SAFE: Sensitivity-Aware Features for Out-of-Distribution Object Detection

<Supplementary Material>

Samuel Wilson[1], Tobias Fischer[1], Feras Dayoub[2], Dimity Miller[1], Niko Sünderhauf[1]
[1]QUT Centre for Robotics, Queensland University of Technology
[2]Australian Institute for Machine Learning, University of Adelaide
s84.wilson@hdr.qut.edu.au

A. Transformer-based Object Detectors

In Section 4 of our manuscript, we provide a comprehensive comparison of our SAFE OOD detector to the current state-of-the-art in OOD object detectors. Specifically, our results in Section 4.3 and Table 1 demonstrated that SAFE achieves state-of-the-art performance across both the ResNet-50 and RegNetX4.0 backbone architectures. In these comparisons, the object detector architecture, Faster-RCNN, remains static.

Notably, our theory in Section 3.1 of the manuscript only requires the presence of our SAFE layers in the *backbone* of the object detector network. Thus, we further test the generalisability of our SAFE OOD object detector by evaluating the performance when applied to a different, transformer-based, object detector. Specifically, we compare SAFE to the benchmark proposed in [3] which performs OOD object detection on the Deformable DETR [19] object detector.

A.1 Experimental Setup

Following [3], we use the same evaluation protocol defined in [4] for our comparisons of SAFE to state-of-the-art OOD object detectors on the transformer-based Deformable DETR [19] object detector.

Datasets Consistent with Section 4.1 of our manuscript, we use PASCAL-VOC [5] and BDD100K [18] as our ID datasets with MS-COCO [11] and OpenImages [9] as our OOD datasets.

Evaluation Metrics Following the standard evaluation protocol [4, 3], we use the same AUROC and FPR95 metrics as defined in Section 4.1 of our manuscript.

Baselines We compare against the following state-of-the-art methods: Mahalanobis Distance [10], Gram Matrices [14], KNN [15], CSI [16], OW-DETR [7], Dismax [12], VOS [4] and SIREN [3]. SIREN can be evaluated with either the original von-Mises Fischer distribution (SIREN-vMF) or in combination with KNN (SIREN-KNN). Performance metrics of baselines are all reported from [3].

Base Network Architecture We implement the Deformable DETR [19] object detector with a ResNet-50 [8] backbone pre-trained on ImageNet [13]. Of the compared methods, CSI [16], OW-DETR [7], Dismax [12], VOS [4] and SIREN [3] all require the object detector to be retrained following a custom loss objective, we identify these methods with a checkmark ✓ in Supplementary Table 1.

Implementation As the Deformable DETR object detector uses a ResNet-based backbone, we follow the same implementation as in Section 4.2 of our manuscript. Specifically, we instantiate a 3-layer auxiliary MLP which takes the object-specific vectors from the SAFE layers as input and outputs a single OOD score for each object. The auxiliary MLP is trained on the *surrogate* task of discriminating clean ID samples from ID samples perturbed with the FGSM [6] adversarial-perturbation. We set the scalar magnitude multiplier for FGSM to be $\epsilon = 8$, the same optimum value found for the ResNet-50 backbone in Section 4.5 and Figure 4 of our manuscript.

A.2 Results and Discussion

Supplementary Table 1 extends the quantitative evaluations from our manuscript, comparing our SAFE OOD detector to the current state-of-the-art in OOD object detection with the transformer-based Deformable DETR object detector [19].

Congruent with the results on the Faster-RCNN model, SAFE achieves state-of-the-art performance, outperforming the previous state-of-the-art SIREN-KNN [3] in 7 out of the 8 total benchmark permutations. In particular, SAFE boasts substantial performance improvements in the PASCAL-VOC setting, with *absolute* reductions in FPR95 of 57.00% for OpenImages and 15.89% for MS-COCO. Furthermore, all methods other than SIREN [3] and Dismax [12] achieve near-

Method	Retrain?	ID: PASCAL-VOC				ID: Berkeley DeepDrive-100K			
		OpenImages		MS-COCO		OpenImages		MS-COCO	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Mahalanobis [10]		49.08	97.88	50.28	97.39	77.98	71.43	76.83	70.86
Gram [14]		38.81	95.29	43.97	94.16	57.14	71.56	60.13	73.81
KNN [15]		59.64	91.36	62.15	91.80	79.64	61.13	80.90	64.75
CSI [16]	✓	51.37	79.16	55.07	84.00	76.42	71.30	77.93	70.27
OW-DETR [7]	✓	57.80	93.82	55.70	93.09	73.78	77.37	70.29	80.78
Dismax [12]	✓	70.66	76.37	75.21	82.05	67.18	81.23	72.14	77.62
VOS [4]	✓	52.77	97.07	54.40	97.46	76.62	72.58	77.33	76.44
SIREN-vMF [3]	✓	71.05±0.1	78.36±1.0	76.10±0.1	75.49±0.8	79.77±1.2	66.31±0.9	80.06±0.5	67.54±1.3
SIREN-KNN [3]	✓	74.93±0.1	65.99±0.5	78.23±0.2	64.77±0.2	89.00±0.4	47.28±0.3	**86.56**±0.1	53.97±0.7
SAFE (ours)		**96.73**±0.7	**8.99**±1.4	**78.88**±1.0	**48.88**±1.5	**94.31**±0.7	**21.10**±2.0	85.95±1.1	**39.18**±2.0

Supplementary Table 1: OOD detection results comparing SAFE to state-of-the-art OOD detectors on the transformer-based Deformable DETR [19] object detector with a ResNet-50 backbone. Comparison metrics are FPR95 and AUROC, directional arrows indicate if higher (↑) or lower (↓) values indicate better performance. **Best** results are shown in **red and bold**, second best results are shown in orange. Methods that require retraining are indicated with a checkmark ✓. Mean and standard deviation over 5 seeds is shown for SAFE in the format of $\mu \pm \sigma$. SAFE provides consistently strong performance

random performance when PASCAL-VOC is the ID dataset – resulting in SAFE reducing the FPR95 of *post-hoc* detectors by 82.37% from 91.36%→8.99% (KNN [15]→SAFE) when OpenImages is the OOD set. Similarly, when BDD100K is the ID set we observe strong reductions in FPR95 with 26.18% for OpenImages and 14.79% in MS-COCO. Notably, the *only* metric in which SAFE does not report the best performance (AUROC for BDD100K→MS-COCO) is less than 1% from achieving state-of-the-art performance.

In summary, SAFE, which does not require retraining, outperforms OOD detectors *that do require retraining*, and significantly outperforms other *post-hoc* OOD detectors across varying backbone and object detector architectures.

B. Alternative Transforms

In Sections 3 & 4 of the manuscript, we posit that the SAFE critical layers, composed of a residual convolution layer followed by batch normalisation, are disproportionately powerful for OOD detection. Specifically, Section 4.4 demonstrated superior sensitivity of the SAFE critical layers when a MLP is trained via a surrogate training task. In the manuscript, the surrogate training task we proposed was adversarial perturbation detection.

Here, our aim is to show that the SAFE critical layers can also detect less targeted input variations, as per our theory from Section 3.1.We thus evaluate the performance of SAFE when the surrogate task is to distinguish clean ID samples from those perturbed by a simple *additive noise transformation*. In particular, we use additive noise sampled from a uniform distribution, $x^+ \sim U(\alpha_w, \alpha_w)$ with magnitude multiplier α_w.

Supplementary Figure 1 visualises the input perturbation signal and resultant effect on an input sample for FGSM [6] (Top) and Noise (Bottom) at varied perturbation magnitudes. For both transforms, we observe that the optimal value found for SAFE ($\epsilon = 8$ for FGSM [6] and $\alpha_w = 30$ for Noise) result in perturbed samples that are imperceptibly different to the original image, whilst large magnitudes that produce perceptible differences result in degraded performance.

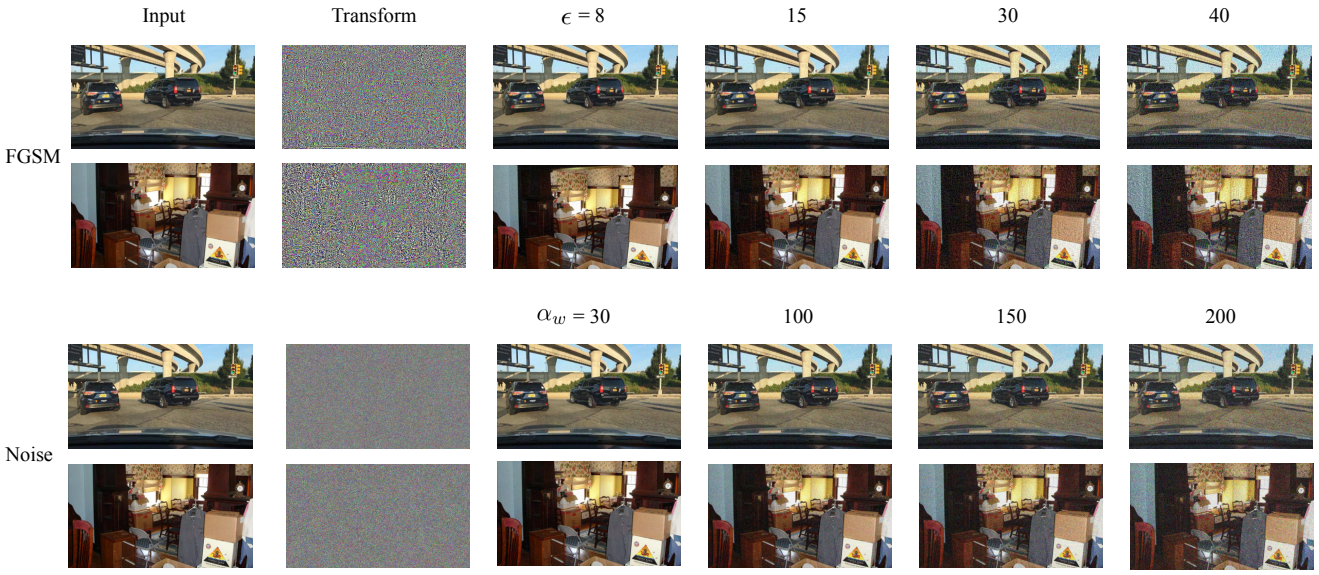
B.1 Robustness to input variations

Supplementary Table 2 compares the performance when using an additive noise transformation (Noise) or an adversarial attack as in the manuscript (FGSM [6]) against the current state-of-the-art, VOS [4]. Consistent with the theory described in Section 3, even in the case where the auxiliary MLP is trained with simple additive noise transform, the SAFE critical layers are *sensitive* enough to produce a strong signal for OOD object detection. While the Noise transformation leads to an approximately 2% AUROC reduction when compared to our proposed adversarial attack transform, it still outperforms the previous state-of-the-art across the majority of benchmarks.

Supplementary Figure 2 provides the same magnitude ablation for the Noise perturbation as Figure 2 did in the manuscript for FGSM [6]. As expected, the existing trend of increasing performance to a peak followed by a decrease in performance for FGSM holds for the Noise transformation.

B.2 Universally high performance of SAFE layers

As discussed in Section 4.3 of our manuscript, prior works [1, 2, 14, 17] in OOD classification have highlighted the variability of layer performance under distributional shift. In this section, we demonstrate that the SAFE layers are high performing irrespective of the input transformation that is



Supplementary Figure 1: Visualisation of the input perturbation functions and impact sample ID images at varying magnitudes. **Top**: Perturbation function is FGSM [6]. **Bottom**: Perturbation function is Noise. Numbers above each image correspond to the magnitude multiplier applied to the image, ϵ for FGSM and α_w for Noise.

PASCAL-VOC				
Method	**OpenImages**		**COCO**	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓
VOS	85.23±0.6	51.33±1.6	**88.70**±1.2	47.53±2.9
SAFE (Noise)	90.22±0.5	24.36±0.9	79.86±0.1	50.00±0.6
SAFE (FGSM)	**92.28**±1.0	**20.06**±2.3	80.30±2.4	**47.40**±3.8
BDD100K				
Method	**OpenImages**		**COCO**	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓
VOS	88.52±1.3	35.54±1.7	86.87±2.1	44.27±2.0
SAFE (Noise)	94.10±0.2	19.58±0.5	88.36±0.4	35.60±0.8
SAFE (FGSM)	**94.64**±0.3	**16.04**±0.5	**88.96**±0.6	**32.56**±0.8

Supplementary Table 2: Comparison of the performance of SAFE utilising either an additive noise transformation (Noise) or adversarial perturbation function (FGSM) against the previous state-of-the-art (VOS) with the ResNet-50 backbone. Comparison metrics are AUROC and FPR95. Directional arrows indicate if higher ↑ or lower ↓ indicates better performance. **Best** results are in **red and bold**, second best results are displayed in orange. SAFE sets a new state-of-the-art in performance with either transformation.

used. This is important as it enables layers selection *a priori* of the OOD data or ID perturbation.

Supplementary Figure 3 provides the same layer-wise performance ablation as in the manuscript's Figure 3 for the additive noise (Noise) transformation. We observe that even under this distributional shift, the SAFE critical layers are among the highest performing layers with little variation in their overall performance. Interestingly, while there are high-performing non-SAFE layers, many of these layers experience a performance drop (most notably FPR95) from the (non-SAFE) high-performing layers when FGSM [6] is used.

References

[1] Vahdat Abdelzad, Krzysztof Czarnecki, Rick Salay, Taylor Denounden, Sachin Vernekar, and Buu Phan. Detecting out-of-distribution inputs in deep neural networks using an early-layer output. *arXiv preprint arXiv:1910.10307*, 2019. 2

[2] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and H.T. Kung. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19217–19227, 2022. 2

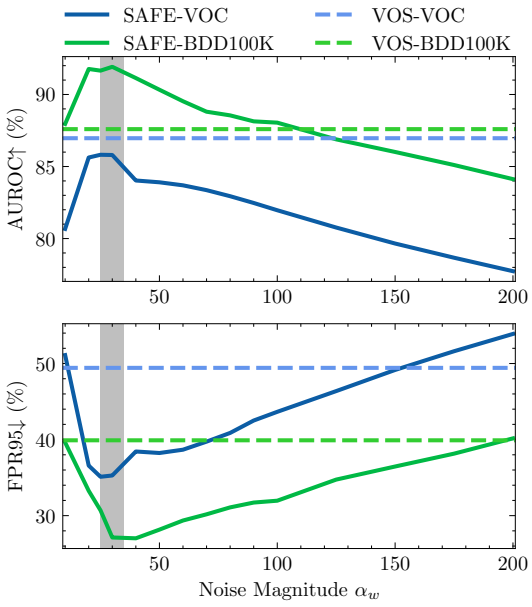
[3] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Advances in Neural Information Processing Systems*, 2022. 1, 2

[4] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 1, 2

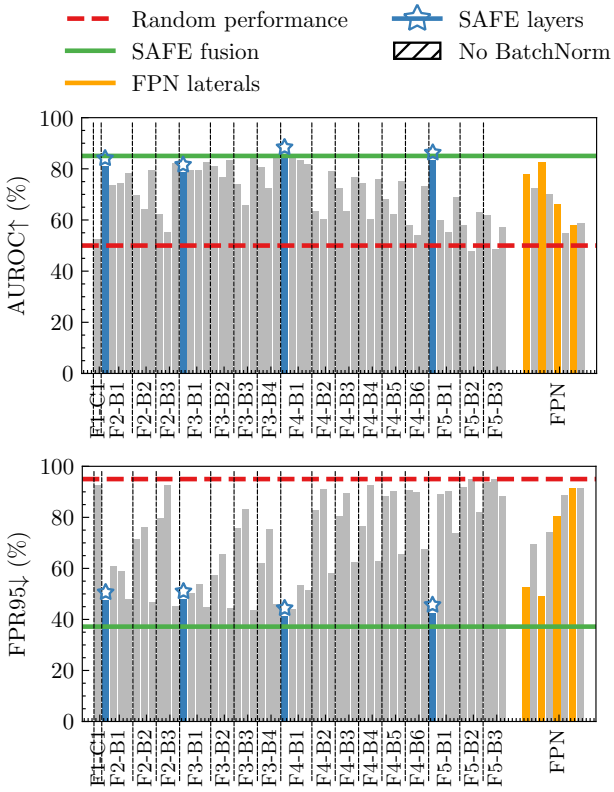
[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1

[6] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 1, 2, 3

[7] Akshita Gupta, Sanath Narayan, K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Con-*



Supplementary Figure 2: OOD detection performance of SAFE with the ResNet-50 backbone as the additive noise transform magnitude α_w is varied. **Top:** Comparison metric is AUROC, higher is better. **Bottom:** Comparison metric is FPR95, lower is better. Individual lines correspond to the average performance over both OOD sets for the given ID set. Dashed lines correspond to the performance of VOS for the respective datasets. A region of consistent high performance for all ID and OOD permutations exists between $\alpha_w \in [25, 35]$ (grey region), the existence of this region for both FGSM and additive noise indicates that SAFE is robust to distributional shifts in the MLP surrogate training.



Supplementary Figure 3: OOD detection performance of individual Conv2d layers in the standard ResNet-50 backbone when PASCAL-VOC is the ID set and the surrogate MLP training task uses an *additive noise transformation*. **Top**: Comparison metric is AUROC, higher is better. **Bottom**: Comparison metric is FPR95, lower is better. Results are reported as averages over both OOD datasets. Layers in blue with a star are the identified critical layers for SAFE. Striped layers belong to the Feature Pyramid Network (FPN) and are the only Conv2d layers that *do not have BatchNorm applied immediately after*. The SAFE layers are consistently among the highest performing even under the effects of a distributional shift in the synthetic outliers.

ference on Computer Vision and Pattern Recognition (CVPR), pages 9235–9244, June 2022. 1, 2

[8] 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 (CVPR)*, pages 770–778, 2016. 1

[9] Ivan Krasin et al. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 1

[10] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeuRIPS)*, page 7167–7177, 2018. 1, 2

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 1

[12] David Macêdo, Cleber Zanchettin, and Teresa Bernarda Ludermir. Distinction maximization loss: Efficiently improving out-of-distribution detection and uncertainty estimation simply replacing the loss and calibrating. *CoRR*, abs/2205.05874, 2022. 1, 2

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[14] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In *Proceedings of the International Conference on Machine Learning (ICML)*,

pages 8491–8501, 2020. 1, 2

[15] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022. 1, 2

[16] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2020. 1, 2

[17] Samuel Wilson, Tobias Fischer, Niko Sünderhauf, and Feras Dayoub. Hyperdimensional feature fusion for out-of-distribution detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2

[18] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020. 1

[19] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 1, 2