Probing Contextual Diversity for Dense Out-of-Distribution Detection – Supplementary Material

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

University of Freiburg

A Application of MOoSe to Transformers

Table 1: **Lawin transformer + MOoSe.** Results of the application of MOoSe to the Lawin transformer on the CAOS benchmark. Compatibly with the results obtained using CNN-based models, the the approach yields improvements on all metrics, for all scoring functions and on both datasets.

Score fn.	Method	StreetHazards AUPR↑	StreetHazards FPR@ 95TPR↓	BDD-Anomaly AUPR↑	BDD-Anomaly FPR@ 95TPR↓
MSP	Global	14.78	21.52	7.46	20.00
	MOoSe	**16.05**	**20.96**	**8.12**	**19.85**
H	Global	18.22	20.81	10.94	17.84
	MOoSe	**19.90**	**20.17**	**11.89**	**17.53**
ML	Global	21.07	18.78	8.17	24.26
	MOoSe	**21.87**	**17.99**	**11.40**	**15.56**

In this section we report the results of our method applied to an attention-based model for semantic segmentation, the Lawin transformer [13]. The Lawin transformer features a spatial pyramid pooling based on multiple attention window sizes instead of dilated convolutions. We apply MOoSe to Lawin as we did for DeepLabV3, by replicating the model's decoder head for the 4 contextual probes, which are each built on top of a single module of the spatial pyramid and all have access to the features from the image-level pooling.

The results for out-of-distribution detection with Lawin on the CAOS benchmark are shown in Table 1. Analogously to the results presented in the main paper, it can be seen that the gains provided by MOoSe affect all metrics on both datasets and for all scoring functions.

Compared to DeepLabV3, Lawin performs substantially better on StreetHazards but worse on BDD-Anomaly. We argue that this could be an effect of the smaller size of anomalous objects in the second dataset, which does not pair well

with the input patching performed by transformers. Regardless, the results show that the advantages of contextual probing are valid for modern attention-based neural networks and are not confined to convolutional neural networks.

B Comparison with Ensembles – ResNet101 Backbone

Shown in Table 2 are the results for MOoSe, ensembles and the single-head baseline on the CAOS benchmark. Unlike the corresponding table in the main paper, all numbers reported in Table 2 are for DeepLabV3 [5] models, which differ in the backbone used: ResNet50 or ResNet101 [7]. On StreetHazards the results are rather consistent over different backbones, on all metrics and especially using entropy and maximum logit scoring functions. On BDD-Anomaly the more powerful backbone has a slight but consistent advantage, possibly due to the higher difficulty of this dataset.

Table 2: **CAOS benchmark, ResNet101 backbone.** Comparison between global head (Global), multi-head ensembles (MH-Ens), standard deep ensembles (DeepEns) and MOoSe on dense out-of-distribution detection. Results are shown for DeepLabV3 with a ResNet50 or ResNet101 backbone. All three scoring functions (maximum softmax probablity (MSP), entropy (H), maximum logit (ML)) are considered. Best results are shown in **bold**, all results are percentages

		StreetHazards				BDD-Anomaly			
		ResNet50		ResNet101		ResNet50		ResNet101	
Score fn.	Method	AUPR	FPR@ 95TPR	AUPR	FPR@ 95TPR	AUPR	FPR@ 95TPR	AUPR	FPR@ 95TPR
MSP	Global	9.11	22.37	9.25	21.70	7.01	22.47	6.87	22.19
	MH-Ens	9.69	21.40	10.39	**19.72**	7.55	25.50	8.49	25.66
	DeepEns	10.22	21.09	10.65	19.77	7.64	**21.53**	8.37	**20.38**
	MOoSe(ours)	**12.53**	**21.05**	**12.86**	22.81	**8.66**	22.49	**8.96**	22.36
H	Global	11.89	22.07	11.92	21.42	10.23	20.64	10.35	20.55
	MH-Ens	12.59	21.10	13.26	**19.09**	10.62	23.51	12.36	23.54
	DeepEns	13.43	20.62	13.91	19.32	11.39	19.31	12.67	**18.05**
	MOoSe(ours)	**15.43**	**19.89**	**15.28**	21.69	**12.59**	**19.27**	**13.25**	19.17
ML	Global	13.57	23.27	13.52	23.64	10.69	15.60	11.89	15.58
	MH-Ens	13.99	21.86	13.80	**17.70**	10.69	20.19	12.23	20.91
	DeepEns	14.57	21.79	14.69	20.71	11.40	14.66	12.87	**13.34**
	MOoSe(ours)	**15.22**	**17.55**	**15.26**	18.62	**12.52**	**13.86**	**13.48**	13.50

C Outlier Exposure

In Table3 (left) we report the results on RoadAnomaly for MOoSe trained using the Outlier Exposure / Entropy training scheme from [4]. Outlier Exposure boosts the results of MOoSe for all three scoring functions, in terms of both AUPR and FPR_{95}. When using Outlier Exposure, the entropy scoring function turns out to outperform max-logit, likely because the negative training loss explicitly encourages entropy to be high for negative pixels.

Table 3: **Left:** results on the RoadAnomaly benchmark, showing the performance of MOoSe trained using Outlier Exposure / Entropy Training. **Right:** results on the SegmentMeIfYouCan benchmark (anomaly track) validation and test.

Score fn.	Method	RoadAnomaly AUPR	RoadAnomaly FPR_{95}
MSP	Global	23.76	51.32
	MOoSe	31.53	43.41
H	Global	32.00	49.14
	MOoSe	41.48	36.78
ML	Global	37.86	39.03
	MOoSe	**43.59**	**32.12**
MSP	MOoSe+OE	44.95	30.18
H	MOoSe+OE	<u>55.86</u>	<u>23.59</u>
ML	MOoSe+OE	53.19	24.38
	DML [2]	37	37
	Std.ML [9]	25.82	49.74

	Method	SegmentMeIfYouCan Pixel AUPR↑	Pixel FPR_{95}↓	Pixel F_1^*↑	Component $\overline{\text{sIoU}}$↑	Component $\overline{F}_1$↑
validation	Global MSP	54.8	38.2	52.7	27.0	14.1
	MOoSe MSP	60.4	35.0	58.0	41.9	22.4
	Global H	62.3	37.4	57.1	32.5	13.2
	MOoSe H	65.7	32.5	65.7	49.8	18.9
	Global ML	67.0	36.4	62.5	35.2	10.4
	MOoSe ML	65.6	33.2	67.0	45.6	18.5
test	MOoSe-H	51.7	44.0	55.0	29.7	10.7
	Resynth.[40]	52.3	25.9	60.5	39.5	12.9
	ObsNet	75.4	26.7	-	44.2	45.1

D SegmentMeIfYouCan Benchmark

Table3 (right) contains the results for the SegmentMeIfYouCan [3] (anomaly track) validation and test benchmarks, the latter having undisclosed ground truth. The benchmark is composed of 100 test images of road scenes containing anomalous objects of various nature, similar to RoadAnomaly in nature. For this benchmark we use the same model used for RoadAnomaly - trained on the BDD100K dataset as explained in the main paper. The explanation of the individual metrics (pixel and component-wise) can be found in the original paper [3].

The results confirm that MOoSe provides the expected gains compared to the base model, while also showing that the entropy scoring function performs as well as max-logit. The Image Resynthesis [11] method outperforms MOoSe on the test benchmark, but as reported in the main paper it has worse results on LostAndFound. The best comparable results on the benchmark are obtained by ObsNet [1], which leverages local adversarial attacks and an external observer network to obtain uncertainty scores. Evaluations of pure Out-of-Distribution detection on other benchmarks are missing.

E Qualitative Examples

In Figure 1 we provide qualitative comparisons between our method and deep ensembles, showing the differences between the various segmentation predictions of the two approaches. Figure 1 resumes the example from StreetHazards shown in Figure 1 of the main paper, in which a large outlier is confidently classified by the global head as belonging to the "car" category. The predictions of the deep ensemble are very similar with little disagreement, essentially failing to detect the outlier.

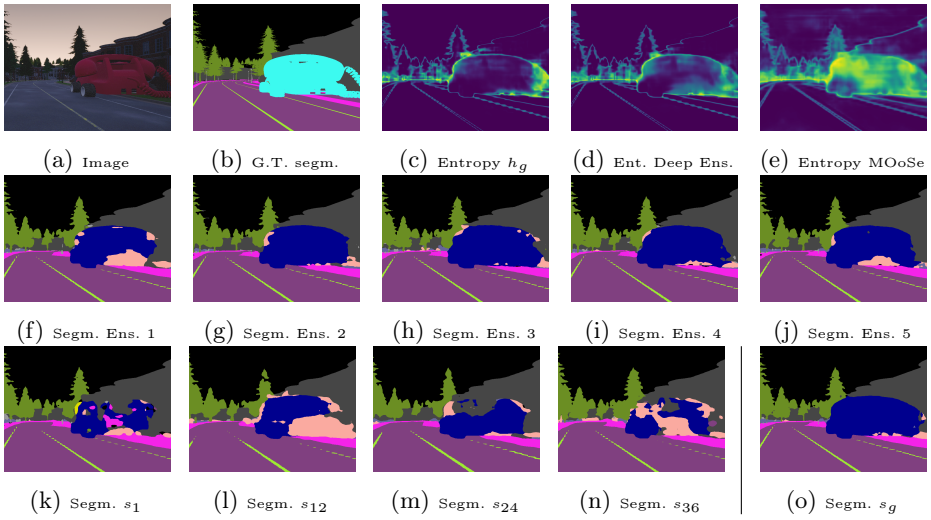


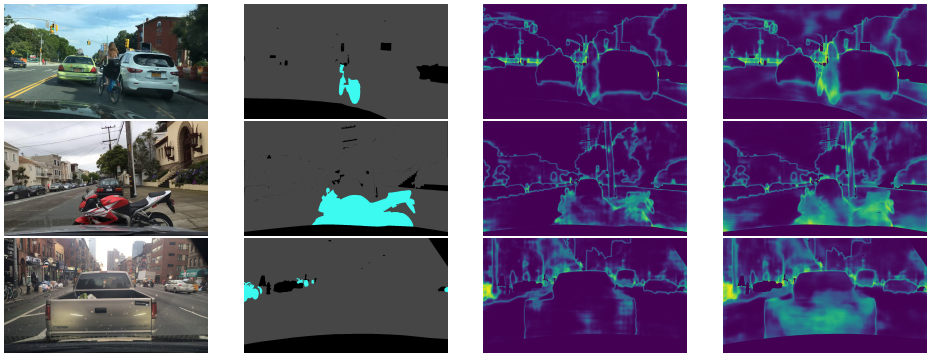
Fig. 1: **Qualitative comparison – MOoSe and deep ensembles.** Test sample from the StreetHazards dataset: a street scene containing an anomalous object (indicated in cyan in the ground truth segmentation (b)). The global head is confident about its prediction, as shown by its entropy heatmap (c). The second row (f-j) shows the segmentation predictions of different members of a deep ensemble model. These mostly agree on the category of the anomalous object and do not improve the entropy heatmap substantially over the single model (d). The contextual predictions of MOoSe are shown in the third row, and manifest clearly higher diversity than ensembles. This results in a different entropy scoremap (e), in which most of the pixels of the anomalous object are highlighted.

In Figure 2 we show the qualitative results for three samples from the BDD-Anomaly dataset, comparing MOoSe and the global head baseline. On the first two samples MOoSe performs better than the baseline for large anomalous objects that are in front of the camera, and which are mostly missed by the global head's entropy scoremap.

Moreover, in all examples it can be seen again that MOoSe tends to be uncertain on particular inlier regions which are otherwise confidently classified by the global head. In the third example in particular, the hatch of the pick-up truck in front of the camera is cause for uncertainty for MOoSe. Although the car category is present in the training set, the rarity of the model and possibly the reflection on the metal are likely sources of doubt for the neural network.

F MOoSe Performance – Standard Deviation

Table 4 shows the average OoD detection performance on the CAOS benchmark [8] for MOoSe (same as Table 1 in the main paper), with the addition of the standard deviation over the 3 runs.



(a) Image (b) OoD ground truth segm. (c) Entropy h_g (d) Entropy MOoSe

Fig. 2: **Qualitative examples – BDD-Anomaly.** Comparison of the OoD scores of MOoSewith those of the global head on three examples from the BDD-Anomaly dataset. In the first example the global head assigns low entropy to the outlier – a bicycle in the center of the roadway. MOoSedoes better in this case, although it is also uncertain about the person riding the bike. In the second example a motorbike (category excluded from the training set and therefore anomalous) is placed right in front of the observer. Neither method produces an entropy scoremap that can clearly identify the object, but the baseline misses a larger number of anomalous pixels. In the example in the third row it can be seen how MOoSewrongly produces high entropy in correspondence of the rear hatch of the pick up truck in front of the camera. Despite this, the model produces higher uncertainty for the true outliers.

Table 4: **Standard deviation.** Results for MOoSeon the CAOS benchmark, using the DeepLabV3 and PSPNet architectures – both with ResNet50 backbones. The average results are the same as in Table 1, but include standard deviations over 3 experiments with different random seeds

	StreetHazards				BDD-Anomaly			
	DeepLabV3		PSPNet		DeepLabV3		PSPNet	
Score fn.	AUPR↑	FPR@ 95TPR↓	AUPR↑	FPR@ 95TPR↓	AUPR↑	FPR@ 95TPR↓	AUPR↑	FPR@ 95TPR↓
MSP	12.53±0.21	21.05±0.25	11.28±0.14	21.94±0.31	8.66±0.25	22.49±0.93	8.11±0.40	24.09±1.35
H	15.43±0.35	19.89±0.25	14.52±0.24	21.20±0.32	12.59±0.42	19.27±1.01	12.35±0.54	20.98±1.63
ML	15.22±0.26	17.55±0.22	15.29±0.14	20.46±0.52	12.52±0.36	13.86±0.82	12.88±0.58	13.94±1.30

G Training Details

G.1 Training of the Base Segmentation Model

The base models for the CAOS benchmark are all trained using a mini-batch of 8 samples, randomly cropped to 512 pixels and flipped horizontally with probability $P = 0.5$. For the DeepLabV3 models we found the best results to be obtained using Stochastic Gradient Descent [10] optimization with a learning rate of 0.05 decreased to 0 over 200 epochs following a cosine schedule and a weight decay factor of $1e^{-4}$. For Lawin we used the AdamW optimizer (weight decay 0.01) with base learning rates $1e^{-6}$ and $1e^{-5}$ for StreetHazards and BDD-Anomaly respectively, and the learning rate of the decoder being 10 times bigger than the learning rate of the backbone.

For the experiments on LostAndFound [12] we used a DeepLabV3+ [6] model pretrained on Cityscapes. The initialization parameters for the model can be found at:
`https://github.com/NVIDIA/semantic-segmentation`.

The models evaluated on RoadAnomaly [11] were trained on the BDD100k [14] dataset using the same training scheme as BDD-Anomaly.

G.2 Training MOoSe

As explained in the main paper, the contextual probes of MOoSe are trained without affecting the main model. We achieve this by only updating the parameters of the probes: this includes the parameters learned by backpropagation/gradient descent, like convolutional filters, and those which are updated following batch statistics.

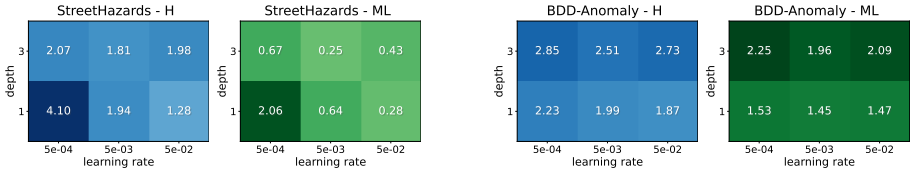
There are 2 hyperparameters introduced by our method: the learning rate for training the probes and their depth factor. In this section, we evaluate the performance of MOoSe for different hyperparameter configurations. The remaining training details, such as batch size and optimizer choice, are unchanged with respect to the training of the main model.

Figure 3 displays the performance improvements yielded by our method over its global head alone for StreetHazards and BDD-Anomaly respectively.

We consider two depth options, $d = \{1, 3\}$, and three learning rate options: $lr = \{5e^{-4}, 5e^{-3}, 5e^{-2}\}$.

As expected, the optimal configurations differ for the two datasets. Once again as a likely result of the lower difficulty of the dataset, the best configuration for StreetHazards is the one with the shallowest heads and the lowest learning rate. However, all configurations produce improvements over the global head, especially when using the entropy scoring function. On BDD-Anomaly the optimal results are obtained with deeper heads and the smallest learning rate, although the performance gains are more consistent across all configurations.

Additionally, the model trained on Cityscapes uses learning rate (probes only) 0.05 and depth 3.



(a) (b)

Fig. 3: **Hyperparameters – (a) StreetHazards.** AUPR improvement of MOoSe compared to the global head, at varying head depths and learning rates. Results for StreetHazards. In blue are the results for the entropy scoring function, in green for maximum logit. **(b) BDD-Anomaly.** AUPR improvement of MOoSe compared to the global head, at varying head depths and learning rates. Results for BDD-Anomaly. In blue are the results for the entropy scoring function, in green for maximum logit.

H Ablation Studies

H.1 Probe Contribution

In Table 5 we report the effect of using only selected subsets of heads for OoD segmentation with MOoSe. While the ranking of the configurations indicates that the number of heads correlates positively with OoD detection performance, we can also observe that different contextual heads contribute in different amounts to the result improvement. In particular it can be seen that h_1 is present in all the 4 best performing configurations, and performs better than h_{12}, h_{24} and h_{36} individually in terms of AUPR. Furthermore, the configurations containing h_1 and none or few more heads have higher FPR95%TPR than the others (underlined in the table), indicating that this head is responsible for increasing the overall recall, and with it the false positives.

Table 5: **Head Contributions.** OoD detection performance of MOoSewhen using selected contextual heads. Results are shown for StreetHazards on a single DeepLabV3-ResNet50 network using entropy scores, the rows within each category are sorted by AUPR. The results indicate that in terms of OoD detection h_1 is the most important individual head It is notable how the contribution of h_1 is more important than that of the other contextual heads in terms of AUPR

h_1	h_{12}	h_{24}	h_{36}	h_g	AUPR↑	FPR@↓ 95%TPR
✓	✓	✓	✓	✓	15.04	18.00
✓	✓			✓	15.05	19.33
✓	✓	✓		✓	15.02	18.24
✓			✓	✓	14.55	22.38
	✓	✓	✓	✓	14.28	17.63
✓				✓	14.17	<u>24.88</u>
		✓	✓	✓	13.84	19.35
	✓			✓	13.62	19.08
			✓	✓	13.52	21.66
		✓		✓	13.26	19.55
				✓	12.23	22.41
✓	✓	✓	✓		14.87	18.08
✓					13.78	<u>27.03</u>
			✓		12.53	22.91
	✓				11.33	21.92
		✓			11.16	21.73

Table 6: OoD detection performance (AUPR) per number of heads (different dilation rates evenly spaced between 1 and 36, same random seed). AUPR increases with the variety of dilation rates but starts saturating with 4 heads (with the current hyperparameters)

	StreetHazards			BDD-Anomaly		
N. probes:	2	4	6	2	4	6
MSP	10.47	12.85	13.31	7.22	8.99	9.01
H	12.73	15.99	15.85	10.63	13.08	12.96
ML	13.49	15.63	15.90	11.25	12.94	12.74

H.2 Number of Probes

In this section we present the results of an ablation study on the effect of the number of spatial pyramid modules – and consequently number of probes – on MOoSe. We train versions of DeepLabV3 featuring 2, 4, or 6 dilated convolutions, on top of which we train contextual probes, following the exact same setup as the experiments in the main paper. The dilation rates are chose to span uniformly the range between 1 and 36, which is the standard range of the spatial pyramid of DeepLabV3.

Results of the ablation are shown in Table 6, where we report the AUPR for each configuration and scoring function - for both datasets of the CAOS benchmark. We observe that the models featuring 2 probes are the ones performing the worst on both datasets and with all scoring functions. When increasing the number of heads the results are mixed and generally closer.

We can conclude that, taking into account consistency with the existing models and computational efficiency, the best practice is to stick with the default number of heads.

References

1. Victor Besnier, Andrei Bursuc, David Picard, and Briot Alexandre. Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.
2. Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15333–15342, 2021.
3. Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation, 2021.
4. Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. *CoRR*, abs/2012.06575, 2020.
5. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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.
6. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.
7. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
8. Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.
9. Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision (ICCV), pages 15425–15434, October 2021.
10. Jack Kiefer, Jacob Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.
11. Krzysztof Maciej Lis, Krishna Kanth Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. *International Conference On Computer Vision (ICCV)*, pages 2152–2161, 2019.
12. 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)*, pages 1099–1106. IEEE, 2016.
13. Haotian Yan, Chuang Zhang, and Ming Wu. Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. *CoRR*, abs/2201.01615, 2022.
14. Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.