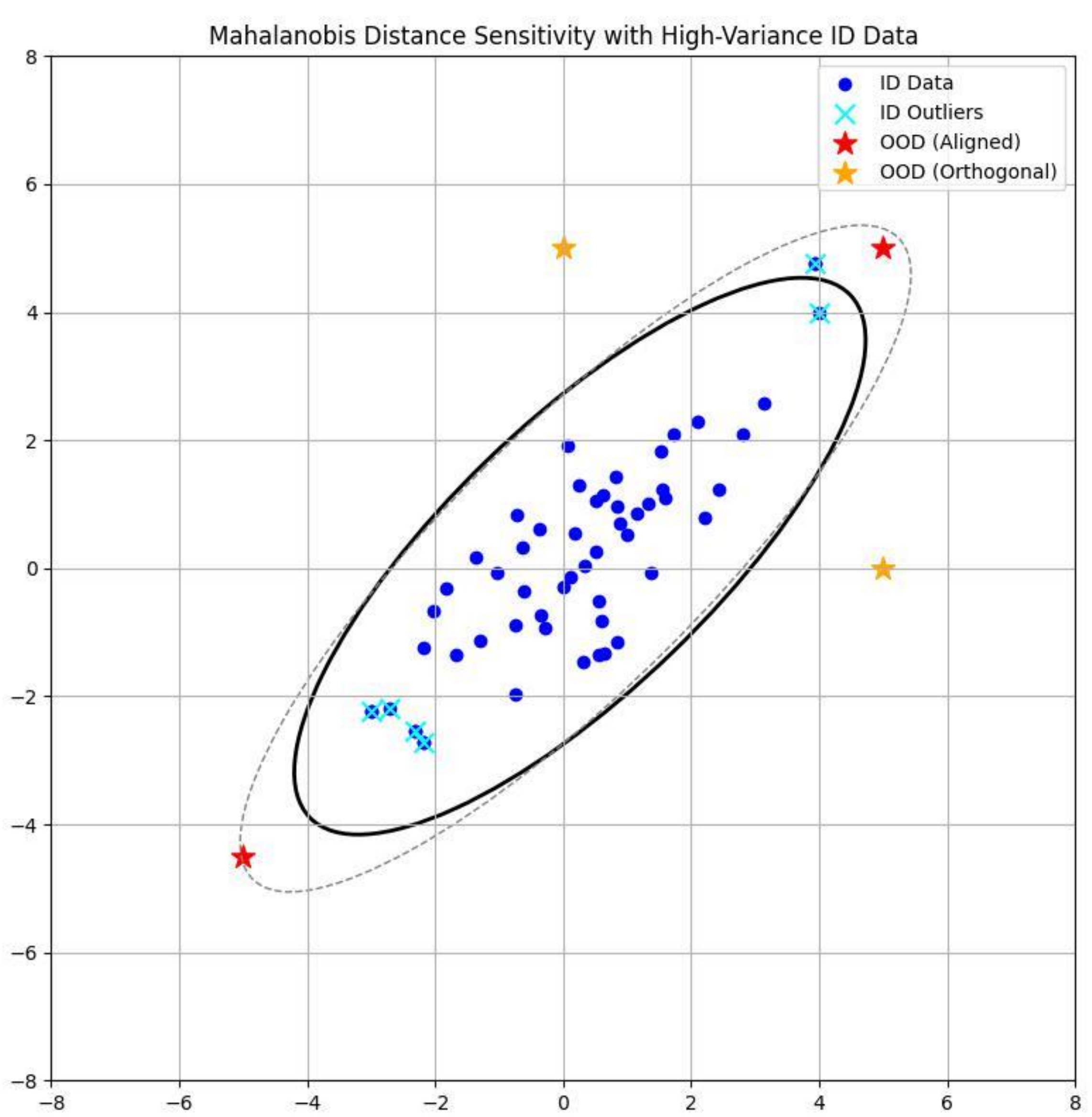
Improving Out-of-Distribution Detection via Dynamic Covariance Calibration

Kaiyu Guo[1,3], Zijian Wang[1], Tan Pan[2,3], Brian C. Lovell[1], Mahsa Baktashmotlagh[1]

[1]University of Queensland [2]Fudan University [3]Shanghai Academy of AI for Science

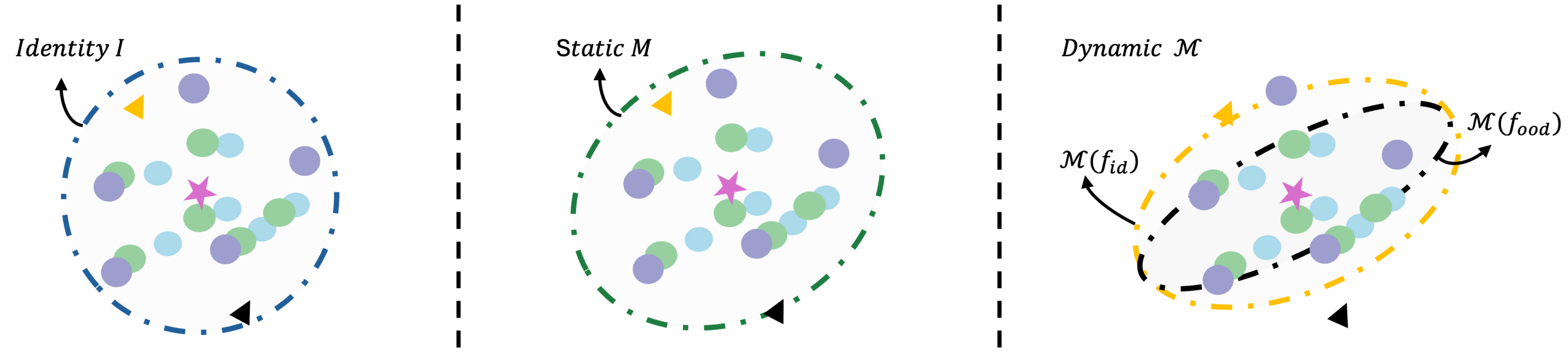
Motivation



Outlier points may affect the calculation of distance for OOD detection.

How can we reduce the effect of outlier features on the information geometry without losing important characteristics of the ID data?

Insight and Method



$$d_I(f, f_a) = \sqrt{(f - f_a)^T I (f - f_a)}$$

Euclidean distance

(1)

$$d_M(f, f_a) = \sqrt{(f - f_a)^T M (f - f_a)}$$

Mahalanobis distance

(2)

$$d_{\mathcal{M}}(f, f_a) = \sqrt{(f - f_a)^T \mathcal{M}(f) (f - f_a)}$$

Dynamic distance

(3)

Training data features · *Real time ID feature f_{id}* · *Real time OOD feature f_{ood}* · *Cluster center f_a (Anchor)*

$$\mathcal{M}(f) = \left(Cov - f_p^\top f_p\right)^{-1}, \quad \text{where } f_p \text{ is the realtime feature projected to the ID residual space}$$

Algorithm 1 OOD Score $s(\cdot)$ Calculation

Input: Feature vector f, basis matrix $\mathbf{B}$, within-class covariance matrix Σ_R, class means $\{\mu_i\}$

Output: Score function $s(\cdot)$

Normalize f using the normalizer

$\mathbf{a} \leftarrow$ torch.einsum$('i, bi \rightarrow b', f, \mathbf{B})$

$adj \leftarrow \mathbf{B}^\top \mathbf{a}^\top \mathbf{a} \mathbf{B}$

$dygeo \leftarrow$ torch.linalg.inv$(\Sigma_R - adj)$

$d \leftarrow f - \{\mu_i\}$

$dists \leftarrow$ torch.einsum$('bi, ij, bj \rightarrow b', d, dygeo, d)$

$score \leftarrow -$torch.min$(\sqrt{dists})$

return $score$

When $(f - f_a)^\top \mathcal{M}(f)(f - f_a)$ is larger than zero

Theorem 4.2. *Given a feature $f \in \mathbf{R}^d$, a non-zero feature $a \in \mathbf{R}^d$, and a symmetric positive definite matrix $\Sigma \in \mathbf{R}^{d \times d}$, we define $p = f^\top \Sigma^{-1} f$, $q = a^\top \Sigma^{-1} a$, and $s = f^\top \Sigma^{-1} a$. Setting $d(f) = (f - a)^\top (\Sigma - f f^\top)^{-1} (f - a)$, if $p < 1$, then $d(f) \geq 0$; if $p > 1$ and $(s-1)^2 \leq (p-1)(q-1)$, then $d(f) \geq 0$.*

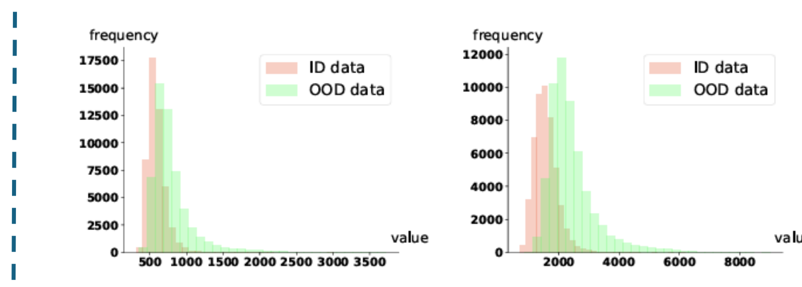
Experiments

Table 1. Comparision with different post-hoc OOD detection methods on **CIFAR** benchmarks. We present the AUROC and FPR95 results on DenseNet and WideResNet and the average results over 2 ID datasets. The results of CIFAR-10/CIFAR-100 are averaged over 6 OOD datasets. The detailed results can be viewed in the appendix.

Method	DenseNet						WideResNet					
	CIFAR-10		CIFAR-100		Avg.		CIFAR-10		CIFAR-100		Avg.	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP (Hendrycks & Gimpel, 2017)	92.5	48.72	74.37	80.13	83.44	64.43	91.07	50.64	76.93	77.48	84	64.06
Energy (Liu et al., 2020)	94.65	26.2	81.17	68.44	87.91	47.32	91.86	33.74	79.83	71.95	85.85	52.85
maxLogit (Basart et al., 2022)	94.64	26.36	81.06	68.53	87.85	47.45	91.84	33.61	79.92	72.37	85.88	52.99
ODIN (Liang et al., 2018)	94.65	26.35	81.06	68.53	87.86	47.44	91.85	33.62	79.93	72.38	85.89	53
Mahalanobis (Lee et al., 2018)	85.9	47.64	77.56	58.08	81.73	52.86	90.88	47.58	79.35	59.63	85.12	53.61
GEM (Morteza & Li, 2022)	88.01	31.73	84.19	56.93	86.1	44.33	93.22	37.28	82.71	57.15	87.97	47.22
KNN (Sun et al., 2022)	96.79	16.16	87.56	42.3	92.18	29.23	93.68	33.56	86.34	48.32	90.01	40.94
ReAct (Sun et al., 2021)	95.76	23.59	82.98	67.38	89.37	45.49	92.09	34.06	80.69	72.26	86.39	53.16
Line (Ahn et al., 2023)	96.99	14.75	88.76	35.11	92.88	24.93	78.94	61.6	66.33	83.45	72.64	72.53
DICE (Sun & Li, 2022)	95.01	21.44	86.55	51.66	90.78	36.55	90.48	34.44	78.44	71.04	84.46	52.74
FDBD (Liu & Qin, 2023)	**97.23**	**13.86**	89.25	50.57	93.24	32.22	92.27	36.87	85.14	65.77	88.71	51.32
ours	96.83	14.63	**92.38**	**29.98**	**94.61**	**22.31**	**96.18**	**18.57**	**89.08**	**44.89**	**92.63**	**31.73**

Table 2. Comparision with different post-hoc OOD detection methods on **ImageNet-1k** benchmark. We present the AUROC and FPR95 results on ViT, ResNet-50, SwinV2-B, and DeiT. We also provide the average results over the 4 pre-trained models. The results of the four pre-pretrained models are averaged over 6 OOD datasets. The detailed results can be viewed in the appendix. As we can not achieve solid results with WDiscOOD on ImageNet-1k pre-trained SwinV2-B/16, the average results are from the other 3 pre-trained models.

Method	Models									
	ViT		ResNet-50		Swin-B		DeiT		Avg.	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP (Hendrycks & Gimpel, 2017)	88.89	43.46	73.97	70.98	81.29	63.49	79.8	66.97	80.99	61.23
Energy (Liu et al., 2020)	94.11	27.56	79.53	65.8	80.07	60.35	71.65	72.65	81.34	56.59
ReAct (Sun et al., 2021)	94.07	27.69	83.34	54.81	85.2	53.53	77.16	68.74	84.94	51.19
ODIN (Liang et al., 2018)	93.73	29.68	79.42	65.77	80.68	58.94	76.07	66.43	82.47	55.2
maxLogit (Basart et al., 2022)	93.73	29.68	79.42	65.78	80.94	59.56	76.43	66.38	82.63	55.35
Mahalanobis (Lee et al., 2018)	94.27	27.11	68.36	80.63	87.86	52.07	83.98	73.86	83.62	58.42
KLMatch (Basart et al., 2022)	87.8	44.39	76.09	69.93	81.83	63.85	82.68	67.24	82.1	61.35
KNN (Sun et al., 2022)	92.6	34.38	84.43	57.46	85.08	65.24	82.75	76.24	86.22	58.33
VIM (Wang et al., 2022)	94.23	27.32	83.93	65.92	86.77	**51.33**	83.91	71.13	87.21	53.93
FDBD (Liu & Qin, 2023)	93.36	31.71	84.47	60.35	86.57	55.75	82.78	71.84	86.79	54.91
Neco (Ammar et al., 2024)	94.38	27.08	75.15	70.27	81.73	54.74	79.2	**62.03**	82.61	53.53
WDiscOOD (Chen et al., 2023)	**94.41**	**26.35**	70	78.71	-	-	83.97	73.83	82.8	59.63
ours	94.27	26.94	**87.43**	**51.77**	**88.1**	51.89	**84.97**	68.29	**88.69**	**49.72**

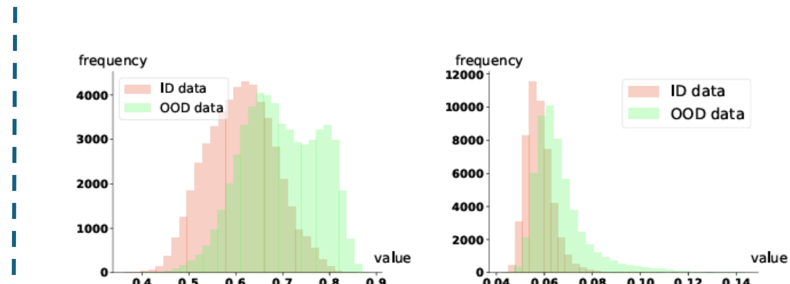


(a) p values on ResNet-50 (b) q values in ResNet-50

Figure 6. p and q values on ImageNet-1k pre-trained ResNet-50

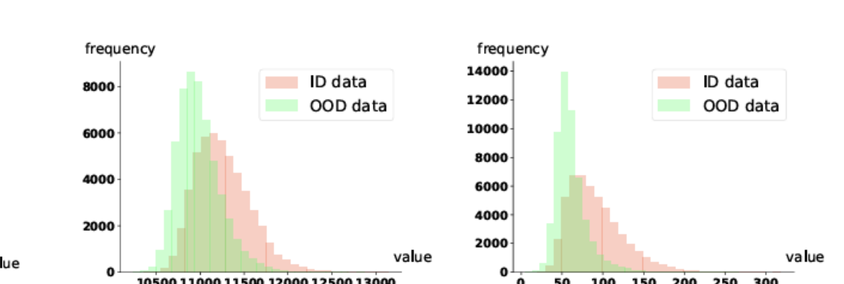
Table 3. **Comparison on DINO** Comparision with different post-hoc OOD detection methods on DINO. We report the averaged AUROC and FPR95 results over 6 OOD datasets. The detailed results can be viewed in the appendix.

Method	DINO	
	AUROC↑	FPR95↓
SSD (Sehwag et al., 2021)	51.5	96.86
Neco (Ammar et al., 2024)	46.97	96.69
KNN (Sun et al., 2022)	84.99	63.88
Mahalanobis (Lee et al., 2018)	91.57	39.27
ours	**91.65**	**38.23**



(a) ViT (b) ResNet-50

Figure 7. The l_2 norms of features extracted by ImageNet-1k pre-trained ResNet-50 and ViT.



(a) s values on ViT (b) s values on ResNet-50

Figure 8. The s values on ImageNet-1k pre-trained ResNet-50 and ViT.