Probing Contextual Diversity for Dense Out-of-Distribution Detection

Silvio Galesso, Maria Alejandra Bravo, Mehdi Naouar, and Thomas Brox

University of Freiburg

Abstract. Detection of out-of-distribution (OoD) samples in the context of image classification has recently become an area of interest and active study, along with the topic of uncertainty estimation, to which it is closely related. In this paper we explore the task of OoD segmentation, which has been studied less than its classification counterpart and presents additional challenges. Segmentation is a dense prediction task for which the model's outcome for each pixel depends on its surroundings. The receptive field and the reliance on context play a role for distinguishing different classes and, correspondingly, for spotting OoD entities. We introduce MOoSe, an efficient strategy to leverage the various levels of context represented within semantic segmentation models and show that even a simple aggregation of multi-scale representations has consistently positive effects on OoD detection and uncertainty estimation.

Keywords: Out-of-Distribution Detection, Semantic Segmentation

1 Introduction

Imagine you see a pattern, an object, or a scene configuration you do not know. You will identify it as novel and it will attract your attention. This ability to deal with an open world and to identify novel patterns at all semantic levels is one of the many ways how human perception differs from contemporary machine learning. Most deep learning setups assume a closed world with a fixed set of known classes to choose from. However, many real-world tasks do not match this assumption. Very often, maximum deviations from the training samples are the most interesting data points.

Accordingly, novelty/anomaly/out-of-distribution detection has attracted more and more interest recently. Outside of data regimes with limited variation, such as in industrial inspection [14,50], the common approaches to identify unseen patterns derive uncertainty estimates from an existing classification model and mark samples with large uncertainty as novel or out-of-distribution [41,31,3,27,56]. This approach comes with a conflict between the classifier focusing on features that help discriminate between the known classes and the need for rich and diverse features that can identify out-of-distribution patterns. This is especially true for semantic segmentation, where a pixel's class is not only defined by its own appearance, but also by the context it appears in. Based on context information only and ignoring appearance, a segmentation model could assume that a large

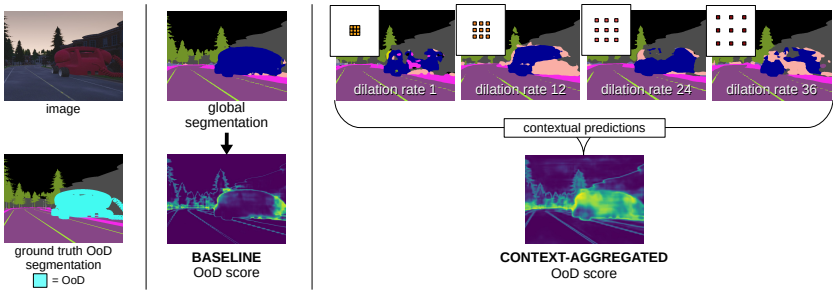


Fig. 1: Our method obtains predictions based on diverse contextual information from different dilated convolutions, exploiting the hierarchical structure of semantic segmentation architectures. On the anomalous pixels (cyan in the ground truth) the contextual predictions diverge, allowing us to improve upon the global model's uncertainty score, which is overconfident in classifying the object as a car. From improved uncertainty we get better out-of-distribution detection. More details in Figure 2 and Section 3.2.

animal (or an oversized telephone, like in the example in Figure 1) in the middle of the road is a vehicle, while based on local appearance only it could believe that pictures on a billboard are, for example, actual people in the flesh.

In order to combine context and local appearance, modern segmentation networks feature modules with different receptive fields and resolutions, designed to extract diverse representations including different amounts of contextual cues. While for known objects the different cues mostly align with the model's notion of a semantic class, in the case of novel objects the representations at multiple context levels tend to disagree. This can be used as an indicator for uncertainty. Indeed, our approach develops on this idea by having multiple heads as probes for comprehensive multi-scale cues, and obtains an aggregated uncertainty estimate for out-of-distribution (OoD) segmentation. We show that this strategy improves uncertainty estimates over using a single global prediction and often even over regular ensembles, while being substantially more efficient than the latter. It also sets up the bar on the common benchmarks for OoD segmentation. We call our model MOoSe, for **M**ulti-head **OoD** **Se**gmentation. Source code available at `https://github.com/MOoSe-ECCV22/moose_eccv2022`.

2 Related Work

Out of Distribution Detection Out-of-distribution detection is closely related to uncertainty estimation. Under the assumption that a model should be uncertain about samples far from its training distribution, model uncertainty can be used as a proxy score for detecting outliers [29,37]. Several methods for OoD detection, including ours, rely on an existing model trained on a semantic task on the in-distribution data, such as image classification or segmentation [29,28].

Different techniques have been developed to improve outlier detection by means of uncertainty scores, either at inference time [41,28] or while learning the representations [27,31,3,56,30].

Another set of methods for uncertainty estimation is inspired by Bayesian neural networks, which produce probabilistic predictions [6]. For example Monte-Carlo Dropout [21,34] approximates the predictive distribution by sampling a finite number of parameter configurations at inference time using random dropout masks. Although arguably not strictly Bayesian, ensembles [37,55] also approximate the predictive distribution by polling a set of independently trained models fixed at inference time. Attributes of the predictive distribution, such as its entropy, can be used as a measure for uncertainty [53,1].

Alternatively, one can directly model the distribution of the training data, and the likelihood estimated by the resulting model can be used to detect outliers [45,62,51,35]. Other approaches rather rely on learning pretext tasks on the in-distribution data as a proxy for density estimation. Examples of such tasks are reconstruction [47,42,57,36,17] and classification of geometric image transformations [22].

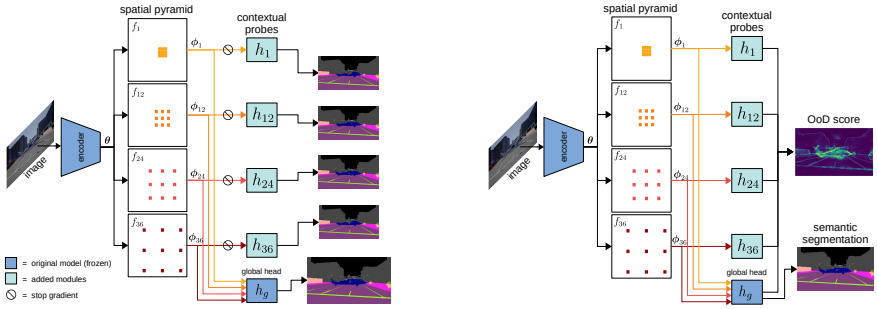
Dense OoD Detection Methods that are effective at recognizing outlier images do not always scale well to dense OoD detection, where individual pixels in each image need to be classified as in-distribution or anomalous. A recent work [28] has found that advanced methods like generative models [2,25,52] and Monte-Carlo dropout [21] are outperformed by metrics derived from the predictions of a pre-trained semantic segmentation network, such as the values of the segmentation logits. Several recent works [23,9,4,7] focus on the improvement of such segmentation by-products. In particular, the practice of Outlier Exposure [30], originally developed for recognition, has recently gained popularity in dense anomaly detection: several approaches revolve around using outlier data during training, either from a real data [9,4,54] or sampled from a generative model [23,36]. While our method does not need outlier exposure to work, we show that it can be beneficially combined with it.

As mentioned above, deep ensembles [37] are a versatile tool and a gold standard for uncertainty estimation, making them a popular choice for anomaly detection [55,38]. Their relative scalability and effectiveness made them a viable option for uncertainty estimation in dense contexts, including anomaly segmentation [19,20]: ensembles are a simple and almost infallible way of improving the quality of uncertainty scores of neural networks.

At the core of ensemble techniques is diversity between models, which is often provided by random weight initialization and data bootstrapping [37,32], sometimes by architectural differences [61]. While these sources of diversity are of proven efficacy and versatility, they are generic and ignore the requirements of the task at hand, introducing significant computational costs. Multi-headed ensembles mitigate this drawback by sharing the largest part of the network and drawing their diversity only from independent, lightweight heads [39,40,46]. Even though it is related to multi-headed ensembles, our method exploits an additional source

of diversity: leveraging a widespread architectural design specific to semantic segmentation, MOoSe captures the variety of contextual information and receptive field within the same model. This allows for performance improvements that are equal or superior to those of bootstrapped ensembles, at a fraction of the computational cost.

3 Multi-Head Context Networks



(a) Training pipeline (b) Evaluation pipeline

Fig. 2: **MOoSe**: illustration of the multi-head architecture based on the DeepLabV3 semantic segmentation model. During training (a) all the probes learn standard semantic segmentation, while the rest of the network (pre-trained) is unaffected. At test time (b) the uncertainties of all heads (contextual and global) are pooled together into an improved scoremap for OoD detection.

3.1 Contextual Diversity in Semantic Segmentation Networks

Consider a semantic segmentation decoder based on the popular spatial pyramid architecture [63,10,11,13], i.e. containing either a Spatial Pyramid Pooling (SPP [59]) or Atrous Spatial Pyramid Pooling (ASPP [10]) structure. Such structures include a series of pooling or dilated convolutional operations applied in parallel to a set of feature embeddings. We refer to these operations as *spatial pyramid modules*. Each spatial pyramid module has a unique pooling scale or receptive field, allowing for scale invariance and providing the decoder with various degrees of context: the larger the receptive field the more global context at the cost of a loss in detail. The outputs of the spatial pyramid modules are concatenated and fed to a segmentation head that produces a single prediction.

In this section we propose a method for extracting the contextual diversity between the representations produced by the different spatial pyramid modules. By exploiting said diversity we are able to improve OoD segmentation performance and network calibration without the need of expensive ensembling of

multiple models. Our method is non-invasive with regard to the main semantic segmentation task and lightweight in terms of computational cost, making it suitable for real-world applications without affecting segmentation performance. In addition, since our approach improves the uncertainty scores of an existing model, it can be easily combined with other state of the art approaches.

3.2 Probing Contextual Diversity

We start from a generic semantic segmentation architecture featuring K *spatial pyramid modules* $\mathcal{F} = \{f_{c_1}, ..., f_{c_K}\}$, each with a different context size c_k, as in Figure 2. The encoder features $\boldsymbol{\theta}$ are fed to the pyramid modules producing a set of contextual embeddings $\Phi = \{\boldsymbol{\phi}_{c_1}, ..., \boldsymbol{\phi}_{c_K}\}$, where $\boldsymbol{\phi}_{c_k} = f_{c_k}(\boldsymbol{\theta})$. The segmentation output is produced by the *global head* h_g, which takes all the concatenated features Φ. We denote the output logits of the global head as $\mathbf{z}_g := h_g(\Phi)$.

Our method consists of a simple addition to this generic architecture, namely the introduction of *probes* to extract context-dependent information from the main model. We pair each spatial pyramid module f_{c_k} with a contextual prediction head h_{c_k}, which is trained to produce segmentation logits $\mathbf{z}_{c_k} = h_{c_k}(\boldsymbol{\phi}_{c_k})$ from the context features of size c_k (and the global pooling features, see [11]). Given the input image $\mathbf{x}$, we denote the prediction distributions as:

$$p(\hat{\mathbf{y}}|\mathbf{x}, h_g) = \text{softmax}(\mathbf{z}_h) \qquad \text{and} \qquad p(\hat{\mathbf{y}}|\mathbf{x}, h_{c_k}) = \text{softmax}(\mathbf{z}_{c_k}) \tag{1}$$

for the global and contextual heads respectively.

We train all the heads using a standard Cross Entropy loss given N classes and the ground truth segmentation $\mathbf{y}$ (we drop the mean operation over the spatial dimensions for simplicity):

$$\mathcal{L}_{\text{CE}}(\mathbf{x}, \mathbf{y}) = -\sum_{k=1}^{K}\sum_{i=1}^{N} \mathbf{y}_i \cdot \log p(\hat{\mathbf{y}}_i|\mathbf{x}, h_{c_k}), \tag{2}$$

although any semantic segmentation objective could be used.

The contextual heads are designed to act as probes and extract information from the context-specific representations. For this reason, we do not back-propagate the gradients coming from the contextual heads to the rest of the network, and only update the weights of the heads themselves. The spatial pyramid modules are distinct operations, each with its specific scope depending on its context size. By stopping the gradients before the spatial pyramid modules we force each head to solve the same segmentation task but using different features, preserving prediction diversity. As a byproduct, our architectural modifications do not interfere with the rest of the network and the main segmentation task.

Head Architecture The architecture of the contextual heads is based on the global head of the base segmentation model. For example, for DeepLabV3 it

consists of a projection block to bring the number of channels down to 256, followed by a sequence of d prediction blocks (3×3 convolution, batch normalization, ReLU), plus a final 1×1 convolution for prediction. The head depth d can be tuned according to the predictive power necessary to process contextual information, depending on the difficulty of the dataset.

3.3 Out-of-Distribution Detection with MOoSe

A model for dense OoD detection should assign to each location in the input image an anomaly score. To obtain per-pixel OoD scores we test three scoring functions, applied to the outputs of the segmentation heads: maximum softmax probability (MSP) [29], prediction entropy (H) [53] and maximum logit (ML) [28]. We adapt each scoring function to work with predictions from multiple heads. The maximum softmax probability is computed on the average predicted distribution over all the heads, including the global head:

$$S_{\text{MSP}} = - \max_{i \in [1,N]} \left[\frac{1}{K+1} \left(p(\hat{\mathbf{y}}_i | \mathbf{x}, h_g) + \sum_{k=1}^{K} p(\hat{\mathbf{y}}_i | \mathbf{x}, h_{c_k}) \right) \right], \tag{3}$$

Similarly, for the entropy we compute the entropy of the expected output distribution:

$$S_{\text{H}} = \mathcal{H} \left[\frac{1}{K+1} \left(p(\hat{\mathbf{y}} | \mathbf{x}, h_g) + \sum_{k=1}^{K} p(\hat{\mathbf{y}} | \mathbf{x}, h_{c_k}) \right) \right], \tag{4}$$

where $\mathcal{H}$ denotes the information entropy. For maximum logit we average the logits over the different heads and compute their negated maximum:

$$S_{\text{ML}} = - \max_{i \in [1,N]} \left[\frac{1}{K+1} \left(\mathbf{z}_{g,i} + \sum_{k=1}^{K} \mathbf{z}_{c_k,i} \right) \right]. \tag{5}$$

All scores should be directly proportional to the model's belief of a pixel belonging to an anomalous object, therefore for MSP and ML the negatives are taken.

4 Experiments

In this section we evaluate our approach on out-of-distribution detection, comparing it to ensembles (Section 4.4) and to the state of the art (Section 4.5).

4.1 Datasets & Benchmarks

StreetHazards [28] is a synthetic dataset for semantic segmentation and OoD detection. It features street scenes in diverse settings, created with the CARLA simulation environment [18]. The 1500 test samples feature instances from 250 different anomalous objects, diverse in appearance, location, and size.

The **BDD-Anomaly** [28] dataset is derived from the BDD100K [60] semantic segmentation dataset by removing the samples containing instances of the motorcycle, bicycle, and train classes, and using them as a test set for OoD segmentation, yielding a 6280/910/810 training/validation/test split. BDD-Anomaly and StreetHazards constitute the CAOS benchmark [28].

Fishyscapes - LostAndFound [49,5] is a dataset for road obstacle detection, designed to be used in combination with the Cityscapes [15] driving dataset. Its test split contains 1203 images of real street scenes featuring road obstacles, whose presence is marked in the segmentation ground truth.

RoadAnomaly [42] consists of 60 real world images of diverse anomalous objects in driving environments, collected from the internet. The images come with pixel-wise annotations of the anomalous objects, making them suitable for testing models trained on driving datasets.

Results for the anomaly track of the SegmentMeIfYouCan [8] benchmark can be found in the supplementary material.

4.2 Evaluation Metrics

We evaluate OoD detection performance using the area under the precision-recall curve (AUPR), and the false positive rate at 95% true positive rate (FPR_{95}). As is customary, anomalous pixels are considered positives. Other works include results for the area under the ROC curve (AUROC), however the AUPR is to be preferred to this metric in the presence of heavy class imbalance, which is the case for anomaly segmentation [16].

Table 1: **CAOS benchmark.** Comparison between single model/global head (Global), multi-head ensembles (MH-Ens), standard deep ensembles (DeepEns) and MOoSe on dense out-of-distribution detection. The results are for DeepLabV3 and PSPNet models with ResNet50 backbones. All three scoring functions (maximum softmax probability (MSP), entropy (H), maximum logit (ML)) are considered. All results are percentages, best results are shown in **bold**

		StreetHazards				BDD-Anomaly			
		DeepLabV3		PSPNet		DeepLabV3		PSPNet	
Score fn.	Method	AUPR↑	FPR_{95} ↓	AUPR↑	FPR_{95} ↓	AUPR↑	FPR_{95} ↓	AUPR↑	FPR_{95} ↓
MSP	Global	9.11	22.37	9.65	22.04	7.01	22.47	6.75	23.63
	MH-Ens	9.69	21.40	9.84	22.49	7.55	25.50	8.07	23.41
	DeepEns	10.22	21.09	10.61	**20.75**	7.64	**21.53**	**8.52**	**21.31**
	MOoSe	**12.53**	**21.05**	**11.28**	21.94	**8.66**	22.49	8.11	24.09
H	Global	11.89	22.07	12.28	21.77	10.23	20.64	9.89	21.69
	MH-Ens	12.59	21.10	12.45	22.29	10.62	23.51	11.73	20.76
	DeepEns	13.43	20.62	13.39	**20.35**	11.39	19.31	12.32	**18.83**
	MOoSe	**15.43**	**19.89**	**14.52**	21.20	**12.59**	**19.27**	**12.35**	20.98
ML	Global	13.57	23.27	13.43	27.71	10.69	15.60	10.68	16.79
	MH-Ens	13.99	21.86	13.64	28.30	10.69	20.19	12.40	15.08
	DeepEns	14.57	21.79	14.14	25.82	11.40	14.66	12.26	13.96
	MOoSe	**15.22**	**17.55**	**15.29**	**20.46**	**12.52**	**13.86**	**12.88**	**13.94**

4.3 Experimental Setup

MOoSe relies on semantic segmentation models to perform dense OoD detection. For the experiments on StreetHazards and BDD-Anomaly we report results for two convolutional architectures, (DeepLabV3 [10] and PSPNet [63], each with ResNet50 and ResNet101 backbones) and one transformer based (Lawin [58], see supplementary material). For the experiments on LostAndFound and RoadAnomaly we use DeepLabV3+ [12] with a ResNet101 backbone, trained on Cityscapes[1] or BDD100k [60] respectively.

Training We build MOoSe on top of fully trained semantic segmentation networks, by adding the prediction heads and training them jointly for segmentation on the respective dataset, using a standard pixel-wise cross-entropy loss. Although nothing prevents from training the whole model together, for fairness of comparison we only apply the loss to the probes. In order to prevent any alteration to the main model while training the heads, we stop gradient propagation through the rest of the network and make sure that the normalization layers would not update their statistics during forward propagation. The heads are trained for 80 epochs, or until saturation of segmentation performance (mIoU).

MOoSe introduces two hyperparameters: learning rate and depth d of the contextual heads. By default we use $d = 1$ for the models trained on StreetHazards and $d = 3$ otherwise. While the performance gains depend on these, we find that our method is robust to configuration changes, as we show in an ablation study in the supplementary material.

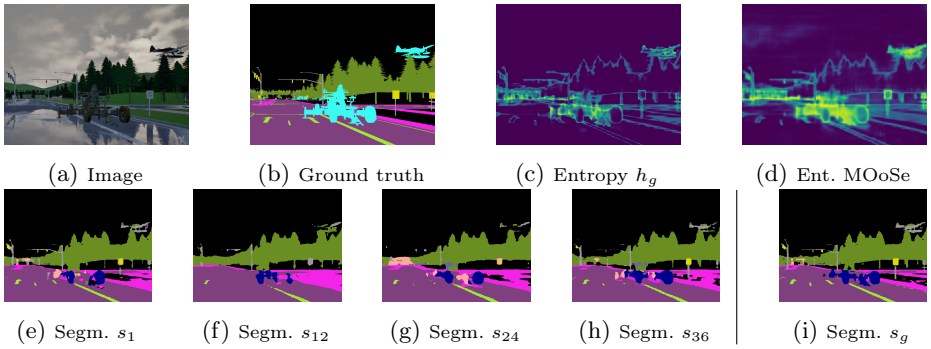
4.4 Comparison with Ensembles

In this section we compare MOoSe with the single prediction baseline (global head) and with two types of ensembles. Deep ensembles [37] (DeepEns) consist of sets of independent segmentation networks, each trained on a different random subset of 67% of the original data, starting from a different random parameter initialization [32]. Similarly, multi-head ensembles (MH-Ens) are trained on random data subsets, but share the same encoder and only feature diverse prediction heads, for increased efficiency.

We compare to ensembles with 5 members/heads to match the number of heads in our method. Additionally, we pick the ensemble member with the median AUPR performance to serve both as the single model baseline and as initialization for MOoSe. The shared backbone of the multi-head ensembles also comes from the same model.

Table 1 shows results for the CAOS benchmark (StreetHazards and BDD-Anomaly) using DeepLabV3 and PSPNet as base architectures, with ResNet50 [26] backbones. We report results for the three OoD scoring functions described in Section 3.3; results for MOoSe are averaged over 3 runs, standard deviations are

[1] Parameters available at:
`https://github.com/NVIDIA/semantic-segmentation`



(a) Image (b) Ground truth (c) Entropy h_g (d) Ent. MOoSe

(e) Segm. s_1 (f) Segm. s_{12} (g) Segm. s_{24} (h) Segm. s_{36} (i) Segm. s_g

Fig. 3: **OoD segmentation & head contribution.** Test image from the StreetHazards dataset: a street scene containing anomalous objects (indicated in cyan in the ground truth (b)). The contextual predictions (e-h) diverge on the outliers, improving the entropy score map (c, d). The example shows an interesting failure case: the street sign (in-distribution) on the right also sparks disagreement between the heads, resulting in increased entropy and thus a false positive.

available in the supplementary material. For all datasets, architectures, scoring functions, and metrics, MOoSe consistently outperforms its respective global head. Similarly, MOoSe outperforms multi-head ensembles, as well as deep ensemble in most cases, while having a smaller computational cost than both.

In accordance with what observed in other works [28], the maximum-logit scoring function tends to outperform entropy, most notably in terms of FPR_{95} and on the BDD-Anomaly dataset. Both scoring functions consistently outperform maximum-softmax probability. Moreover, maximum-logit appears to combine well with MOoSe by effectively reducing false positives. Results for models using the ResNet101 backbone are available in the supplementary material.

Figure 3 shows an example for OoD segmentation on a driving scene. The top row compares the entropy obtained using the global head (3c) and our multi-head approach (3d). The probes of MOoSe disagree on the nature of the anomalous objects in the image, and its aggregated entropy score is able to outline the anomalous objects more clearly than the global head. However, prediction disagreement also produces false positives for smaller inlier objects, such as the street sign on the right, highlighting a possible failure mode of our approach.

Computational costs In Table 2 we compare our method against ensembles in terms of computational costs, reporting the number of parameters of each model and the estimated runtime of a forward pass. We consider DeepLabV3 and PSPNet with ResNet50 and MOoSe head depth 1. Deep ensembles have the highest parameter count and runtime, 5 times that of a single network. MOoSe compares favorably to both ensembles on all architectures. The larger size and runtime of PSPNet compared to DeepLab is due to its higher dimensional representations, which can be reduced with projection layers before the probes.

Table 2: **Computational costs.** Estimated computational costs of MOoSe in comparison with ensembles. We report the number of parameters (in millions) and the estimated forward pass runtime on StreetHazards (in milliseconds), estimated on a single Nvidia RTX2080Ti GPU using the PyTorch [48] benchmarking utilities

Architecture		Single	MH-Ens	DeepEns	MOoSe
DeepLabV3	parameters (M)	40	104	198	43
	runtime (ms)	113	286	583	121
PSPNet	parameters (M)	47	139	233	94
	runtime (ms)	107	246	542	183

4.5 Comparison with the State of The Art

Here we compare MOoSe with the best approaches for dense OoD detection that do not require negative training data (see Section 4.5).

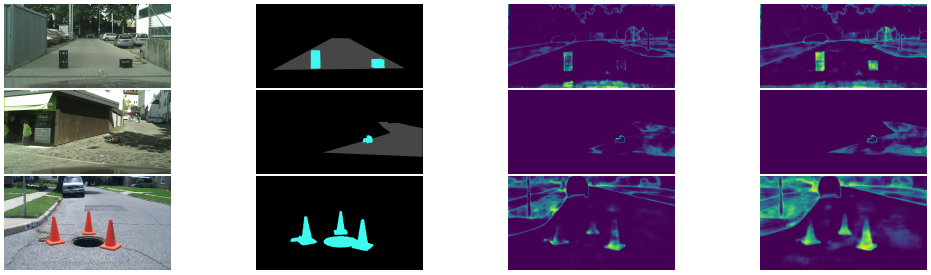
The CAOS Benchmark On StreetHazards and BDD-Anomaly we compare with TRADI [20], SynthCP [57], OVNNI [19], Deep Metric Learning (DML) [7], and the approach by Grcic et al. [23] that uses outlier exposure with generated samples. TRADI and OVNNI require multiple forward passes per sample, increasing the evaluation run-time (or memory requirements) considerably. Table 3(a) shows that MOoSe compares favorably to existing works on both datasets and on all metrics. We note that, given its non-invasive nature, MOoSe is compatible with other approaches, and can for example be combined with the loss of DML.

Table 3: **State-of-the-art comparison. Left - CAOS benchmark:** MOoSe in combination with the max-logit scoring function, outperforms all other methods on StreetHazards, except for DML in terms of FPR_{95}. On BDD-Anomaly MOoSe performs the best in both metrics. **Right**: MOoSe yields improvements on both **Fishyscapes LostAndFound (FS - LaF)** and **RoadAnomaly**, but on the former benchmark is outperformed by Standardized Max-Logits (Std.ML).

	Street Hazards		**BDD Anomaly**	
	AUPR	FPR_{95}	AUPR	FPR_{95}
TRADI[20]	7.2	25.3	5.6	26.9
SynthCP[57]	9.3	28.4	-	-
OVNNI[19]	12.6	22.2	6.7	25.0
Grcic[23]	12.7	25.2	-	-
DML[7]	14.7	**17.3**	-	-
MOoSe ML	**15.22**	17.55	**12.52**	**13.86**

		FS - LaF		**RoadAnomaly**	
	Method	AUPR	FPR_{95}	AUPR	FPR_{95}
MSP	Global	3.06	37.46	23.76	51.32
	MOoSe	7.13	33.72	31.53	43.41
H	Global	6.23	37.34	32.00	49.14
	MOoSe	12.08	32.58	<u>41.48</u>	<u>36.78</u>
ML	Global	10.25	37.45	37.86	39.03
	MOoSe	<u>13.64</u>	<u>32.32</u>	**43.59**	**32.12**
Resynth. [42]		5.70	48.05	-	-
DML [7]		-	-	37	37
Std.ML [33]		**31.05**	**21.52**	25.82	49.74

LostAndFound, RoadAnomaly We extend our evaluation of MOoSe to other real world benchmarks, Fishyscapes LostAndFound and RoadAnomaly, using



(a) Image (b) OoD G.T. (c) Entropy h_g (d) Ent. MOoSe

Fig. 4: **OoD segmentation.** Examples on LostAndFound and RoadAnomaly, anomalous objects shows in cyan in the second column. The first row shows an example in which our model is able to recognize anomalous objects in their entirety, where the global head fails. The second row shows a failure case, where our model only marks the borders of the obstacle on the road and produces false positives, performing worse than the global head. The example in the last row is from RoadAnomaly: MOoSe detects the traffic cones better than the global head, but introduces noise in the background and still fails to detect the manhole.

models trained on Cityscapes and BDD100k [60] respectively, as described in Section 4.3. We report our results in Table 3(b) and include results for the comparable (not needing negative training data) state-of-the-art methods DML, Standardized Max-Logits [33] and Image Resynthesis [42]. Similarly to the CAOS benchmark, we can observe that the adoption of MOoSe improves OoD detection performance on both benchmarks, regardless of the chosen scoring function. In the supplementary material we include results for the SegmentMeIfYouCan benchmark (similar to RoadAnomaly), where Image Resynthesis is currently SOTA. On the other hand, Table 3(b) shows that Image Resynthesis does not perform well on LostAndFound. Standardized Max-Logits has remarkable results on LostAndFound but not on RoadAnomaly, where MOoSe works best.

Figure 4 shows examples of OoD detection on LostAndFound and RoadAnomaly. In the first example MOoSe improves over the entropy heatmap of the global head. In the second example, however, it can be seen that our method still fails to detect the obstacles and produces more false positives. Increased false positives are visible in the background of the third example too, although here MOoSe also improves the detection of some anomalous objects.

Outlier Exposure Several methods for anomaly segmentation rely on negative training data from a separate source. While this technique introduces some drawbacks, such as a reliance on the choice of the negative data and a potential negative impact on segmentation, it has been shown to improve OoD detection on the common benchmarks. Following the procedure described in [9] as "entropy training", we investigate whether our method can also benefit from outlier exposure. Indeed, results show that outlier exposure boosts MOoSe+ML to 53.19

AUPR (+22%) and 24.38 FPR_{95} (-24%) on RoadAnomaly. Full results on all scoring functions are available in the supplementary material.

5 Analysis

Our approach relies on a collection of different predictions to improve OoD detection. Previous literature on network ensembles puts the spotlight on diversity [43,39], emphasizing that multiple estimators can be helpful only if their predictions are diverse and each contributes with useful information for the cumulative decision. In this section we address some points to better understand the working principle of the method and verify its underlying hypotheses. Specifically, we investigate: 1) the effect of MOoSe on prediction diversity, 2) whether contextual aggregation can be responsible for prediction diversity, and 3) how this translates into better OoD detection.

5.1 Quantifying Diversity: Variance and Mutual Information

We are interested in comparing MOoSe to the closely related ensembles in terms of prediction diversity, of which a simple metric is variance. We compare the average variance of the output distributions of MOoSe and ensembles on StreetHazards and BDD-Anomaly validation, as reported in Table 4 (left). On both datasets our method's predictions have higher variance than both ensembles.

Variance, however, gives us no insights on what the predictions disagree upon, and is therefore of limited interest. From the literature on Bayesian networks we can borrow a more informative metric: the mutual information (MI) between the model distribution and the output distribution [44]. Consider an ensemble of K networks, or a multi-head model with K heads. Each model or head produces a prediction $p(\hat{y}|x,k)$. We can compute the mutual information between the distribution of the models k and the distribution of their predictions as:

$$\text{MI}(\hat{y},k|x) = \mathcal{H}\left[\frac{1}{K}\sum_{k=1}^{K} p(\hat{y}|x,k)\right] - \frac{1}{K}\sum_{k=1}^{K}\mathcal{H}\left[p(\hat{y}|x,k)\right], \tag{6}$$

which is the entropy of the expected output distribution minus the average entropy of the output distributions. MI is high for a sample x if the predictions are *individually confident but also in disagreement with each other*. This tells us how much additional information the diversity brings to the overall model: if all the predictions are equally uncertain about the same samples they disagree on, then aggregating them will not affect the aggregated uncertainty estimate.

In Table 4 (left) we report the average MI on StreetHazards and BDD-Anomaly validation, comparing again MOoSe and ensembles. Similarly to variance, our method's predictions have higher MI than both ensemble types, indicating that contextual probing not only produces more diversity in absolute terms, but also that this diversity adds more information to the model's predictive distribution.

Finally, in Table 4 (left) we report the Expected Calibration Error [24] of all methods, to show that even if ensembles are better calibrated than the baseline, it is MOoSe that performs the best at uncertainty estimation overall.

Table 4: **Left - variance and Mutual Information (MI)** show higher diversity for MOoSe than for ensembles, while lower ECE shows that our approach yields better calibrated predictions. All metrics are computed on DeepLabV3-ResNet50. **Right - single-dilation:** OoD detection results (AUPR) for single dilation models (SD) with different dilation rates (1, 12, 24, 36), compared to standard multi-dilation MOoSe. First row shows results for the global head, second row adds the probes, bottom two rows show the absolute and relative improvement

	StreetHazards			BDD-Anomaly		
Method	Var.↑	MI↑	ECE↓	Var.↑	MI↑	ECE↓
Global	-	-	.038	-	-	.123
MH-Ens	0.20	.004	.039	0.30	.022	.104
DeepEns	0.54	.012	.032	1.19	.054	.103
MOoSe	1.05	.034	.031	1.34	.062	.093

	SD1	SD12	SD24	SD36	MOoSe
Var.	0.41	1.05	0.93	0.83	1.34
Global	8.2	8.1	9.1	8.6	10.2
+probes	10.1	10.1	10.6	10.5	13.1
Chng.	1.9	2.0	1.5	1.9	2.9
Chng. %	22.7	24.9	16.4	22.3	27.9

5.2 Context as a Source of Diversity

In the previous section we showed that our approach produces highly diverse predictions. In this section we investigate the source of this diversity: our hypothesis is that each head relies differently on contextual information depending on the dilation rate of their respective spatial pyramid module, resulting in diverse predictive behaviors.

We test this hypothesis by evaluating the ability of each head to perform semantic segmentation when *only* contextual information is available. We corrupt the pixels of the foreground classes in BDD-Anomaly[2] with random uniform noise while leaving the background pixels unchanged, then we evaluate how well each head can still classify the corrupted foreground pixels by relying on the context. An example of the process can be seen in Figure 5 (left). Figure 5 (right) shows the mIoU on the noisy foreground as a percentage of the foreground mIoU on the original clean image. We can observe that dilation rate and robustness to foreground corruption are proportional to each other at multiple noise levels, as further illustrated by the qualitative example in the figure. The different result quality for different dilation rates confirm the validity of contextual aggregation as a source of prediction diversity, as anticipated by the comparison with regular (non-contextual) ensembles on variance and mutual information in Section 5.1.

5.3 Effect of Contextual Diversity on OoD Detection

The results presented in Section 4 already show that contextual probing improves performance on the task. Moreover, results obtained from the application of MOoSe to transformer-based models (7.3% average AUPR increase on StreetHazards across scoring functions), which are available in the supplementary material, indicate that the principle is applicable across architectures and its gains are not an artifact of CNNs.

[2] Pole, traffic light, traffic sign, person, car, truck, bus.

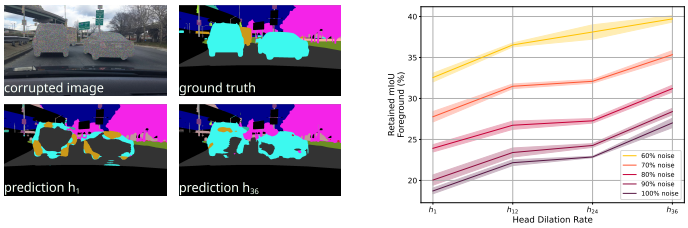


Fig. 5: **Left - example of foreground corruption:** the cars are corrupted with noise and the head with the largest dilation rate (h_{36}) can still largely segment them, unlike the no-dilation head h_1. **Right - corruption robustness:** We evaluate semantic segmentation of each probe on the corrupted foreground objects. The retained mIoU ($\text{mIoU}_{corrupt}/\text{mIoU}_{clean}$) increases with dilation rate, indicating more reliance on context. Results for BDD-Anomaly on DeepLabV3.

The last point to address is the contribution of contextual diversity to out-of-distribution detection. To quantify this contribution, we performed an ablation study removing receptive field diversity from DeepLabV3 by using the same dilation rate for all the convolutions in the spatial pyramid module. We train several versions of this single-dilation (SD) MOoSe, each with a different dilation rate, and present the comparison with standard MOoSe in Table 4 (right). Firstly, all single dilation models have lower prediction variance than regular MOoSe. Secondly, although the single-dilation models still outperform their global head, MOoSe yields larger gains than all SD models, both in absolute and relative terms. While these results confirm that contextual diversity is crucial for the success of our method, they also show that there are more contributing factors, compatibly with the known benefits of ensembles.

6 Conclusion

In this work we proposed a simple and effective approach for improving dense out-of-distribution detection by leveraging the properties of segmentation decoders to obtain a set of diverse predictions. Our experiments showed that MOoSe yields consistent gains on a variety of datasets and model architectures, and that it compares favorably with computationally much more expensive ensembles. We showed that our approach also outperforms other state-of-the-art approaches, and that due to its simplicity it could be easily combined with them. Even though we tested our method on various architectures, and despite the versatility of the main idea, one current limitation of MOoSe is its reliance on a specific architectural paradigm: the spatial pyramid. We also identified false positives among small objects to be an inherent failure mode of our approach, which potentially could be mitigated by combining MOoSe with alternative concepts that act at a single contextual scale.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. arXiv preprint arXiv:2011.06225 (2020)
2. Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In: International MICCAI Brainlesion Workshop. pp. 161–169. Springer (2018)
3. Bergman, L., Hoshen, Y.: Classification-based anomaly detection for general data. In: International Conference on Learning Representations (2019)
4. Bevandić, P., Krešo, I., Oršić, M., Šegvić, S.: Simultaneous semantic segmentation and outlier detection in presence of domain shift. In: German Conference on Pattern Recognition. pp. 33–47. Springer (2019)
5. Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: The fishyscapes benchmark: Measuring blind spots in semantic segmentation. arXiv preprint arXiv:1904.03215 (2019)
6. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International Conference on Machine Learning. pp. 1613–1622. PMLR (2015)
7. Cen, J., Yun, P., Cai, J., Wang, M.Y., Liu, M.: Deep metric learning for open world semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15333–15342 (2021)
8. Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Fua, P., Salzmann, M., Rottmann, M.: Segmentmeifyoucan: A benchmark for anomaly segmentation (2021)
9. Chan, R., Rottmann, M., Gottschalk, H.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. CoRR **abs/2012.06575** (2020), `https://arxiv.org/abs/2012.06575`
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
11. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
12. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
13. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: 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. pp. 12475–12485 (2020)
14. Cohen, N., Hoshen, Y.: Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357 (2020)
15. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
16. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)

17. Di Biase, G., Blum, H., Siegwart, R., Cadena, C.: Pixel-wise anomaly detection in complex driving scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16918–16927 (2021)
18. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
19. Franchi, G., Bursuc, A., Aldea, E., Dubuisson, S., Bloch, I.: One versus all for deep neural network incertitude (OVNNI) quantification. CoRR **abs/2006.00954** (2020), `https://arxiv.org/abs/2006.00954`
20. Franchi, G., Bursuc, A., Aldea, E., Dubuisson, S., Bloch, I.: Tradi: Tracking deep neural network weight distributions. In: European Conference on Computer Vision (ECCV) 2020. Springer (2020)
21. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
22. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: NeurIPS (2018)
23. Grcić, M., Bevandić, P., Šegvić, S.: Dense open-set recognition with synthetic outliers generated by real nvp. arXiv preprint arXiv:2011.11094 (2020)
24. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
25. Haselmann, M., Gruber, D.P., Tabatabai, P.: Anomaly detection using deep learning based image completion. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1237–1242. IEEE (2018)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
27. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41–50 (2019)
28. Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132 (2019)
29. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proceedings of International Conference on Learning Representations (2017)
30. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: International Conference on Learning Representations (2018)
31. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. Advances in Neural Information Processing Systems (NeurIPS) (2019)
32. Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., Brox, T.: Uncertainty estimates and multi-hypotheses networks for optical flow. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 652–667 (2018)
33. Jung, S., Lee, J., Gwak, D., Choi, S., Choo, J.: Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15425–15434 (October 2021)

34. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: British Machine Vision Conference 2017, BMVC 2017 (2017)
35. Kirichenko, P., Izmailov, P., Wilson, A.G.: Why normalizing flows fail to detect out-of-distribution data. arXiv preprint arXiv:2006.08545 (2020)
36. Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. arXiv preprint arXiv:2104.02939 (2021)
37. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017)
38. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
39. Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., Batra, D.: Why m heads are better than one: Training a diverse ensemble of deep networks. arXiv preprint arXiv:1511.06314 (2015)
40. Li, H., Ng, J.Y.H., Natsev, P.: Ensemblenet: End-to-end optimization of multi-headed models. arXiv preprint arXiv:1905.09979 (2019)
41. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018)
42. Lis, K.M., Nakka, K.K., Fua, P., Salzmann, M.: Detecting the unexpected via image resynthesis. International Conference On Computer Vision (ICCV) pp. 2152–2161 (2019). https://doi.org/10.1109/ICCV.2019.00224, `http://infoscience.epfl.ch/record/269093`
43. Liu, L., Wei, W., Chow, K.H., Loper, M., Gursoy, E., Truex, S., Wu, Y.: Deep neural network ensembles against deception: Ensemble diversity, accuracy and robustness. In: 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS). pp. 274–282. IEEE (2019)
44. Malinin, A., Mlodozeniec, B., Gales, M.: Ensemble distribution distillation. arXiv preprint arXiv:1905.00076 (2019)
45. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? arXiv preprint arXiv:1810.09136 (2018)
46. Narayanan, A.R., Zela, A., Saikia, T., Brox, T., Hutter, F.: Multi-headed neural ensemble search. In: Workshop on Uncertainty and Robustness in Deep Learning (UDL@ICML'21) (2021)
47. Nguyen, D.T., Lou, Z., Klar, M., Brox, T.: Anomaly detection with multiple-hypotheses predictions. In: International Conference on Machine Learning. pp. 4800–4809. PMLR (2019)
48. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)
49. Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., Mester, R.: Lost and found: detecting small road hazards for self-driving vehicles. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1099–1106. IEEE (2016)
50. Rudolph, M., Wandt, B., Rosenhahn, B.: Same same but differnet: Semi-supervised defect detection with normalizing flows. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1907–1916 (2021)

51. Schirrmeister, R., Zhou, Y., Ball, T., Zhang, D.: Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. Advances in Neural Information Processing Systems **33** (2020)
52. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging. pp. 146–157. Springer (2017)
53. Smith, L., Gal, Y.: Understanding measures of uncertainty for adversarial example detection. arXiv preprint arXiv:1803.08533 (2018)
54. Vojir, T., Sipka, T., Aljundi, R., Chumerin, N., Reino, D.O., Matas, J.: Road anomaly detection by partial image reconstruction with segmentation coupling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15651–15660 (2021)
55. Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., Willke, T.L.: Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
56. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al.: Contrastive training for improved out-of-distribution detection. arXiv preprint arXiv:2007.05566 (2020)
57. Xia, Y., Zhang, Y., Liu, F., Shen, W., Yuille, A.L.: Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In: European Conference on Computer Vision. pp. 145–161. Springer (2020)
58. Yan, H., Zhang, C., Wu, M.: Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. CoRR **abs/2201.01615** (2022), `https://arxiv.org/abs/2201.01615`
59. Yoo, D., Park, S., Lee, J.Y., So Kweon, I.: Multi-scale pyramid pooling for deep convolutional representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 71–80 (2015)
60. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 **2**(5), 6 (2018)
61. Zaidi, S., Zela, A., Elsken, T., Holmes, C., Hutter, F., Teh, Y.W.: Neural ensemble search for performant and calibrated predictions. Workshop on Uncertainty and Robustness in Deep Learning (UDL@ICML'20) (2020)
62. Zhang, H., Li, A., Guo, J., Guo, Y.: Hybrid models for open set recognition. In: European Conference on Computer Vision. pp. 102–117. Springer (2020)
63. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)