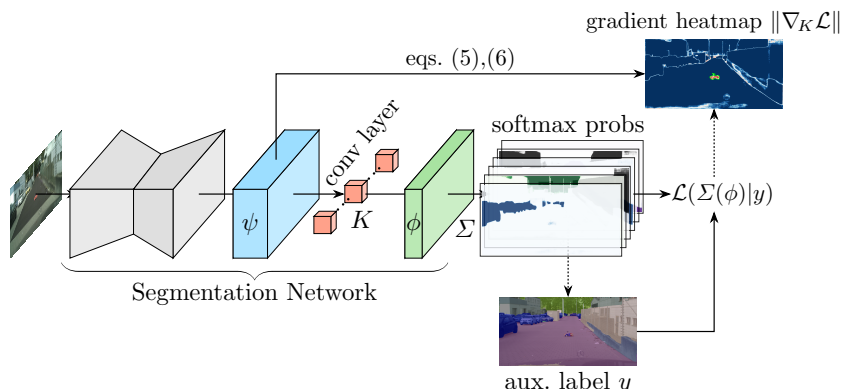
Figure 2: Schematic illustration of the computation of pixel-wise gradient norms for a semantic segmentation network with a final convolution layer. Auxiliary labels may be derived from the softmax prediction or supplied in any other way (e.g., as a uniform all-warm label). We circumvent direct back propagation per pixel by utilizing eqs. (5), (6).

are far less computationally expensive than a forward pass through the DNN. This means that computationally, this method is preferable over prediction sampling like MC Dropout or ensembles. We abbreviate our method using the **p**ixel-wise **g**radient **n**orms obtained from eqs. (5) and (6) by PGN_{oh} and PGN_{uni}, respectively. The particular value of p is denoted by superscript, e.g., $PGN_{oh}^{p=0.3}$ for the ($p = 0.3$)-seminorm of gradients obtained from y^{oh}.

4 EXPERIMENTS

In this section, we present the experimental setting first and then evaluate the uncertainty estimation quality of our method on pixel and segment level. We apply our gradient-based method to OoD segmentation, show some visual results and explicitly measure the runtime of our method.

4.1 Experimental Setting

Datasets. We perform our tests on the Cityscapes (Cordts et al., 2016) dataset for semantic segmentation in street scenes and on four OoD segmentation datasets[1]. The Cityscapes dataset consists of 2,975 training and 500 validation images of dense urban traffic in 18 and 3 different German towns, respectively. The LostAndFound (LAF) dataset (Pinggera et al., 2016) contains 1,203 validation images with annotations for the road surface and the OoD objects, i.e., small obstacles on German roads in front of the ego-car. A filtered version (LAF test-NoKnown) is provided in (Chan et al., 2021a). The Fishyscapes LAF dataset (Blum et al., 2019a)

includes 100 validation images (and 275 non-public test images) and refines the pixel-wise annotations of the LAF dataset distinguishing between OoD object, background (Cityscapes classes) and void (anything else). The RoadObstacle21 dataset (Chan et al., 2021a) (412 test images) is comparable to the LAF dataset as all obstacles appear on the road, but it contains more diversity in the OoD objects as well as in the situations. In the RoadAnomaly21 dataset (Chan et al., 2021a) (100 test images), a variety of unique objects (anomalies) appear anywhere on the image which makes it comparable to the Fishyscapes LAF dataset.

Segmentation Networks. We consider a state-of-the-art DeepLabv3+ network (Chen et al., 2018) with two different backbones, WideResNet38 (Wu et al., 2016) and SEResNeXt50 (Hu et al., 2018). The network with each respective backbone is trained on Cityscapes achieving a mean IoU value of 90.58% for the WideResNet38 backbone and 80.76% for the SEResNeXt50 on the Cityscapes validation set. We use one and the same model trained exclusively on the Cityscapes dataset for both tasks, the uncertainty estimation and the OoD segmentation, as our method does not require additional training. Therefore, our method leaves the entire segmentation performance of the base model completely intact.

4.2 Numerical Results

We provide results for both, error detection and OoD segmentation in the following.

Pixel-Wise Uncertainty Evaluation. In order to assess the correlation of uncertainty and prediction errors on the pixel level, we resort to the commonly

---

[1]Benchmark: http://segmentmeifyoucan.com/

Table 1: Pixel-wise uncertainty evaluation results for both backbone architectures and the Cityscapes dataset in terms of ECE and AuSE.

	WideResNet		SEResNeXt	
	ECE $\downarrow$	AuSE $\downarrow$	ECE $\downarrow$	AuSE $\downarrow$
Ensemble	0.0173	0.4543	0.0279	0.0482
MC Dropout	0.0444	0.7056	0.0091	0.5867
Maximum Softmax	**0.0017**	<u>0.0277</u>	**0.0032**	**0.0327**
Entropy	0.0063	0.0642	0.0066	0.0623
$PGN_{oh}^{p=2}$ (ours)	<u>0.0019</u>	**0.0268**	<u>0.0039</u>	<u>0.0365</u>

used sparsification graphs (Ilg et al., 2018). Sparsification graphs normalized w.r.t. the optimal oracle (sparsification error) can be compared in terms of the so-called area under the sparsification error curve (AuSE). The closer the uncertainty estimation is to the oracle in terms of Brier score evaluation, i.e., the smaller the AuSE, the better the uncertainty eliminates false predictions by the model. The AuSE metric is capable of grading the uncertainty ranking, however, does not address the statistics in terms of given confidence. Therefore, we resort to an evaluation of calibration in terms of expected calibration error (ECE, (Guo et al., 2017)) to assess the statistical reliability of the uncertainty measure.

As baseline methods we consider the typically used uncertainty estimation measures, i.e., mutual information computed via samples from deep ensembles and MC dropout as well as the uncertainty ranking provided by the maximum softmax probabilities native to the segmentation model and the softmax entropy. An evaluation of calibration errors requires normalized scores, so we normalize our gradient scores according to the highest value obtained on the test data for the computation of ECE.

The resulting metrics are compiled in Table 1 for both architectures evaluated on the Cityscapes val split. We see that the calibration of our method is roughly on par with the stronger maximum softmax baseline (which was trained to exactly that aim via the negative log-likelihood loss) and outperforms the other three baselines. In particular, we achieve superior values to MC Dropout which represents the typical uncertainty measure for semantic segmentation.

Segment-wise Prediction Quality Estimation. To reduce the number of FP predictions and to estimate the prediction quality, we use meta classification and meta regression introduced by (Rottmann et al., 2020). As input for the post-processing models, the authors use information from the network's softmax output which characterize uncertainty and geometry of a given segment like the segment size. The degree of randomness in semantic segmentation predic-
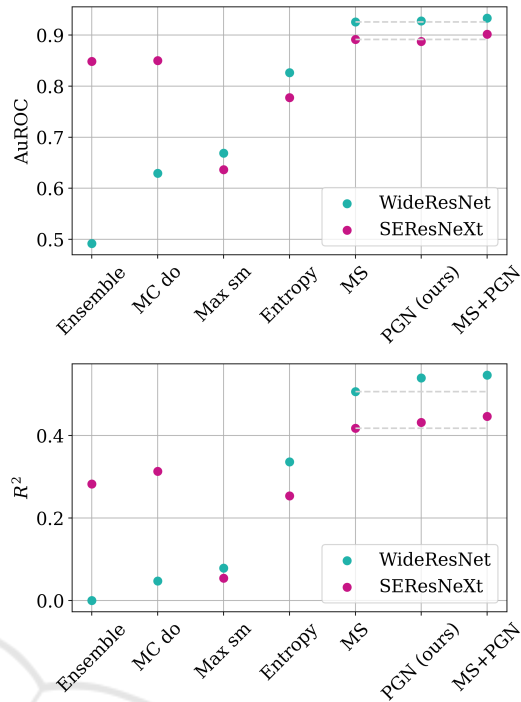

Figure 3: Segment-wise uncertainty evaluation results for both backbone architectures and the Cityscapes dataset in terms of classification AuROC and regression R^2 values. From left to right: ensemble, MC Dropout, maximum softmax, mean entropy, MetaSeg (MS) approach, gradient features obtained by predictive one-hot and uniform labels (PGN), MetaSeg in combination with PGN.

tion is quantified by pixel-wise quantities, like entropy and probability margin. Segment-wise features are generated from these quantities via average pooling. These MetaSeg features used in (Rottmann et al., 2020) serve as a baseline in our tests.

Similarly, we construct segment-wise features from our pixel-wise gradient p-norms for $p \in \{0.1, 0.3, 0.5, 1, 2\}$. We compute the mean and the variance of these pixel-wise values over a given segment. These features per segment are considered also for the inner and the boundary since the gradient scores may be higher on the boundary of a segment, see Figure 1 (bottom). The inner of the segment consists of all pixels whose eight neighboring pixels are also elements of the same segment. Furthermore, we define some relative mean and variance features over the inner and boundary which characterizes the degree of fractality. These hand-crafted quantities form a structured dataset where the columns correspond to features and the rows to predicted segments which serve as input to the post-processing models.

We determine the prediction accuracy of semantic segmentation networks with respect to the ground
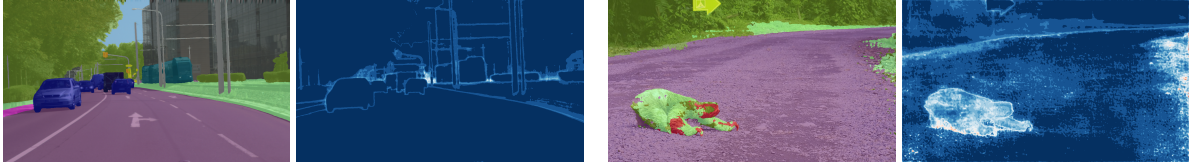
Figure 4: Semantic segmentation prediction and $\text{PGN}_{uni}^{p=0.5}$ heatmap for the Cityscapes dataset (left) and the RoadAnomaly21 dataset (right) for the WideResNet backbone.

Table 2: OoD segmentation benchmark results for the LostAndFound and the RoadObstacle21 dataset.

	OoD data	OoD arch	adversarial	LostAndFound test-NoKnown					RoadObstacle21				
				AuPRC ↑	FPR$_{95}$ ↓	$\overline{\text{sIoU}}$ ↑	$\overline{\text{PPV}}$ ↑	$\overline{F_1}$ ↑	AuPRC ↑	FPR$_{95}$ ↓	$\overline{\text{sIoU}}$ ↑	$\overline{\text{PPV}}$ ↑	$\overline{F_1}$ ↑
Void Classifier	✓			4.8	47.0	1.8	35.1	1.9	10.4	41.5	6.3	20.3	5.4
Maximized Entropy	✓	(✓)		77.9	9.7	**45.9**	63.1	**49.9**	**85.1**	**0.8**	**47.9**	**62.6**	48.5
SynBoost	✓	✓		**81.7**	**4.6**	36.8	**72.3**	48.7	71.3	3.2	44.3	41.8	37.6
Image Resynthesis		✓		57.1	8.8	27.2	30.7	19.2	37.7	4.7	16.6	20.5	8.4
Embedding Density		✓		61.7	10.4	37.8	35.2	27.6	0.8	46.4	35.6	2.9	2.3
ODIN			✓	52.9	30.0	39.8	49.3	34.5	22.1	15.3	21.6	18.5	9.4
Mahalanobis			✓	55.0	12.9	33.8	31.7	22.1	20.9	13.1	13.5	21.8	4.7
Ensemble		(✓)		2.9	82.0	6.7	7.6	2.7	1.1	77.2	8.6	4.7	1.3
MC Dropout		(✓)		36.8	35.6	17.4	34.7	13.0	4.9	50.3	5.5	5.8	1.1
Maximum Softmax				30.1	33.2	14.2	**62.2**	10.3	15.7	16.6	19.7	15.9	6.3
Entropy				52.0	30.0	40.4	53.8	42.4	**20.6**	**16.3**	**21.4**	**19.5**	**10.4**
$\text{PGN}_{uni}^{p=0.5}$ (ours)				**69.3**	**9.8**	**50.0**	44.8	**45.4**	16.5	19.7	19.5	14.8	7.4

truth via the segment-wise intersection over union (IoU, (Jaccard, 1912)). On the one hand, we perform the direct prediction of the IoU (meta regression) which serves as prediction quality estimate. On the other hand, we discriminate between IoU = 0 (FP) and IoU > 0 (true positive) (meta classification). We use linear classification and regression models.

For the evaluation, we use AuROC (area under the receiver operating characteristic) for meta classification and determination coefficient R^2 for meta regression. As baselines, we employ an ensemble, MC Dropout, maximum softmax probability, entropy and the MetaSeg framework. A comparison of these methods and our approach for the Cityscapes dataset is given in Figure 3. We outperform all baselines with the only exception of meta classification for the SEResNeXt backbone where MetaSeg achieves a marginal 0.41 pp higher AuROC value than ours. Such a post-processing model captures all the information which is contained in the network's output. Therefore, matching the MetaSeg performance by a gradient heatmap is a substantial achievement for a predictive uncertainty method. Moreover, we enhance the MetaSeg performance for both networks and both tasks combining the MetaSeg features with PGN by up to 1.02 pp for meta classification and up to 3.98 pp for meta regression. This indicates that gradient features contain information which is partially orthogonal to the information contained in the softmax output. Especially, the highest AuROC value of 93.31%

achieved for the WideResNet backbone, shows the capability of our approach to estimate the prediction quality and filter out FP predictions on the segment-level for in-distribution data.

OoD Segmentation. Figure 4 shows segmentation predictions of the pre-trained DeepLabv3+ network with the WideResNet38 backbone together with its corresponding $\text{PGN}_{uni}^{p=0.5}$-heatmap. The two panels on the left show an in-distribution prediction on Cityscapes where uncertainty is mainly concentrated around segmentation boundaries which are always subject to high prediction uncertainty. Moreover, we see some false predictions in the far distance around the street crossing which can be found as a region of high gradient norm in the heatmap. In the two panels to the right, we see an OoD prediction from the RoadAnomaly21 dataset of a sloth crossing the street which is classified as part vegetation, terrain and person. The outline of the segmented sloth can be seen brightly in the gradient norm heatmap to the right indicating clear separation.

Our results in OoD segmentation are based on the evaluation protocol of the official SegmentMeIfYouCan benchmark (Chan et al., 2021a). Evaluation on the pixel-level involves the threshold-independent area under precision-recall curve (AuPRC) and the false positive rate at the point of 0.95 true positive rate (FPR$_{95}$). The latter constitutes an interpretable choice of operating point for each method where a minimum

Table 3: OoD segmentation benchmark results for the Fishyscapes LostAndFound and the RoadAnomaly21 dataset.

	OoD data	OoD arch	adversarial	Fishyscapes LostAndFound					RoadAnomaly21				
				AuPRC $\uparrow$	FPR$_{95}$ $\downarrow$	$\overline{\text{sIoU}}$ $\uparrow$	$\overline{\text{PPV}}$ $\uparrow$	$\overline{F_1}$ $\uparrow$	AuPRC $\uparrow$	FPR$_{95}$ $\downarrow$	$\overline{\text{sIoU}}$ $\uparrow$	$\overline{\text{PPV}}$ $\uparrow$	$\overline{F_1}$ $\uparrow$
Void Classifier	✓			11.7	15.3	9.2	39.1	14.9	36.6	63.5	21.1	22.1	6.5
Maximized Entropy	✓	(✓)		44.3	37.7	21.1	**48.6**	30.0	**85.5**	**15.0**	**49.2**	**39.5**	**28.7**
SynBoost	✓	✓		**64.9**	30.9	**27.9**	**48.6**	**38.0**	56.4	61.9	34.7	17.8	10.0
Image Resynthesis		✓		5.1	29.8	5.1	12.6	4.1	52.3	25.9	39.7	11.0	12.5
Embedding Density		✓		8.9	42.2	5.9	10.8	4.9	37.5	70.8	33.9	20.5	7.9
ODIN			✓	15.5	38.4	9.9	21.9	9.7	33.1	71.7	19.5	17.9	5.2
Mahalanobis			✓	32.9	**8.7**	19.6	29.4	19.2	20.0	87.0	14.8	10.2	2.7
Ensemble		(✓)		0.3	90.4	3.1	1.1	0.4	17.7	91.1	16.4	20.8	3.4
MC Dropout		(✓)		14.4	47.8	4.8	18.1	4.3	28.9	69.5	20.5	17.3	4.3
Maximum Softmax				5.6	40.5	3.5	9.5	1.8	28.0	72.1	15.5	15.3	5.4
Entropy				14.0	39.3	8.0	17.5	7.7	31.6	71.9	15.7	18.4	4.2
PGN$_{uni}^{p=0.5}$ (ours)				**26.9**	**36.6**	**14.8**	**29.6**	**16.5**	**42.8**	**56.4**	**25.8**	**21.8**	**9.7**

Table 4: Runtime measurements in seconds per frame for each method; standard deviations taken over samples in the Cityscapes validation dataset.

sec. per frame $\downarrow$	WideResNet	SEResNeXt
Softmax	3.02 ± 0.18	1.08 ± 0.04
MC Dropout	7.52 ± 0.35	2.31 ± 0.10
PGN$_{oh}$ + PGN$_{uni}$ (ours)	3.05 ± 0.16	1.09 ± 0.05

true positive fraction is dictated. On segment-level, an adjusted version of the mean intersection over union (sIoU) representing the accuracy of the segmentation obtained by thresholding at a particular point, positive predictive value (PPV or precision) playing the role of binary instance-wise accuracy and the F_1-score. The latter segment-wise metrics are averaged over thresholds between 0.25 and 0.75 with a step size of 0.05 leading to the quantities $\overline{\text{sIoU}}$, $\overline{\text{PPV}}$ and $\overline{F_1}$.

Here, we compare the gradient scores of PGN$_{uni}$ as an OoD score against all other methods from the benchmark which provide evaluation results on all given datasets. We restrict ourselves to PGN$_{uni}$ since we suspect it performs especially well in OoD segmentation as explained in Section 3. The results are based on evaluation files submitted to the public benchmark and are, therefore, deterministic. As baselines, we also include the same methods as for error detection before since these constitute a fitting comparison. Note, that the standard entropy baseline is not featured in the official leaderboard, so we report our own results obtained by the softmax entropy with the WideResNet backbone which performed better.

The results verified by the official benchmark are compiled in Table 2 and Table 3. We separate both tables into two halves. The bottom half contains methods with similar requirements as our method while the top half contains the stronger OoD segmentation methods which may have heavy additional requirements such as auxiliary data, architectural changes, retraining or the computation of adversarial examples.

Note, the best performance value of both halves of the table is marked. In the lower part, a comparison with methods which have similarly low additional requirements is shown. We mark deep ensembles and MC dropout as requiring architectural changes since they technically require re-training in order to serve as numerical approximations of a Bayesian neural network. PGN performs among the strongest methods, showing superior performance on LostAndFound test-NoKnown, Fishyscapes and the RoadAnomaly dataset. While not clearly superior to the Entropy baseline on RoadObstacles, PGN still yields stronger performance than the deep ensemble and MC dropout baselines. Especially in Table 3, the methods in the lower part are significantly weaker w.r.t. most of the computed metrics.

The upper part of the table provides results for other OoD segmentation methods utilizing adversarial samples (ODIN (Liang et al., 2018) and Mahalanobis (Lee et al., 2018)), OoD data (Void Classifier (Blum et al., 2019b), Maximized Entropy (Chan et al., 2021b) and SynBoost (Biase et al., 2021)) or auxiliary models (SynBoost, Image Resynthesis (Lis et al., 2019) and Embedding Density (Blum et al., 2019b)). Note, that the Maximized Entropy method does not modify the network architecture, but uses another loss function requiring retraining. In several cases, we find that our method is competitive with some of the stronger methods. We outperform the two adversarial-based methods for the LostAndFound as well as the RoadAnomaly21 dataset. In detail, we obtain AuPRC values up to 22.8 pp higher on segment-level and $\overline{F_1}$ values values up to 24.8 pp higher on pixel-level. For the other two datasets we achieve similar results. Furthermore, we beat the Void Classifier method, that uses OoD data during training, in most cases. We improve the AuPRC metric by up to 64.5 pp and the $\overline{\text{sIoU}}$ metric by up to 48.2 pp, both for the LostAndFound dataset. In addition, our gradient

norm outperforms in many cases the Image Resynthesis as well as the Embedding Density approach which are based on an external reconstruction model followed by a discrepancy network and normalizing flows, respectively. Summing up, we have shown superior OoD segmentation performance in comparison to the other uncertainty based methods and outperform some of the more complex approaches (using OoD data, adversarial samples or generative models).

Computational Runtime. Lastly, we demonstrate the computational efficiency of our method and show runtime measurements for the network forward pass ("Softmax"), MC dropout (25 samples) and computing both, PGN_{oh} and PGN_{uni}, in Table 4. While sampling MC dropout requires over twice the time per frame for both backbone networks, the computation of $PGN_{oh} + PGN_{uni}$ only leads to a marginal computational overhead of around 1% due to consisting only of tensor multiplications. Note, that the entropy baseline requires roughly the same compute as Softmax and deep ensembles tend to be slower than MC dropout due to partial parallel computation. The measurements were each made on a single Nvidia Quadro P6000 GPU.

Limitations. We declare that our method is merely on-par with other uncertainty quantification methods for pixel-level considerations but outperforming well-known methods such as MC Dropout. For segment-wise error detection, our approach showed improved results. Moreover, there have been some submissions to the SegmentMeIfYouCan benchmark (requiring OoD data, re-training or complex auxiliary models) which outperform our method. However, a direct application of the methods is barely possible if suitable OoD data have to pass into the training or auxiliary models are used in combination with the segmentation network, which also increases the runtime during application, in comparison to our simple and flexible method. Note, our approach has one light architectural restrictions, i.e., the final layer has to be a convolution, but this is in general common for segmentation models.

5 CONCLUSION

In this work, we presented an efficient method of computing gradient uncertainty scores for a wide class of deep semantic segmentation models. Moreover, we appreciate a low computational cost associated with them. Our experiments show that large gradient norms obtained by our method statistically correspond to erroneous predictions already on the pixel-level and can be normalized such that they yield similarly calibrated confidence measures as the maximum softmax score of the model. On a segment-level our method shows considerable improvement in terms of error detection. Gradient scores can be utilized to segment out-of-distribution objects significantly better than sampling- or any other output-based method on the SegmentMeIfYouCan benchmark and has competitive results with a variety of methods, in several cases clearly outperforming all of them while coming at negligible computational overhead. We hope that our contributions in this work and the insights into the computation of pixel-wise gradient feedback for segmentation models will inspire future work in uncertainty quantification and pixel-wise loss analysis.

REFERENCES

Biase, G. D., Blum, H., Siegwart, R. Y., and Cadena, C. (2021). Pixel-wise anomaly detection in complex driving scenes. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16913–16922.

Blum, H., Sarlin, P.-E., Nieto, J., Siegwart, R., and Cadena, C. (2019a). Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Blum, H., Sarlin, P.-E., Nieto, J. I., Siegwart, R. Y., and Cadena, C. (2019b). The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, pages 3119 – 3135.

Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Fua, P., Salzmann, M., and Rottmann, M. (2021a). Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chan, R., Rottmann, M., and Gottschalk, H. (2021b). Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

DeVries, T. and Taylor, G. W. (2018). Leveraging uncertainty estimates for predicting segmentation quality.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48, pages 1050–1059.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. (2020). Your classifier is secretly an energy based model and you should treat it like one. *International Conference on Learning (ICLR)*.

Grcic, M., Bevandić, P., and Segvic, S. (2021). Dense anomaly detection by robust learning on synthetic negative data.

Grcic, M., Bevandić, P., and Segvic, S. (2022). Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *European Conference on Computer Vision (ECCV)*, pages 500–517.

Grcic, M., Šarić, J., and Šegvić, S. (2023). On advantages of mask-level recognition for outlier-aware segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2937–2947.

Gudovskiy, D., Okuno, T., and Nakata, Y. (2023). Concurrent misclassification and out-of-distribution detection for semantic segmentation via energy-based normalizing flow.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks.

Hoebel, K., Andrearczyk, V., Beers, A., Patel, J., Chang, K., et al. (2020). An exploration of uncertainty information for segmentation quality assessment.

Hornauer, J. and Belagiannis, V. (2022). Gradient-based uncertainty for monocular depth estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 613–630. Springer.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks.

Huang, C., Wu, Q., and Meng, F. (2016). Qualitynet: Segmentation quality evaluation with deep convolutional networks. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4.

Huang, R., Geng, A., and Li, Y. (2021). On the importance of gradients for detecting distributional shifts in the wild. In *Neural Information Processing Systems (NeurIPS)*.

Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., and Brox, T. (2018). Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 652–667.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Neural Information Processing Systems (NeurIPS)*, page 6405–6416.

Lee, H., Kim, S. T., Navab, N., and Ro, Y. (2020). Efficient ensemble model generation for uncertainty estimation with bayesian approximation in segmentation.

Lee, J., Prabhushankar, M., and Alregib, G. (2022). Gradient-based adversarial and out-of-distribution detection. In *International Conference on Machine Learning*.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Neural Information Processing Systems (NeurIPS)*.

Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning (ICLR)*.

Lis, K., Honari, S., Fua, P., and Salzmann, M. (2020). Detecting road obstacles by erasing them.

Lis, K., Nakka, K., Fua, P., and Salzmann, M. (2019). Detecting the unexpected via image resynthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liu, Y., Ding, C., Tian, Y., Pang, G., Belagiannis, V., Reid, I., and Carneiro, G. (2023). Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation.

Maag, K. (2021). False negative reduction in video instance segmentation using uncertainty estimates. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*.

Maag, K., Chan, R., Uhlemeyer, S., Kowol, K., and Gottschalk, H. (2022). Two video data sets for tracking and retrieval of out of distribution objects. In *Asian Conference on Computer Vision (ACCV)*, pages 3776–3794.

Maag, K., Rottmann, M., and Gottschalk, H. (2020). Time-dynamic estimates of the reliability of deep semantic

segmentation networks. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI).*

Maag, K., Rottmann, M., Varghese, S., Hüger, F., Schlicht, P., and Gottschalk, H. (2021). Improving video instance segmentation by light-weight temporal uncertainty estimates. In *International Joint Conference on Neural Network (IJCNN).*

MacKay, D. J. C. (1992). A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.

Mukhoti, J. and Gal, Y. (2018). Evaluating bayesian deep learning methods for semantic segmentation.

Nayal, N., Yavuz, M., Henriques, J. F., and Güney, F. (2023). Rba: Segmenting unknown regions rejected by all.

Oberdiek, P., Rottmann, M., and Gottschalk, H. (2018). Classification uncertainty of deep neural networks based on gradient information. *Artificial Neural Networks and Pattern Recognition (ANNPR).*

Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., and Mester, R. (2016). Lost and found: detecting small road hazards for self-driving vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*

Rai, S., Cermelli, F., Fontanel, D., Masone, C., and Caputo, B. (2023). Unmasking anomalies in road-scene segmentation.

Riedlinger, T., Rottmann, M., Schubert, M., and Gottschalk, H. (2023). Gradient-based quantification of epistemic uncertainty for deep object detectors. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3921–3931.

Rottmann, M., Colling, P., Hack, T., Hüger, F., Schlicht, P., et al. (2020). Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *IEEE International Joint Conference on Neural Networks (IJCNN) 2020.*

Schubert, M., Kahl, K., and Rottmann, M. (2021). Metadetect: Uncertainty quantification and prediction quality estimates for object detection. In *International Joint Conference on Neural Networks (IJCNN).*

Tian, Y., Liu, Y., Pang, G., Liu, F., Chen, Y., and Carneiro, G. (2022). Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision (ECCV).*

Vojir, T., Šipka, T., Aljundi, R., Chumerin, N., Reino, D. O., and Matas, J. (2021). Road anomaly detection by partial image reconstruction with segmentation coupling. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15651–15660.

Vojíř, T. and Matas, J. (2023). Image-consistent detection of road anomalies as unpredictable patches. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5480–5489.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., and Xiao, B. (2021). Deep high-resolution representation

learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Wickstrøm, K., Kampffmeyer, M., and Jenssen, R. (2019). Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis*, 60:101619.

Wu, Z., Shen, C., and Hengel, A. (2016). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition.*