FEVER-OOD: Free Energy Vulnerability Elimination for Robust Out-of-Distribution Detection

Brian K.S. Isaac-Medina*,†, Mauricio Che‡, Yona Falinie A. Gaus*, Samet Akçay§ and Toby P. Breckon*
*Durham University, †OpenWorks Engineering, ‡University of Vienna, §Intel

ICCV HONOLULU HAWAII OCT 19-23, 2025

Motivation: Free-Energy Score

The Free Energy score is an effective measure of uncertainty for OOD detection[1,2,3,4]. For a classifier, it is:

$$F(x) = -\log \sum_{k=1}^{K} \exp\left(\mathrm{W}_{cls_k}^{\top} \cdot h(x)\right)$$

x: Input sample.
K: Total in-distribution categories.
$h(x) \in \mathbb{R}^{d'}$: Feature vector of x.
$\mathrm{W}_{cls} \in \mathbb{R}^{d' \times K}$: Last linear layer of the classifier
$\mathrm{W}_{cls_k}^{\top} \in \mathbb{R}^{d'}$: k-th column of W_{cls}.

Null Space Vulnerabilities (NSV)

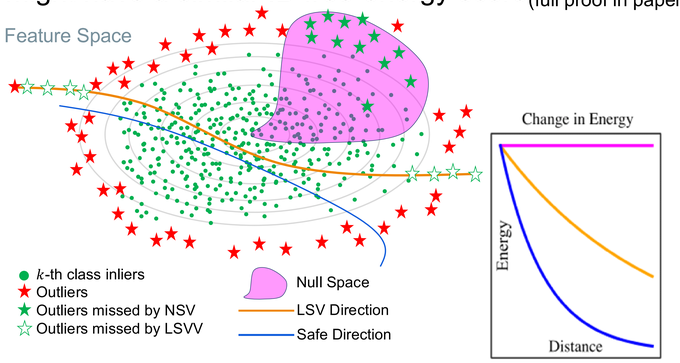
If $\mathbf{v} = h(x)$, then it can be decomposed in

$$\mathbf{v} = \mathrm{Null}(\mathrm{W}_{cls})\mathbf{v}_0 + \mathrm{Null}(\mathrm{W}_{cls})^{\perp}\mathbf{v}_{\perp}$$

where $\mathbf{v}_0$ lies in the null space of W_{cls} and $\mathbf{v}_{\perp}$ is orthogonal to $\mathbf{v}_0$. Therefore, the free energy is dominated by $\mathbf{v}_{\perp}$ which may lie in the ID space.

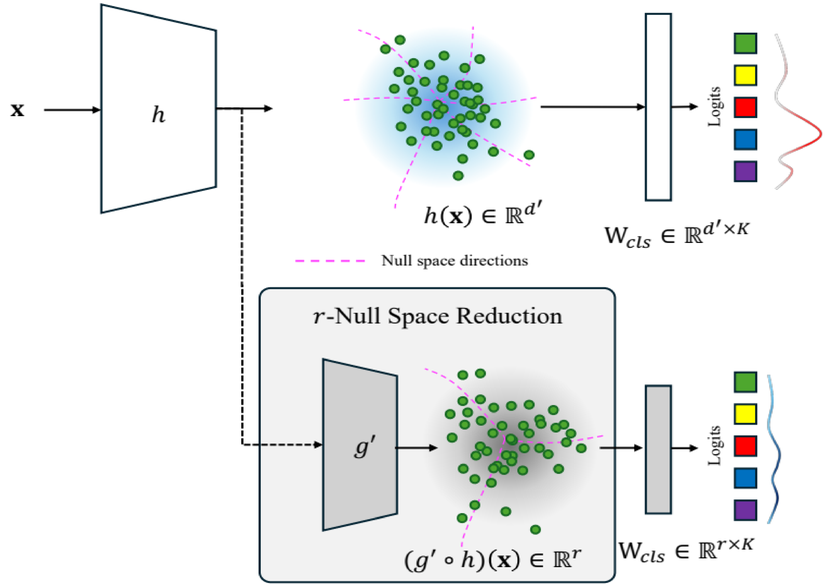
Least Singular Value Vulnerabilities (LSVV)

Similarly, if a significant part of $\mathbf{v}_{\perp}$ lies in the direction of the least singular vector of W_{cls}, then it might have a similar ID free-energy score(full proof in paper).



FEVER-OOD: Mitigating the Vulnerabilities

Null space reduction (NSR): adding an extra linear-layer $g' : \mathbb{R}^{d'} \to \mathbb{R}^{r}$, with $r < d'$ such that the nullity of W_{cls} decreases.



Least Singular Value Regularisation: we propose two regularisation loses to increase the least singular value σ_{min}:

Least Singular Value Regulariser (LSVR)

$$\mathcal{L}_{LSV} = \sigma_{min}^{-1}(\mathrm{W}_{cls})$$

Conditioning Number Regulariser (CNR)

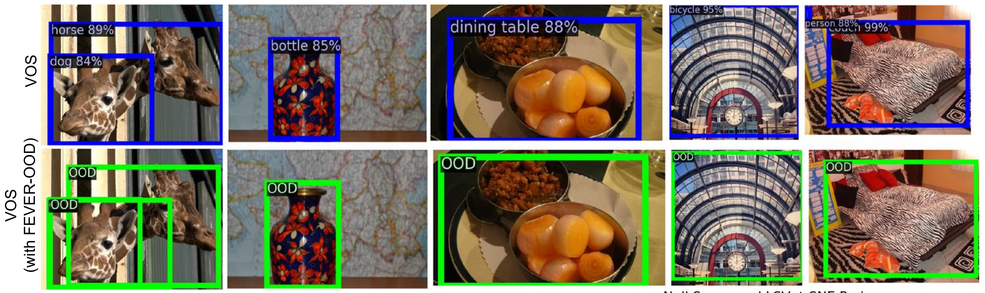
$$\mathcal{L}_{CN} = \sigma_{max}(\mathrm{W}_{cls})/\sigma_{min}(\mathrm{W}_{cls})$$

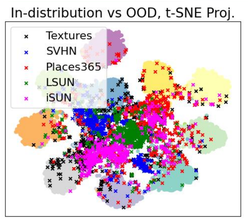
Results: FEVER-OOD achieved SOTA in OOD detection for both tasks

Dataset	Method	FEVER-OOD		Average	
		NSR	Reg	FPR95	AUROC
CIFAR-10	VOS	-	-	33.66	92.15
		96	1.0	**28.22**	**94.78**
	FFS	-	-	34.82	90.52
		96	1.0	**31.04**	**94.03**
CIFAR-100	VOS	-	-	71.01	81.02
		114	0.01	**68.57**	**82.83**
	FFS	-	-	71.89	80.89
		114	0.001	**66.14**	**83.47**
	Dream -OOD	-	-	50.96	88.06
		100	-	**42.77**	**89.98**
Image Net-100	Dream -OOD	-	-	39.33	91.85
		100	0.01	**36.50**	**92.74**

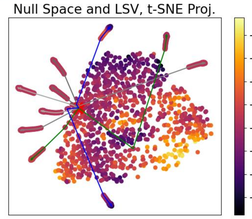
Method	FEVER-OOD		Average			
			OOD Dataset			
			MS-COCO		OpenImages	
	NSR	Reg	FPR95	AUROC	FPR95	AUROC
VOS	-	-	50.29	87.77	53.09	86.58
	768	-	**47.47**	88.15	**49.36**	86.32
	256	-	47.88	**88.49**	52.41	86.39
	-	-	50.77	87.18	53.78	85.34
FFS	512	-	**46.93**	**88.82**	53.89	86.28
	256	-	51.18	88.74	53.67	86.56
	256	0.01	47.93	88.04	**47.93**	**84.95**

↑ Results with PASCAL VOC AS ID dataset & MS-COCO, OpenImages as OOD dataset
← Results with CIFAR-10, CIFAR-100, ImageNet-100 as ID dataset & Textures, SVHN, Places365, LSUN, iSUN as OOD dataset

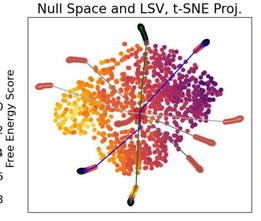




Feature Space t-SNE Projection for models trained on CIFAR-10.



Free Energy visualization of the feature space



Free Energy visualization of the feature space with FEVER-OOD

Conclusions:

Vulnerabilities: NSV (overlapping ID–OOD scores) and LSVV (energy similarity).
Approach: Extra layer with CNR, compact features with LSVR and CNR regularizers for clearer score separation and stable energy.
Results: 10.13% ↓ FPR, +1.6% ↑ AUROC on Dream-OOD (ImageNet-100), new SOTA.
Generalization: Works across varied detection architectures.

[1] Liu et al., Energy-based Out-of-distribution Detection, NeurIPS 2020
[2] Du et al., VOS: Learning What You Don't Know by Virtual Outlier Synthesis, ICML 2022
[3] Kumar et al., FFS: Normalizing Flow based Feature Synthesis for Outlier-Aware Object Detection, CVPR 2023
[4] Du et al., Dream the Impossible (Drean): Outlier Imagination with Diffusion Models, NeurIPS 2023