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

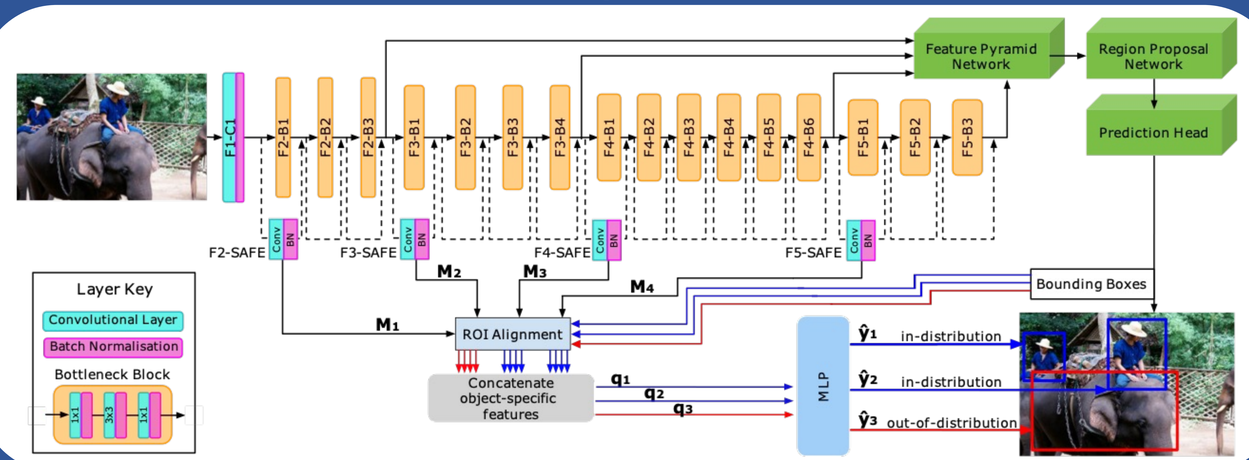
Samuel Wilson[1], Tobias Fischer[1], Feras Dayoub[2], Dimity Miller[1], Niko Sünderhauf[1]

[1]QUT Centre For Robotics, [2]Australian Institute for Machine Learning

1) Task: Out-of-Distribution Object Detection

Out-**o**f-**D**istribution (**OOD**) object detection challenges deep networks to alert when test-time object detections do not belong to the training distribution.

2) Overview: Sensitivity-Aware Features (SAFE)



3) Identifying Sensitive Features

We identify two characteristics that signify a layer is *sensitive* to OOD samples:

1) Residual connections ensure that input distances are preserved in the hidden space, making the network *sensitive* to input changes:

$$L_1 \cdot \|x - x^*\|_I \leq \|f(x) - f(x^*)\|_F$$

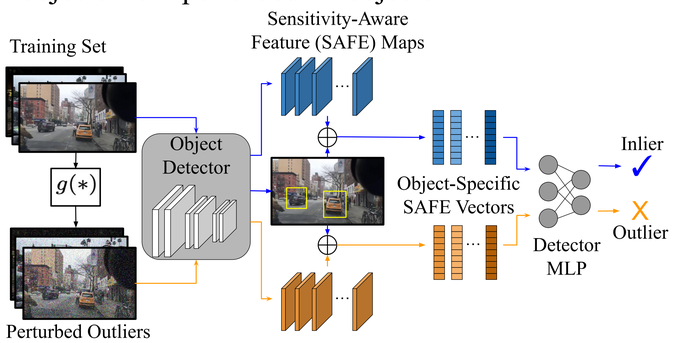
2) Batch Normalisation layers trigger abnormal activations when presented with OOD data:

$$\text{BatchNorm}(z;\gamma,\beta,\epsilon) = \frac{z - \mathbb{E}_{in}[z]}{\sqrt{\mathbb{V}_{in}[z] + \epsilon}} \cdot \gamma + \beta$$

Residual connections followed immediately by **batch normalisation** (SAFE layers) exhibit both characteristics, thus making them *sensitive* to OOD data. Such connections already exist in pretrained ResNet-like architectures.

4) Surrogate Training

A **Multi**Layer Perceptron (**MLP**) is trained to distinguish SAFE features of clean in-distribution (**ID**) objects from perturbed ID objects:



At test-time, the **surrogate**-trained MLP flags OOD objects by assigning them a higher score than known in-distribution objects:

$$\hat{y}_d = f_\beta(\boldsymbol{q}_d) \quad \hat{y} \in [0, 1]$$

5) Experimental Results

SAFE more than *halves* the false positive rate (FPR95) of previous state-of-the-art methods.

Method	ID: PASCAL-VOC				ID: Berkley DeepDrive-100K			
	OpenImages		MS-COCO		OpenImages		MS-COCO	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP [2]	81.91	73.13	83.45	70.99	77.38	79.04	75.87	80.94
ODIN [5]	82.59	63.14	82.20	59.82	76.61	58.92	74.44	62.85
Energy [6]	82.98	58.69	83.69	56.89	79.60	54.97	77.48	60.06
KNN [7]	85.08	55.73	86.07	54.50	88.37	44.50	87.45	47.28
G-ODIN [3]	79.23	70.28	83.12	59.57	87.18	50.17	85.22	57.27
CSI [8]	82.95	57.41	81.83	59.91	87.99	37.06	84.09	47.10
GAN-Syn [4]	82.67	59.97	83.67	60.93	81.25	50.61	78.82	57.03
VOS-RN50 [1]	85.23	51.33	88.70	47.53	88.52	35.54	86.87	44.27
VOS-RX4.0 [1]	87.59	48.33	**89.00**	47.77	92.13	27.24	89.08	36.61
SAFE-RN50	92.28	20.06	80.30	47.40	94.64	16.04	88.96	32.56
SAFE-RX4.0	**94.38**	**17.69**	87.03	**36.32**	**95.97**	**13.98**	**93.91**	**21.69**

SAFE layers are consistently among the most powerful layers across distributional shifts.