Identifying Out-of-Domain Objects with Dirichlet Deep Neural Networks

Ahmed Hammam
Stellantis, Opel Automobile GmbH
Institute of Measurement and Control Systems,
Karlsruhe Institute of Technology
Germany
ahmedmostafa.hammam@external.stellantis.com

Frank Bonarens
Stellantis, Opel Automobile GmbH
Germany
frank.bonarens@stellantis.com

Seyed Eghabl Ghobadi
Technische Hochschule Mittelhessen
Germany
seyed.eghbal.ghobadi@stellantis.com

Christoph Stiller
Institute of Measurement and Control Systems,
Karlsruhe Institute of Technology
Germany
stiller@kit.edu

Abstract

Deep neural networks are usually trained on a closed set of classes, which makes them distrustful when handling previously-unseen out-of-domain (OOD) objects. In safety-critical applications such as perception for automated driving, detecting and localizing OOD objects is crucial, especially if they are positioned in the driving path. In the context of this contribution, OOD objects refer to objects that were not represented in the training dataset. We propose a Dirichlet deep neural network for instance segmentation with inherent uncertainty modeling based on Dirichlet distributions and the Intermediate Layer Variational Inference (ILVI). A thorough analysis shows that our method delivers reliable uncertainty estimates to its predictions whilst identifying OOD instances. The model-agnostic approach can be applied to different instance segmentation models as demonstrated for two different state-of-the-art deep neural networks. Superior results can be shown on the out-of-domain Lost and Found dataset compared to state-of-the-art approaches, whilst also achieving improvements on the in-domain Cityscapes dataset.

1. Introduction

Deep learning has revolutionized computer vision, offering groundbreaking advancements in various domains including medical imaging [1] and automated driving (AD) [2]. In the field of AD systems, deep neural networks (DNNs) have emerged as the predominant method, finding widespread applications in sensor fusion [3], path planning [4] and image semantic segmentation [5].

Despite the remarkable achievements of deep learning techniques in various tasks, a key insufficiency of DNNs is the lack of out-of-domain (OOD) detection capability which is essential to ensure the reliability and safety of machine learning systems [6]. This is particularly important for an AD system, where the ability to identify OOD objects in the driving path is safety critical.

In recent years, researchers have addressed the aforementioned limitations by incorporating the capability of expressing uncertainty in the predictions made by DNNs [7]. Uncertainty modeling has paved the way for the exploration and implementation of various techniques to quantify uncertainty estimations. These estimations not only inherently deliver evidence for the reliability of predictions but also serve as a means to identify anomalies in the network's output and identify OOD objects present in the input data [8].

Recent studies showed several drawbacks of depending on uncertainty estimation as a method to detect OOD objects, as the overconfidence in the DNN predictions could highly mislead the uncertainty estimation, hindering the DNN from being able to have high uncertainty estimation on OOD objects [9]. In recent studies [10, 11], it has been shown that the usage of Dirichlet DNNs and Intermediate layer variational inference (ILVI) [12] contributes to the improvement of uncertainty estimation and overcomes the overconfidence of the DNNs for semantic segmentation tasks. Figure 1 showcases the capabilities of our proposed method, highlighting its superior performance compared to the baseline Cross Entropy (CE) DNN and the current state-of-the-art approaches. Our approach excels in accurately

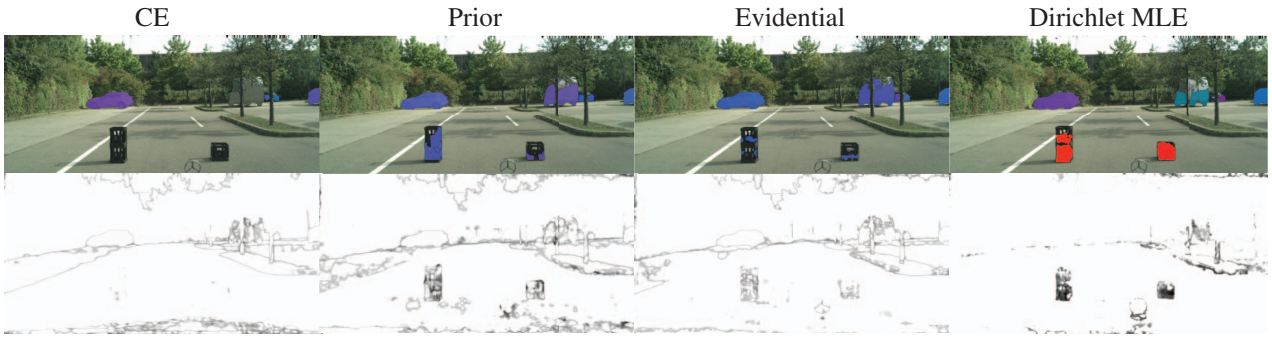


Figure 1: Sample result representing the instance segmentation output for all approaches (top row). Uncertainty estimates are shown for the baseline and the two state-of-the-art approaches whereas Dirichlet strength for our Dirichlet MLE approach is shown (bottom row). The baseline fails to detect the OOD objects and lacks high uncertainty estimates for them. State-of-the-art approaches partially detect OOD objects but with only a few pixels exhibiting high uncertainty. In contrast, our proposed method effectively detects the OOD objects with low Dirichlet strength.

identifying and segmenting out-of-domain (OOD) objects, surpassing the limitations faced by the other methods.

In this work, we build on these studies by evolving this approach into identifying OOD objects in the driving scene, without using OOD object annotations for training. We develop a DNN architecture that is able to segment instances whilst being able to distinguish them as in-domain (ID) and OOD objects. This is achieved by using two properties from the Dirichlet distribution: predictive entropy and Dirichlet strength. Following [10, 11], we train the DNN using Maximum Likelihood Estimation (MLE) [13] and incorporate the ILVI approach for the identification of OOD objects. In the context of this work, an ID dataset refers to datasets having only ID objects, whereas an OOD dataset refers to a dataset having both ID and OOD objects. Furthermore, ID objects with different poses and orientations are also counted as OOD.

The remainder of the paper is organized as follows: Section 2 presents related work for uncertainty modeling approaches for identifying OOD objects, whilst the proposed architecture and the approach used is discussed and explained in Section 3. The experiments conducted to test our approach are presented in Section 4 and a conclusion of the work is discussed in Section 5.

2. Related Work

Anomaly detection was first investigated for the task of image classification. This involved postprocessing techniques that aimed to modify the confidence scores generated by a classification DNN [14]. While these methods were initially developed for identifying anomalies at the image level, they can be conveniently adjusted for the task of semantic segmentation. This adaptation involves treating each pixel in an image as a potential anomaly.

A recent approach in anomaly segmentation involves the use of generative models to reconstruct or resynthesize the original input image. The idea behind this approach is that the reconstructed images will better retain the visual characteristics of regions that contain familiar objects compared to those with unfamiliar objects. By identifying discrepancies between the original image and its reconstructed version at the pixel level, anomaly detection can be performed [15].

Recent work has been focusing on modeling the DNN outputs as Dirichlet distributions which would in return improve its uncertainty and OOD detection. Dirichlet Prior Networks build upon the framework introduced in [16] by modeling the predicted logits as the concentration parameters of a Dirichlet distribution. This distribution serves as a prior for the categorical distribution. The intention behind this extension is to capture the distributional uncertainty through the spread of the Dirichlet prior. In order to achieve this, authors of [16] train the DNN to mimic a Dirac distribution for correct predictions and flat distribution for incorrect predictions. To do so they propose to learn the parameters of a Dirichlet distribution by training the DNN using Kullback-Leibler (KL) divergence.

Similar to Dirichlet Prior networks, the authors in [17] propose the Evidential networks, which is a method that combines Dirichlet Prior networks with the likelihood to maximize the whole posterior. Inspired by Dempster–Shafer Theory of Evidence (DST), they treat the predictions of the DNN as subjective opinions and train the DNN to gather evidence supporting these opinions. Additionally, a penalty term is introduced to penalize the DNN for incorrect detections and encourage it to exhibit high uncertainty in such cases. This ensures that the network learns both accurate classifications and appropriate representations of uncertainty for incorrect predictions.

Inspired by previous studies of [16,17], our goal is to utilize Dirichlet models to enhance the identification of OOD objects using instance segmentation whilst also improving reliable uncertainty estimation and maintaining segmentation performance for ID classes. Optimizing the reliability of uncertainty estimation in the Dirichlet DNN by formulating its loss function using KL divergence is often considered challenging [16]. In this work, we adopt an alternative approach by directly formulating our loss function to maximize the likelihood of the Dirichlet concentration parameters, as suggested in [10, 11].

3. Methodology

3.1. Dirichlet DNN Architecture

The architecture, presented in Figure 2a, comprises a shared backbone that takes the input image and passes the extracted features to the ILVI module, as shown in Figure 2b. The ILVI module acts as a regularizer by adding stochasticity in the DNN avoiding overfitting and overconfidence. The output of the ILVI is passed on to the three following decoders in parallel: the semantic segmentation decoder, the instance segmentation decoder, and the depth decoder.

The semantic segmentation decoder and the Dirichlet layer, Figure 2c, are trained together to model the semantic segmentation output as a Dirichlet distribution, which enables the identification of OOD instances, as shown in Figure 2. This decoder generates three results: semantic segmentation, uncertainty estimation and Dirichlet strength based on the per-pixel Dirichlet distribution. The instance segmentation decoder generates the center points and the masks of the instances. Lastly, the depth decoder outputs per-pixel depth estimates only for the objects in the scene.

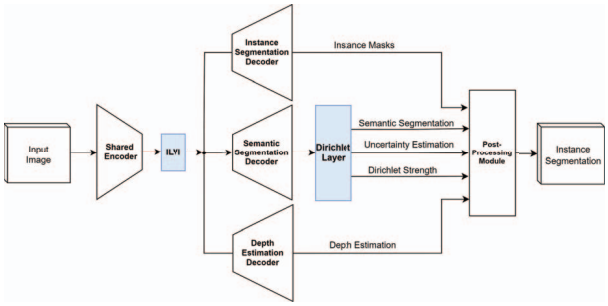
All generated outputs from the three decoders are passed on to the last post-processing module. The final result is an instance segmentation output where all objects in the scene are classified, as ID or OOD objects with additional attributes.

3.2. Semantic Segmentation Decoder

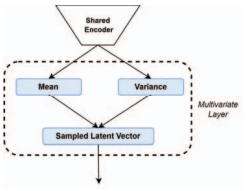
The semantic segmentation component of the DNN relies on the modeling of semantic segmentation using Dirichlet distributions. In this section, we elaborate on the Dirichlet modeling approach and subsequently provide an explanation of our method for identifying (OOD) objects using Dirichlet distributions.

Dirichlet Modeling: Given the probability simplex as $\mathcal{S} = \{(\theta_1, \ldots, \theta_k) : \theta_i \geq 0, \sum_i \theta_i = 1\}$, the Dirichlet distribution is a probability density function on vectors $\theta \in \mathcal{S}$ and categorized by concentration parameters $\alpha = \{\alpha_1, \ldots, \alpha_K\}$ as:

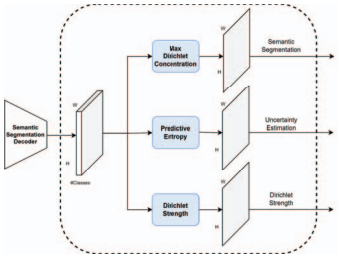
$$\text{Dir}(\theta; \alpha) = \log \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \quad (1)$$



(a) Dirichlet MLE DNN architecture



(b) ILVI layer



(c) Dirichlet layer

Figure 2: The Dirichlet MLE DNN is illustrated, emphasizing its key components for enhancing OOD identification. The ILVI layer introduces a multivariate layer structure, while the Dirichlet layer handles semantic segmentation, uncertainty estimation, and Dirichlet strength calculation.

where the normalizing constant $\frac{1}{B(\alpha)}$ denotes the multivariate Beta function $B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$, $\alpha_0 = \sum_{i=1}^{K} \alpha_i$ and Gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$, and θ denotes the ground truth probability distribution [13]. To model the Dirichlet distribution, the concentration parameters α correspond to each class output from the semantic segmentation decoder as follows: $\alpha = f_w(x)$, where α changes with each input x.

To train the Dirichlet distributions, we propose a direct maximization of the likelihood, inspired by the works of [10, 11]. Unlike the Dirichlet Prior and Evidential approaches, our method eliminates the constraints of the KL-divergence term in the loss function, allowing the DNN to explore the weight space more freely. This leads to improved segmentation performance and enhanced reliability in uncertainty estimation by encouraging a sharper concentration of Dirichlet parameters for correct predictions and flatter distributions for incorrect predictions.

Training a Dirichlet DNN with maximum likelihood estimation (MLE) can be done by minimizing the negative log-likelihood [13] as follows:

$$F(\alpha;\theta) = \log \prod \text{Dir}(\theta;\alpha) = \log \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$$
$$= \log\Gamma\left(\sum_{i=1}^{K} \alpha_j\right) - \sum_{i=1}^{K} \log\Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1)\log\theta_i \tag{2}$$

We aim to train the DNN to produce reliable uncertainty estimations by treating the DNN's correct and incorrect predictions separately. Our primary objective is to obtain accurate predictions with low uncertainties, while assigning high uncertainty to incorrect predictions.

To ensure high certainty for correct predictions, the DNN should exhibit a strong concentration towards the correct class. This can be achieved by maximizing the likelihood using the ground truth label probability and employing a *one-hot vector*. Conversely, for incorrect predictions, high uncertainty is achieved by maximizing the likelihood using an equal probability vector with equal probabilities assigned to all classes.

To address these cases, we extend the formulation presented in Equation 2 for the semantic segmentation as follows:

$$\mathcal{L}_{sem} = F(\alpha_{correct};\theta_{correct}) + F(\alpha_{incorrect};\theta_{incorrect}), \tag{3}$$

where $\alpha_{correct}$ and $\alpha_{incorrect}$ are the network's concentration parameters representing the correct and incorrect DNN predictions respectively, and $\theta_{correct}$ and $\theta_{incorrect}$ represent the ground truth probability distribution for the correct classes and the equal probability vector to yield high uncertainty respectively. As stated previously, it is worth mentioning that the DNN is not trained on OOD objects.

In this work, we model the uncertainty estimation of the Dirichlet distribution using the predictive entropy:

$$\hat{\mathbb{H}}[y|x] = -\sum_{c} (p(y=c|x,w))\log(p(y=c|x,w)) \tag{4}$$

where y is the output variable, c ranges over all the classes, $p(y=c|x,w) = \frac{\alpha_c}{\sum \alpha}$ is the probability of the input x being class c, and w are the model parameters. The class of each pixel for the semantic segmentation output is determined according to the highest concentration value of the Dirichlet distribution.

OOD Identification Using Dirichlet Strength: The key enabler of our approach to identifying OOD objects is the use of Dirichlet strength α_0. Dirichlet distributions applied to DNN architectures exhibit high Dirichlet strength on ID instances. This work builds on the characteristic of Dirichlet strength to exhibit high values for ID instances that still may have higher uncertainty values, while for OOD instances, the objective is to achieve low Dirichlet strength values. We leverage this key feature in Dirichlet DNNs by extending it in our approach to produce two results for each output instance: an uncertainty estimate and a Dirichlet strength. Uncertainty estimation in this case resembles the inter-class entropy for pixel classification, whereas Dirichlet strength would directly reflect whether an instance refers to an ID or OOD object.

For full functionality of the system, both results are needed to identify ID and OOD instances. Based upon a threshold to distinguish between ID and OOD instances using Dirichlet strength, above the threshold is defined as ID and below it as OOD instance. For ID instances, uncertainty estimation using the predictive entropy, presented in equation 3, is used due to its higher reliability over the use of Dirichlet Strength.

3.3. Intermediate Layer Variational Inference

The concept of Intermediate Layer Variational Inference (ILVI), presented in Figure 2b, modifies a latent layer in the network to take the shape of a multivariate Gaussian distribution with mean and variance, instead of using single point estimates. Studies showed that by adopting this method, stochasticity is introduced allowing for the sampling of points from this layer and consequently improving the uncertainty estimation of the DNN and also its generalization performance [12].

Training a network with this approach would encourage the use of the reparametrization trick from the variational autoencoder implementation [18] by having $\theta = \mu + \sigma \odot \epsilon$, where μ and σ are the mean and standard deviation respectively, and $\odot$ is the pointwise multiplication. This allows the mean and log-variance vectors to remain as the learnable parameters of the network while maintaining the stochasticity of the entire system via the random variable $\epsilon \sim \mathcal{N}(0, I)$ [12].

3.4. Depth Decoder

The depth decoder in the architecture is trained on the disparity maps in the training dataset, where it comprises a per-pixel depth. This is achieved by using a decoder similar in structure to the semantic segmentation decoder but estimates depth values for each pixel rather than a class. The decoder is trained to estimate the depth for pixels that belong to detected instances only. The mean value is calculated for each instance and categorized based on the safety-relevant zones as described in [19] within the post-processing module. According to the scope of this work, it is sufficient to estimate the depth category for each detected instance.

3.5. Architecture Variants

To evaluate the performance of this architecture, we examined two different DNN models. The first model utilized the lightweight MobileNetV3 [20] as the shared backbone. It incorporated semantic and instance segmentation decoders inspired by Panoptic Deeplab [21]. In contrast, the second model employed the computationally more intensive EfficientNet [22] as the shared backbone. It incorporated semantic and instance heads inspired by Efficient Panoptic Segmentation [23].

The MobileNet variant uses the MobileNet backbone and each decoder incorporates an Atrous Spatial Pyramid Pooling (ASPP) module for processing and decoding the ILVI outputs. The instance segmentation module produces two outputs: center points for the instances in the scene and the instance masks for it.

The EfficientNet variant in our architecture utilizes the EfficientNet as the shared backbone, complemented by a 2-way Feature Pyramid Network (FPN). For the instance segmentation head, we employ a variant of the Mask R-CNN architecture [24], which comprises two stages: the Region Proposal Network (RPN) module followed by the Region of Interest (ROI) align module. This approach leverages object proposals to extract features from the FPN encodings, enhancing the accuracy of the instance segmentation process.

3.6. Post-Processing Module

The post-processing module incorporates all outputs and calculates for each instance identified by the instance decoder the following attributes: instance mask, instance class, mean distance, mean uncertainty estimate and mean Dirichlet strength. According to a Dirichlet strength threshold, an instance is assumed ID or OOD if it's higher or lower than the threshold respectively. The threshold was manually set to 60% following to experimental studies to best separate between ID and OOD instances. The depth category is also calculated in the post-processing module as described before and used to eliminate instances not relevant according to the scope of this work.

3.7. Loss Function

Each module in the architecture contributes to the loss function. The loss function looks as follows:

$$\mathcal{L} = \lambda_{weight}\mathcal{L}_{sem} + \mathcal{L}_{ILVI} + \mathcal{L}_{depth} + \mathcal{L}_{ins} \quad (5)$$

To mitigate the issue of overconfidence caused by the imbalanced class distribution in the training dataset, an additional weighting factor λ_{weight} is introduced, which is multiplied by the semantic segmentation loss $\mathcal{L}_{sem}$ (equation 3). This weighting factor is computed as $\lambda_{weight_c} = 1 - \frac{N_c}{\sum N}$, where N represents the total pixel count and N_c represents the pixel count for class c.

Furthermore, the loss function $\mathcal{L}_{ILVI}$, described in detail in [12], is utilized to encourage the layer to model a multivariate Normal Gaussian distribution, promoting more reliable uncertainty estimation. In addition, the depth loss $\mathcal{L}_{depth}$ utilizes the L1 loss to train the DNN based on the per-pixel ground-truth instance depth information.

The instance segmentation loss $\mathcal{L}_{ins}$ varies between the two DNN variants. For the MobileNet variant, it consists of two components: the center points loss $\mathcal{L}_{center}$ (using the mean squared error loss) and the instance offset loss $\mathcal{L}_{offset}$ (using the L1 loss) [21]. On the other hand, for the EfficientNet variant, the instance segmentation loss is composed of five terms: $\mathcal{L}_{ins} = \mathcal{L}_{os} + \mathcal{L}_{op} + \mathcal{L}_{cls} + \mathcal{L}_{bbx} + \mathcal{L}_{mask}$, the objectness score loss $\mathcal{L}_{os}$, the object proposal loss $\mathcal{L}_{op}$, the classification loss $\mathcal{L}_{cls}$, the bounding box loss $\mathcal{L}_{bbx}$, and the mask segmentation loss $\mathcal{L}_{mask}$. These loss functions are adapted from the Mask R-CNN approach [23, 24].

4. Experiments and Results

In this section, the Dirichlet MLE is experimented alongside the Prior network, Evidential network, and Cross Entropy (CE) as the baseline. Experiments with the MobileNet variant are used to compare the performance between of the Dirichlet MLE and the other approaches. On the other hand, to verify the applicability of the approach on other models, only the CE and Dirichlet MLE are implemented for the EfficientNet variant.

4.1. Datasets

The DNN is trained using the Cityscapes dataset, comprising 3475 finely annotated images; 2975 for training and 500 validation images used for evaluation [25]. The KITTI Vision dataset [26] is also used to assess the generalization capabilities of the DNN and to assure comparable performance on another in-domain dataset.

Moreover, the Fishyscapes Lost and Found dataset [27] is used for evaluating the identification of OOD objects. This dataset well suits the scope our work to assess the OOD instance detection performance as it contains real images with real OOD objects, unlike other datasets using synthetically rendered images or augmenting real images with synthetic objects.

4.2. Segmentation Performance

Results in Table 1 show the performance of the DNN on diverse datasets to ensure its generalization capabilities on ID and OOD datasets.

Instance segmentation performance is represented by average precision (AP). Further evaluation metrics are AP 50% for an overlap value of 50%, AP 50m and AP 100m where instances up to 50m and 100m, respectively, are only included. Additionally, the semantic segmentation perfor-

Table 1: Instance and semantic segmentation performance comparison for all approaches on in-domain (ID) datasets. Results indicate the high performance of the Dirichlet MLE on both datasets and in both variants.

	Cityscapes					KITTI		
	AP	AP 50%	AP 50m	AP 100m	mIoU	AP	AP 50%	mIoU
MobileNet Variant								
CE	22.9	38.2	35.8	31.9	65.2	23.5	39.1	47.6
Prior	23.1	39.8	37.5	33.2	66.7	22.1	38.7	48.2
Evidential	23.8	42.9	39.8	35.6	68.1	22.9	40.6	50.1
Dirichlet MLE	**26.3**	**45.1**	**43.5**	**42.2**	**69.1**	**23.8**	**42.1**	**51.3**
EfficientNet Variant								
CE	31.8	48.1	45.5	41.6	72.5	22.4	43.2	54.1
Prior	31.5	47.5	45.9	42.1	72.7	22.2	43.5	54.6
Evidential	32.3	50.7	47.5	43.9	73.8	24.4	44.6	55.9
Dirichlet MLE	**32.5**	**51.3**	**48.1**	**44.2**	**74.1**	**24.6**	**45.1**	**56.9**

Table 2: Calibration and Accuracy vs. Certainty Metrics. Dirichlet MLE shows improved uncertainty estimation performance when compared to the other approaches, indicating reliable estimates for in-domain (ID) data.

	Calibration (↓)	Accuracy vs. Certainty (%)(↑)		
	ECE	P(A|C)	P(U|I)	AvU
MobileNet Variant				
CE	4.9	50.9	24.4	51.8
Prior	**4.7**	72.7	17.5	38.5
Evidential	5.2	53.2	63.1	61.2
Dirichlet MLE (Uncertainty)	4.8	**85.4**	**78.1**	**70.1**
EfficientNet Variant				
CE	4.9	56.7	20.6	54.2
Dirichlet MLE (Uncertainty)	**4.8**	**81.2**	**72.4**	**60.1**

mance is shown in terms of mean intersection over union (mIoU).

The Dirichlet MLE DNN exhibits superior performance in segmentation and instance detection tasks on the Cityscapes dataset. This superiority is evident in both the lightweight MobileNet version and the EfficientNet version of the network. Furthermore, the performance is also observed in the semantic and instance segmentation results on the KITTI dataset.

4.3. Uncertainty Estimation Performance

It is essential to make sure that improving OOD detection does not hinder the uncertainty estimation on ID data. For that, two sets of experiments are conducted to examine the uncertainty estimation on the Cityscapes validation dataset for the calibration, and accuracy vs. certainty metrics. Since the experiments are done on ID data, uncertainty estimation using predictive entropy will be used for the Dirichlet MLE.

Calibration: A DNN should be able to provide a calibrated confidence measure in addition to its prediction. In other words, the probability associated with the predicted class label should reflect its ground truth correctness likelihood. The applied calibration metric is the expected calibration error (ECE) using uncertainty estimates.

Accuracy vs. Certainty: It is essential for a DNN equipped with uncertainty estimation to deliver reliable certainty on its correct and incorrect predictions. Accordingly, three conditional probabilities are needed for this evaluation test: $p(accurate|certain) = \frac{n_{ac}}{n_{ac}+n_{ic}}$, $p(uncertain|inaccurate) = \frac{n_{iu}}{n_{ic}+n_{iu}}$ and $AvU = \frac{n_{ac}+n_{iu}}{n_{ac}+n_{au}+n_{ic}+n_{iu}}$. To calculate the probabilities, four fundamental components of the metrics are first calculated: accurate and certain (n_{ac}), accurate and uncertain (n_{au}), inaccurate and certain (n_{ic}), and inaccurate and uncertain (n_{iu}) are calculated. The metric AvU, which stands for accuracy vs. uncertainty, provides insights into the probability of obtaining accurate and certain predictions or inaccurate and uncertain predictions from the network [28].

With varying uncertainty thresholds and the calculation of the conditional probabilities, each value is recorded at increasing steps of uncertainty thresholds. The area under the curve is then calculated, with higher values corresponding to better performance. In this work, we formulate the metrics to accommodate instances where the accuracy is perceived as whether the instance has a 50% overlap or more with its ground-truth counterpart mask, and the uncertainty value is taken as the mean value of the instance-uncertainty mask.

Results Overview: Tables 2 shows the calibration result indicating that all approaches have comparable instance calibration performance when compared to the baseline. It is worth noting that no post-calibration methods are incorporated.

The accuracy vs. certainty results show strongly improved performance for the Dirichlet approach, especially for the P(U|I) metric. Taking a closer look, we can see that we have a high improvement in the P(U|I) whilst still maintaining high performance on the other two metrics. This is not frequently observed as any method trying to improve uncertainty representation would come to a cost of reduced performance on the other two metrics.

4.4. Distributional Separation Efficiency

An ideal network should be able to show high certainty in its correct predictions and low certainty in its incorrect predictions. Accordingly, we aim to quantify the efficiency of the DNN to differentiate between correct and incorrect predictions by plotting their corresponding certainty distribution for both cases. The Cityscapes and the Lost and Found datasets are both used to compare the output characteristics of each approach. The distributions are then compared using the Wasserstein distance metric, where a high value indicates dissimilar distinctive distributions and vice versa for a low Wasserstein distance value.

Distribution plots are presented in Figure 3 and their corresponding Wasserstein distance values in Table 4. Even though the two state-of-the-art approaches have great improvements in the distribution separation for the ID Cityscapes, the Dirichlet uncertainty and strength show even better separation. Greater separation is further distinctive for the OOD Lost and Found dataset. Both Dirichlet uncertainty and Dirichlet strength outperform the other three approaches. This can be also seen in Figure 3 where the separation of both Dirichlet representations have significant separation distinction when compared to the other separation plots. With regards to EfficientNet Variant, the same results can be seen where Dirichlet outperforms the CE baseline in both ID and OOD datasets.

4.5. Depth Estimation

In this section, we assess the depth categorization of the detected true positive instances. The DNN outputs an estimated mean depth for each instance, for each instance categorized in 4 categories as shown in Table 5, according to [19]. Results from Table 5 show that the DNN has sufficiently high categorization quality as needed for the scope of this work.

Figure 3: Distribution Separation Plots for in-domain (ID) and out-of-domain (OOD) dataset. Dirichlet MLE demonstrates superior separation for both uncertainty and Dirichlet strength.

Table 5: Depth categorization performance of the DNN architecture at different depth ranges.

Distance Range	True Positive Categorization Quality (%)(↑)
0- 7.5 m	96.5
7.5 - 15 m	88.2
15 - 25 m	83.1
25 - 37.5 m	79.3

4.6. OOD Instance Segmentation and Identification Performance

Being able to identify OOD instances in the scene, not only on the pixel level, is crucial for the safety of an AD system. For that, the OOD Detection Rate (DR) of instance segmentation is assessed with regard to 25m, 50m and 100m. In this context, DR is the ratio of how many OOD objects have been detected, where detected OOD instances indicates 50% overlap or more between the instance detection and the ground truth OOD object.

To evaluate the DNN's ability to identify OOD objects, we calculate the identification rate of OOD objects (iOOD). In this context, the identifier for each approach is based on the Dirichlet MLE using Dirichlet strength, and the baseline and state-of-the-art approaches using predictive entropy respecitively. For the Dirichlet MLE, the average Dirichlet strength of each object is calculated at different distances of 25m, 50m, and 100m. Again, for the other approaches, predictive entropy is used.

Table 3 shows the results for DR and iOOD, and demonstrates the significance of the proposed Dirichlet approach. It exhibits a higher detection rate of the OOD objects, with significantly high iOOD values. This reflects the sample results in Figure 1, where one sample from the Lost and Found dataset is presented with their respective uncertainty result for all four approaches. The OOD results for the Dirichlet MLE reflect the results shown in Table 3, where the DNN is able to segment the OOD objects and also with very high iOOD.

4.7. Discussion

The results demonstrate the effectiveness of the Dirichlet MLE DNN outperforming state-of-the-art and the baseline DNNs. The baseline DNN does not support the identification of OOD objects, whilst the state-of-the-art approaches only partially segment the OOD objects or do not deliver high uncertainty values on pixel levels compared to other areas of the image.

The generated samples presented in Figure 4 demonstrate the output of the Dirichlet MLE. It is observed that the differences between uncertainty estimation and Dirichlet strength for ID samples are minimal. However, noticeable differences emerge for OOD samples, where OOD objects exhibit low levels of Dirichlet strength, while uncertainty estimation reflects low certainty not only for OOD

Table 3: Out-of-domain (OOD) Instance Detection and Identification Performance. Dirichlet MLE shows an improved detection rate (DR) over the other approaches, but an even greater improvement in the identification rate of OOD objects (iOOD).

	Detection Rate (DR) (%)(↑)			Identification rate of OOD objects (iOOD) (%)(↑)		
	DR_{25m}	DR_{50m}	DR_{100m}	$iOOD_{25m}$	$iOOD_{50m}$	$iOOD_{100m}$
MobileNet Variant						
CE	45.2	34.2	24.9	14.2	16.8	17.5
Prior	50.2	37.6	29.7	18.6	19.3	19.1
Evidential	53.2	39.9	33.3	11.7	13.3	13.3
Dirichlet MLE (Dir. Strength)	**59.5**	**43.7**	**38.3**	**65.4**	**56.1**	**47.1**
EfficientNet Variant						
CE	40.2	27.1	23.9	8,9	8.3	6.2
Dirichlet MLE (Dir. Strength)	**46.7**	**32.2**	**28.1**	**66.1**	**54.3**	**43.7**

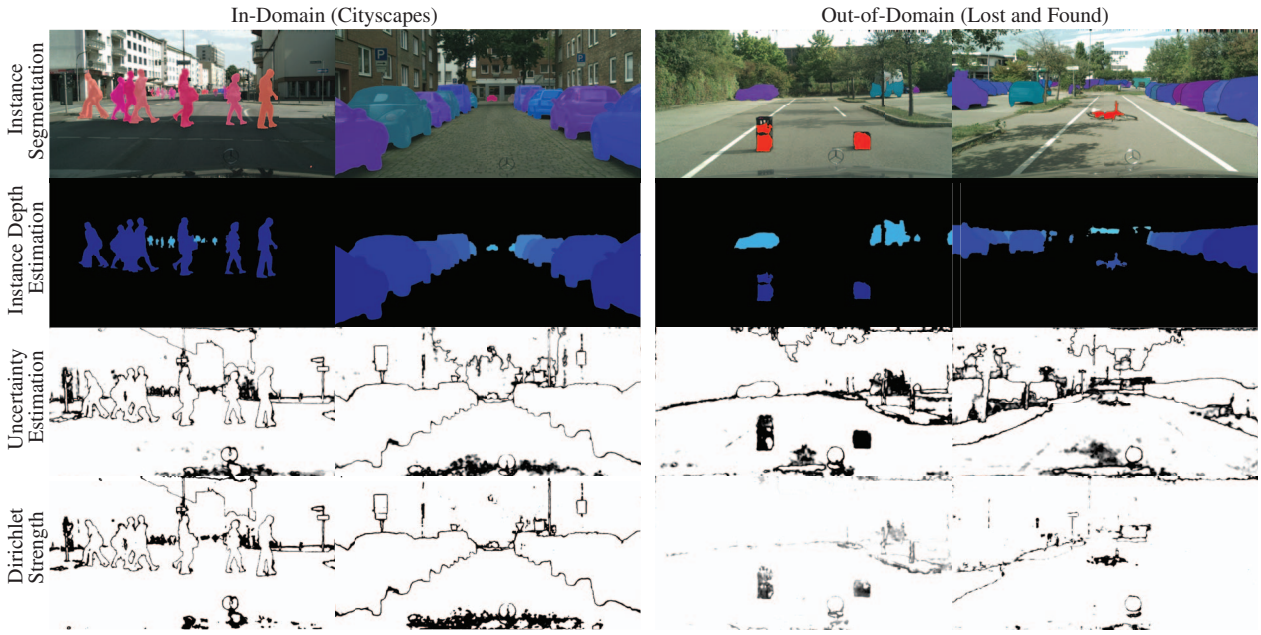


Figure 4: Sample results from the Dirichlet MLE approach for both ID and OOD data. For ID dataset, uncertainty estimation and Dirichlet strength have high levels of uncertainty estimates on similar features in the image. On the other hand, for the OOD samples, Dirichlet strength is low only on the OOD instances unlike uncertainty estimation having high uncertainty on both ODD and ID instances.

objects but also for other instances in the scene. This highlights the complementary nature of both methods, emphasizing the need to utilize both for accurate OOD object identification and reliable uncertainty estimation.

The Dirichlet MLE DNN exhibits exceptional performance in accurately estimating low certainty for incorrect predictions and vice versa. In addition to achieving comparable or even enhanced results for ID objects, the Dirichlet MLE DNN surpasses other network architectures by effectively distinguishing between ID and OOD objects. This distinction is reinforced by its improved separation capabilities, highlighting the superior discriminative power of the Dirichlet MLE DNN. The Dirichlet MLE DNN with its combination of instance segmentation and depth branches together with the post-processing module delivers superior OOD instance results.

5. Conclusion

In this study, we propose an instance segmentation architecture that effectively identifies OOD objects in AD systems. Our framework combines the Dirichlet MLE approach and the ILVI method, resulting in superior OOD instance detection while maintaining robustness on ID data. By leveraging the Dirichlet strength, we successfully differentiate between ID and OOD instances, as demonstrated through comprehensive comparisons with state-of-the-art approaches and a baseline method. Notably, we have evaluated the performance using different backbones and instance segmentation models within a model-agnostic architecture.

Table 4: Distributional Separation Efficiency Results. The Dirichlet MLE has a high Wasserstein distance reflecting its efficient OOD object identification performance.

	Wasserstein Distance (↑)	
	Cityscapes	Lost and Found
MobileNet Variant		
CE	1.1	1.3
Prior	2.3	1.8
Evidential	3.3	0.9
Dirichlet MLE (Uncertainty)	**4.9**	5.2
Dirichlet MLE (Dir. Strength)	4.2	**7.1**
EfficientNet Variant		
CE	0.9	0.5
Dirichlet MLE (Uncertainty)	**3.6**	4.1
Dirichlet MLE (Dir. Strength)	3.2	**5.9**

By adopting our approach, the AD maneuver planner benefits from accurate instance segmentation results, reliable certainty estimation, identification of OOD instances, and estimated depth ranges. These findings highlight the potential of the Dirichlet MLE DNN architecture to enhance the perception capabilities of AD systems, contributing to safer and more efficient AD systems. Future research can further explore applications and potential enhancements of this architecture to address the evolving challenges in autonomous driving.

ACKNOWLEDGMENT

This work is partly funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) and partly financed by the European Union in the frame of NextGenerationEU within the project ”Solutions and Technologies for Automated Driving in Town” (FKZ 19A22006P).

References

[1] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[2] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

[3] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022.

[4] Szilárd Aradi. Survey of deep reinforcement learning for motion planning of autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[5] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[6] Timo Sämann, Peter Schlicht, and Fabian Hüger. Strategy to increase the safety of a dnn-based perception for had systems. *arXiv preprint arXiv:2002.08935*, 2020.

[7] Yarin Gal. Uncertainty in deep learning. 2016.

[8] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021.

[9] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022.

[10] Ahmed Hammam, Frank Bonarens, Seyed Eghbal Ghobadi, and Christoph Stiller. Predictive uncertainty quantification of deep neural networks using dirichlet distributions. In *Proceedings of the 6th ACM Computer Science in Cars Symposium*, pages 1–10, 2022.

[11] Ahmed Hammam, Frank Bonarens, Seyed Eghbal Ghobadi, and Christoph Stiller. Towards improved intermediate layer variational inference for uncertainty estimation. In *European Conference on Computer Vision*, pages 1–14. Springer, 2022.

[12] Ahmed Hammam, Seyed Eghbal Ghobadi, Frank Bonarens, and Christoph Stiller. Real-time uncertainty estimation based on intermediate layer variational inference. In *Computer Science in Cars Symposium*, pages 1–9, 2021.

[13] Thomas Minka. Estimating a dirichlet distribution, 2000.

[14] Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. In *International Conference on Learning Representations*, 2020.

[15] Giancarlo Di Biase et al. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16918–16927, June 2021.

[16] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[17] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] D Brüggemann, Robin Chan, Hanno Gottschalk, and Stefan Bracke. Software architecture for human-centered reliability assessment for neural networks in autonomous driving. In *11th IMA International Conference on Modelling in Industrial Maintenance and Reliability (MIMAR)*, volume 2, page 11, 2021.

[20] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, October 2019.

[21] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020.

[22] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[23] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision (IJCV)*, 2021.

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

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

[26] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[27] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[28] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.