GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection Supplementary Material

A. Experimental Setup

Datasets Specifications of the datasets used in our experiments are summarized in Table 1. ImageNet-1K represents ID data, and ImageNet-O, OpenImage-O, iNaturalist, and Textures are the OOD datasets. We also provide additional results for two datasets used in the earlier work of GradNorm [9] — SUN [22] and Places [23].

Input Images An input image to BiT [10] is resized to 480×480. For ViT [4], it is resized to 384×384. And the size of input images to the remaining four architectures RepVGG [3], Swin [15], DeiT [19], and ResNet-50-D [5] is resized to 224×224.

Dataset	Class / Image Distribution	# Images
ImageNet-1K (val) [17]	predefined (ID) class list	50,000
ImageNet-O [8]	natural adversarial images	2,000
OpenImage-O [21]	natural (OOD) class distribution	17,632
iNaturalist [20]	predefined (OOD) class list	10,000
Textures [2]	predefined (OOD) class list	5,160
SUN [22]	predefined (OOD) class list	10,000
Places [23]	predefined (OOD) class list	10,000

Table 1. *Specifications of ID/OOD datasets.*

ReAct [18] vs. ReAct* Here we clarify the difference between the original ReAct [18] and our local version, ReAct*. To use the consistent notation with the main paper, $\mathbf{z}$ denotes the feature from the penultimate layer, b and b^* denote the clipping threshold of ReAct [18] and ReAct*, respectively. N is the number of samples in the training dataset, and m is the dimensionality of the extracted feature. ReAct [18] is defined as following,

$$\texttt{ReAct}(\mathbf{z}; b) = \min(\mathbf{z}, b) \quad (1)$$
$$\text{s.t.} \quad \frac{\textbf{card}\left(\{i : \mathbf{z}_{\text{train}}(i) < b\}\right)}{mN} = q,$$

where $\mathbf{z}_{\text{train}} = \text{flatten}(\mathbf{Z}_{\text{train}})$ is the flattened array of features $\mathbf{Z}_{\text{train}} \in \mathbb{R}^{N \times m}$ extracted from the training data, $\textbf{card}(\cdot)$ is the cardinality and q is a pre-defined quantile, *e.g.*, $q = 0.99$. Intuitively, Eq. 1 indicates that ReAct [18]

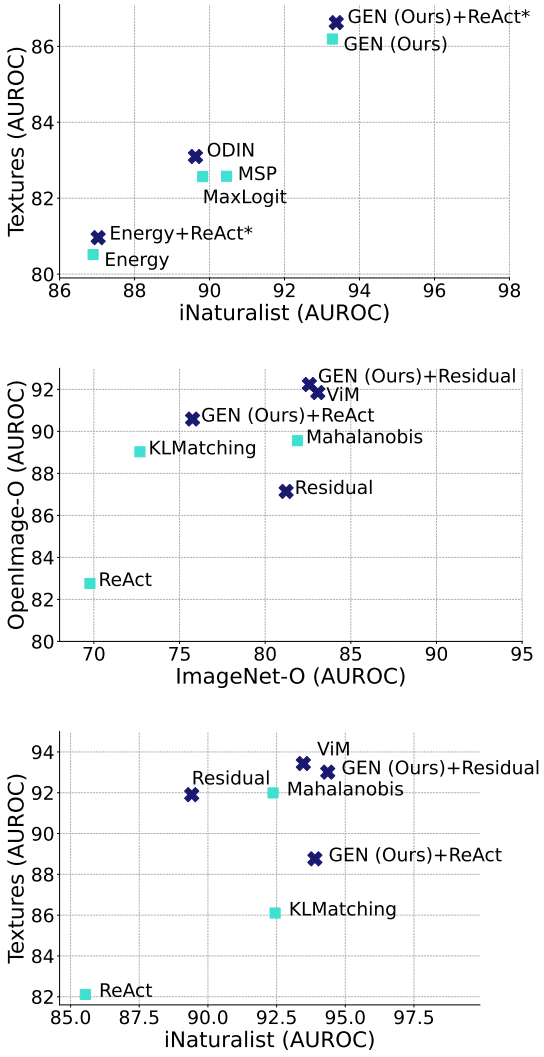

Figure 1. *Performance of (top) Post-hoc OOD Detection Methods and (mid—bottom) Methods Requiring ID Train Data Applied to 6 Classifiers Trained on ImageNet-1K.* Reported are AUROC values (%) averaged across the classifiers. Methods marked with light squares use information from logits / probabilities. Methods marked with dark crosses also use information from features.

employs all feature information extracted from the whole training data to find the optimal clipping threshold b. Instead, ReAct* chooses the clipping threshold based on the feature extracted from the current input only. ReAct* is defined as following,

$$\texttt{ReAct}^*(\mathbf{z}; b) = \min(\mathbf{z}, b^*) \quad (2)$$
$$\text{s.t.} \quad \frac{\mathbf{card}\{i : \mathbf{z}_i < b^*\}}{m} = q,$$

where $\mathbf{z} \in \mathbb{R}^m$ is an output of the penultimate layer applied to the input sample.

Combining Different Scores While ViM [21] suggests employing a scaled addition of the residual and the energy score (which requires estimation of a normalization parameter), we decide to multiply the residual with our post-hoc score, GEN. This avoids the need to estimate an additional scalar parameter and appears also beneficial given the normalization property of the geometric mean.

B. Averaged Performance Across Models

In Fig. 1, we report average AUROC across six classifiers for the remaining two datasets as well as the other non-post-hoc methods. One can see that GEN outperforms all the post-hoc methods on iNaturalist and Texture datasets as well as OpenImage-O and ImageNet-O shown in the main paper (see Fig. 1 of the main paper). One can also notice that GEN combined with Residual [21] is very competitive to ViM [21] on all the OOD datasets.

C. Detailed OOD Detection Performance Results

We provide an extended version of Tables 3 and 4 of the main paper reporting *Per-Dataset Performance* and *Average Performance* of OOD detection methods, respectively. Due to the page capacity limitation, we split the extended results into two tables. Table 3 shows the detailed OOD detection performance on each architecture and each OOD dataset for the post-hoc methods, and Table 4 — for the methods that require ID training data. In addition, the averaged performance across all six classifiers is reported in the bottom-most block of both tables — these results are graphically visualized in Fig. 1 of the main paper and Fig. 1 in this supplementary.

Recall that we rerun the experimental evaluation of OOD detection methods according to the protocol in ViM [21] with the exception of ODIN [13], and we obtained slightly better results than reported in ViM [21]. For ODIN [13], both the code and tuned hyperparameters (scale of the perturbation ε and temperature T) were not provided by ViM, therefore its results were taken from ViM [21] paper.

D. Extended Results for Effective Value of M and γ

This section contains a more detailed evaluation for our GEN score using varying choices for M and γ. In particular, we illustrate the results for the four remaining architectures RepVGG [3], ViT [4], DeiT [19], and ResNet-50-D [5]. The results for varying $M \in \{2, 10, 50, 100, 200, 500, 700, 800, 900, 1000\}$ are depicted in Fig. 2, where it can be seen that using more logit information causes OOD detection performance to degrade for most architectures except for ViT [4]. Besides, setting $M = 100$ seems perform well in terms of AUROC and FPR95 generally. The results of using different $\gamma = \{0.1, 0.3, 0.5, 0.9\}$ are shown in Fig. 3. The top row shows that using larger γ barely improves the performance in terms of AUROC. The same observation can be made regarding FPR95, which is shown in the bottom row.

E. Performance on Unseen Datasets

We perform OOD detection on two completely unseen OOD datasets from SUN [22] and Places [23]. Importantly, the overlapped classes between SUN / Places and ImageNet-1K are removed as provided by [9]. We use the previously validated hyperparameters $M = 100$ and $\gamma = 0.1$. The results can be found in Table 2 indicating a consistently better performance of GEN.

F. Using the Top Logits for the Energy Score

We empirically verify the hypothesis that using only the partial information from the largest logits is beneficial. In particular, the smallest logits seem to introduce noise that might be especially detrimental for OOD detection in large scale and fine-grained classification tasks with a large number of semantic classes. The main paper has a respective evaluation for our proposed score w.r.t.

OOD Method	SUN		Places		Average	
	A↑	F↓	A↑	F↓	A↑	F↓
Averaged						
MSP [7]	83.97	64.39	82.18	69.48	83.08	66.93
MaxLogit [6]	81.86	62.34	79.48	67.38	80.67	64.86
Energy [14]	79.53	65.13	76.68	70.72	78.11	67.93
GradNorm [9]	54.91	78.64	51.34	83.65	53.13	81.14
GEN (Ours)	**84.99**	**61.34**	**82.79**	**65.98**	**83.89**	**63.66**
KL Matching [6]	82.76	69.70	81.26	72.20	82.01	70.95
Mahalanobis [12]	81.88	72.25	79.40	75.36	80.64	73.81
ReAct [18]	77.61	65.08	74.25	71.42	75.93	68.25
pNML [1]	84.46	58.23	82.05	64.90	83.26	61.57
Residual [21]	78.53	77.66	75.52	80.40	77.03	79.03
ViM [21]	84.93	64.97	82.06	69.45	83.50	67.21
GEN (Ours) + Residual [21]	**88.54**	**52.37**	**84.79**	**64.05**	**86.67**	**58.21**

Table 2. *OOD Detection Performance on Unseen Datasets.*

Architecture + OOD Method	OpenImage-O		Textures		iNaturalist		ImageNet-O		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
BiT-S-R101x1										
MSP [7]	83.05	76.21	79.76	77.13	87.90	64.53	57.16	96.90	76.97	78.69
MaxLogit [6]	82.33	79.75	81.65	73.59	86.78	70.52	62.99	96.90	78.44	80.19
Energy [14]	80.59	82.00	81.10	73.91	84.52	74.93	63.56	96.35	77.44	81.80
GradNorm [9]	70.68	79.34	**83.12**	**55.72**	86.13	**58.34**	53.73	**91.90**	73.42	**71.33**
ODIN [13]	**85.64**	**72.83**	81.60	74.07	86.73	70.75	63.00	96.85	79.24	78.63
ReAct*	80.83	81.85	81.44	73.74	84.77	74.80	63.63	96.30	77.67	81.67
Shannon Entropy	83.98	80.48	81.30	76.32	88.73	69.66	60.42	97.30	78.61	80.94
GEN (Ours)	83.77	80.43	81.48	77.93	88.67	68.32	66.09	97.30	80.00	81.00
GEN (Ours) + ReAct*	83.99	80.35	81.80	77.87	**88.90**	68.03	**66.18**	97.25	**80.22**	80.88
DeiT										
MSP [7]	83.85	61.65	81.98	64.46	88.27	52.02	63.66	86.75	79.44	66.22
MaxLogit [6]	80.01	60.44	80.42	61.10	85.24	52.60	61.40	83.35	76.77	64.37
Energy [14]	74.56	66.36	77.41	64.77	78.64	65.80	60.63	**82.60**	72.81	69.88
GradNorm [9]	27.63	97.96	38.96	94.75	28.56	98.90	33.06	98.25	32.05	97.47
ODIN [13]	80.19	59.53	81.26	59.38	85.36	51.81	61.70	84.95	77.13	63.92
ReAct*	74.57	66.35	77.42	64.81	78.67	65.62	60.62	82.70	72.82	69.87
Shannon Entropy	84.71	57.54	83.50	59.05	89.29	47.55	64.93	83.00	80.61	61.78
GEN (Ours)	**88.34**	**55.63**	**86.49**	**56.36**	**92.29**	**42.52**	**71.33**	84.20	**84.61**	**59.68**
GEN (Ours) + ReAct*	88.33	55.72	86.48	56.45	92.27	42.68	**71.33**	84.25	84.60	59.77
RepVGG										
MSP [7]	84.72	64.04	78.58	72.69	87.10	55.02	61.67	91.55	78.02	70.83
MaxLogit [6]	84.48	65.45	76.31	76.71	86.21	62.15	62.89	89.90	77.47	73.55
Energy [14]	83.36	70.08	74.51	82.87	83.92	75.49	63.38	88.00	76.29	79.11
GradNorm [9]	52.48	94.81	58.25	91.30	53.40	98.20	47.79	95.60	52.98	94.98
ODIN [13]	85.22	63.48	76.77	76.14	86.37	61.40	62.50	89.70	77.72	72.68
ReAct*	84.66	69.23	76.39	82.46	84.30	74.84	65.05	**87.75**	77.60	78.57
Shannon Entropy	85.82	64.09	78.86	74.92	87.77	58.55	63.60	89.70	79.01	71.81
GEN (Ours)	87.46	59.86	80.98	67.42	90.56	45.32	66.33	91.40	81.33	66.00
GEN (Ours) + ReAct*	**88.66**	**59.31**	**83.20**	**67.07**	**91.00**	**44.78**	**68.67**	91.40	**82.88**	**65.64**
ResNet-50-D										
MSP [7]	84.56	63.55	82.71	64.71	88.57	50.38	56.14	93.75	77.99	68.10
MaxLogit [6]	81.90	65.04	79.17	66.16	86.39	53.35	54.40	92.55	75.47	69.28
Energy [14]	76.72	75.07	73.85	75.48	80.44	71.54	53.99	89.95	71.25	78.01
GradNorm [9]	38.85	97.75	54.68	90.41	41.74	98.06	40.88	98.10	44.04	96.08
ODIN [13]	81.53	64.49	80.21	63.93	86.48	52.58	52.87	93.25	75.27	68.56
ReAct*	77.01	74.88	74.32	75.12	80.59	70.94	54.27	**89.85**	71.55	77.70
Shannon Entropy	85.12	62.40	83.18	62.77	89.23	48.67	57.75	91.80	78.82	66.41
GEN (Ours)	88.09	**58.59**	86.43	**57.25**	**92.25**	**39.97**	64.24	92.50	82.75	**62.08**
GEN (Ours) + ReAct*	**88.14**	58.82	**86.50**	57.48	92.23	40.36	**64.34**	92.50	**82.80**	62.29
Swin										
MSP [7]	91.38	34.81	85.31	51.74	94.76	22.97	78.86	63.90	87.58	43.36
MaxLogit [6]	92.09	26.70	84.81	47.23	95.71	15.34	81.07	52.10	88.42	35.34
Energy [14]	91.24	26.92	82.80	51.57	95.19	15.49	82.00	**45.85**	87.81	34.96
GradNorm [9]	45.52	77.94	37.12	93.02	33.79	88.81	50.27	78.05	41.68	84.45
ODIN [13]	91.38	28.42	85.74	44.59	94.24	19.65	80.62	53.65	88.00	36.58
ReAct*	91.23	26.98	82.79	51.69	95.18	15.50	82.00	45.90	87.80	35.02
Shannon Entropy	93.16	25.61	87.15	43.84	95.95	16.21	82.13	51.95	89.60	34.40
GEN (Ours)	**94.70**	**22.60**	**89.43**	**40.95**	**97.25**	**11.55**	**84.45**	54.00	**91.46**	**32.28**
GEN (Ours) + ReAct*	94.69	22.62	89.42	41.01	**97.25**	11.56	84.44	54.00	91.45	32.30
ViT-B/16										
MSP [7]	92.17	34.96	87.13	48.45	96.13	19.14	81.88	65.00	89.33	41.89
MaxLogit [6]	96.73	16.58	93.05	30.27	98.57	6.53	89.88	44.00	94.56	24.34
Energy [14]	**96.99**	**14.78**	**93.42**	**28.14**	**98.66**	6.04	**90.49**	**41.20**	**94.89**	**22.54**
GradNorm [9]	93.79	20.94	89.76	34.26	97.34	8.54	80.38	50.90	90.32	28.66
ODIN [13]	96.86	15.68	93.01	30.60	98.57	6.58	89.85	44.15	94.57	24.25
ReAct*	96.98	14.87	93.41	28.35	**98.66**	6.01	**90.49**	42.10	**94.89**	22.83
Shannon Entropy	94.81	22.24	89.82	38.18	97.92	8.71	85.10	52.50	91.91	30.41
GEN (Ours)	96.60	17.13	92.35	34.01	98.63	**5.83**	89.67	47.60	94.31	26.14
GEN (Ours) + ReAct*	96.60	17.19	92.35	34.07	98.63	5.85	89.67	47.80	94.31	26.23
Averaged										
MSP [7]	86.62	55.87	82.58	63.20	90.45	44.01	66.56	82.97	81.55	61.51
MaxLogit [6]	86.26	52.33	82.57	59.18	89.82	43.41	68.77	76.47	81.85	57.85
Energy [14]	83.91	55.87	80.52	62.79	86.89	51.55	69.01	**73.99**	80.08	61.05
GradNorm [9]	54.82	78.12	60.31	76.58	56.83	75.14	51.02	85.47	55.75	78.83
ODIN [13]	86.80	50.74	83.10	58.12	89.62	43.79	68.42	77.09	81.98	57.44
ReAct*	84.21	55.69	80.96	62.70	87.03	51.29	69.34	74.10	80.39	60.94
Shannon Entropy	81.98	52.06	83.97	59.18	91.48	41.56	68.99	70.71	83.09	57.63
GEN (Ours)	89.83	49.04	86.19	**55.65**	93.27	35.59	73.69	77.83	85.74	54.53
GEN (Ours) + ReAct*	**90.07**	**49.00**	**86.62**	55.66	**93.38**	**35.54**	**74.11**	77.87	**86.04**	**54.52**

Table 3. *Performance of Post-hoc Methods.* *BiT-S-R101x1*, *DeiT*, *RepVGG*, *ResNet-50-D*, *Swin*, and *ViT-B/16*are included along with the averaged performance across models. The ID dataset is ImageNet-1K, the OOD datasets are OpenImage-O, Textures, iNaturalist and ImageNet-O. Units for AUROC and FPR95 are percentages. The best performing method is in bold, the second best is underlined.

Architecture + OOD Method	OpenImage-O		Textures		iNaturalist		ImageNet-O		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
BiT-S-R101x1										
KL Matching [6]	87.94	54.92	86.91	50.89	92.95	**33.19**	65.76	86.80	83.39	56.45
Mahalanobis [12]	82.62	66.24	97.33	13.95	85.79	64.71	80.37	70.20	86.53	53.77
ReAct [18]	82.54	79.06	84.85	68.60	86.89	70.16	64.81	95.65	79.77	78.37
pNML [1]	88.62	55.27	93.59	22.25	**93.12**	38.21	67.27	86.35	85.65	50.52
Residual [21]	80.20	68.05	97.67	11.14	76.93	80.18	81.58	65.60	84.09	56.24
ViM [21]	89.96	49.01	**98.92**	**4.63**	89.38	55.09	83.85	**61.25**	90.53	42.50
GEN (Ours) + ReAct [18]	87.44	70.07	88.35	63.86	91.63	53.30	70.81	93.80	84.56	70.26
GEN (Ours) + Residual [21]	**91.75**	**43.83**	98.54	5.78	92.25	47.13	**83.88**	63.70	**91.61**	**40.11**
DeiT										
KL Matching [6]	87.29	60.58	84.88	63.35	90.56	50.45	71.09	84.25	83.46	64.66
Mahalanobis [12]	89.18	64.84	83.60	77.13	91.55	58.78	**75.98**	90.25	85.08	72.75
ReAct [18]	75.95	64.80	78.03	64.28	80.63	62.22	61.17	**82.25**	73.95	68.39
pNML [1]	86.68	57.86	86.02	**56.32**	90.54	47.45	69.10	83.95	83.09	61.39
Residual [21]	88.16	68.56	82.70	77.58	91.30	58.45	74.58	91.30	84.19	73.97
ViM [21]	89.21	63.84	84.43	73.12	92.13	52.86	75.34	89.40	85.28	69.81
GEN (Ours) + ReAct [18]	88.43	**55.84**	**86.46**	56.90	92.30	**43.06**	71.42	84.45	84.65	**60.06**
GEN (Ours) + Residual [21]	**89.46**	61.96	84.90	70.04	**92.58**	49.05	75.42	89.00	**85.59**	67.51
RepVGG										
KL Matching [6]	86.49	57.53	83.20	61.92	89.06	42.24	66.42	84.90	81.29	61.65
Mahalanobis [12]	85.16	66.18	92.69	32.13	89.14	58.92	76.65	81.95	85.91	59.80
ReAct [18]	67.37	96.93	68.25	94.13	66.25	99.19	59.79	94.90	65.42	96.29
pNML [1]	88.75	**49.92**	86.02	44.22	89.91	46.67	68.23	80.65	83.23	55.37
Residual [21]	81.70	66.73	93.03	28.66	86.05	62.45	75.06	79.90	83.96	59.44
ViM [21]	88.68	53.82	**93.68**	**23.88**	91.33	46.91	**76.90**	**79.20**	**87.65**	**50.95**
GEN (Ours) + ReAct [18]	**88.99**	52.85	90.35	48.82	91.82	**36.76**	74.13	85.90	86.32	56.08
GEN (Ours) + Residual [21]	**88.99**	53.89	92.73	28.00	**92.16**	42.80	76.09	82.00	87.49	51.67
ResNet-50-D										
KL Matching [6]	87.13	60.88	86.06	61.92	90.48	47.66	66.96	88.85	82.66	64.83
Mahalanobis [12]	88.69	58.71	94.15	28.14	89.51	62.34	80.10	76.35	88.11	56.38
ReAct [18]	81.63	66.16	84.68	54.17	84.55	60.71	59.86	84.75	77.68	66.45
pNML [1]	88.72	**47.86**	91.28	32.62	91.36	39.53	65.39	80.80	84.19	**50.20**
Residual [21]	86.47	62.86	94.63	25.66	84.70	75.79	**81.10**	**73.45**	86.72	59.44
ViM [21]	90.00	53.50	**95.84**	**20.48**	89.29	64.43	80.98	74.70	89.03	53.28
GEN (Ours) + ReAct [18]	89.20	55.86	89.17	50.93	**92.72**	**38.48**	67.24	91.05	84.58	59.08
GEN (Ours) + Residual [21]	**90.18**	53.41	95.24	23.51	90.67	58.33	80.19	78.50	**89.07**	53.44
Swin										
KL Matching [6]	91.86	39.93	86.82	53.24	94.75	27.76	81.78	67.30	88.80	47.06
Mahalanobis [12]	94.35	34.85	89.95	49.09	98.69	5.38	85.43	73.65	92.11	40.74
ReAct [18]	91.83	25.92	83.33	50.54	95.90	13.84	82.26	45.75	88.33	34.01
pNML [1]	95.53	**19.29**	91.55	**33.29**	97.84	8.98	87.22	**45.05**	93.03	**26.65**
Residual [21]	94.44	33.40	91.36	43.26	98.90	4.79	86.66	68.65	92.84	37.53
ViM [21]	**95.93**	24.43	**92.40**	37.98	**99.29**	**2.62**	**88.74**	59.00	**94.09**	31.01
GEN (Ours) + ReAct [18]	95.09	21.94	89.71	41.22	97.75	9.45	84.84	56.10	91.85	32.18
GEN (Ours) + Residual [21]	95.73	25.06	92.23	37.66	99.13	3.10	88.07	61.50	93.79	31.83
ViT-B/16										
KL Matching [6]	93.46	29.58	88.75	43.84	96.88	15.03	84.14	55.70	90.81	36.04
Mahalanobis [12]	**97.33**	14.32	94.21	25.27	**99.53**	**2.15**	**92.78**	37.00	95.96	19.69
ReAct [18]	97.24	**13.99**	93.54	27.62	99.01	4.21	90.74	41.90	95.13	21.93
pNML [1]	95.38	20.33	90.98	34.53	98.18	7.69	86.44	49.95	92.75	28.12
Residual [21]	91.86	36.41	92.04	34.73	98.58	6.56	88.35	48.30	92.71	31.50
ViM [21]	97.30	14.39	**95.31**	**20.14**	99.41	2.56	92.61	**36.75**	**96.16**	**18.46**
GEN (Ours) + ReAct [18]	96.77	16.37	92.41	33.70	98.95	4.34	89.79	47.95	94.48	25.59
GEN (Ours) + Residual [21]	97.29	14.17	94.41	25.17	99.38	2.67	91.83	40.75	95.73	20.69
Averaged										
KL Matching [6]	89.03	50.57	86.10	55.86	92.45	36.05	72.69	77.97	85.07	55.11
Mahalanobis [12]	89.56	50.86	91.99	37.62	92.37	42.05	81.89	71.57	88.95	50.52
ReAct [18]	82.76	57.81	82.11	59.89	85.54	51.72	69.77	74.20	80.05	60.91
pNML [1]	90.61	**41.76**	89.91	37.20	93.49	**31.42**	73.94	71.12	86.99	45.38
Residual [21]	87.14	56.00	91.90	36.84	89.41	48.04	81.22	71.20	87.42	53.02
ViM [21]	91.85	43.16	**93.43**	**30.04**	93.47	37.41	**83.07**	**66.72**	90.45	44.33
GEN (Ours) + ReAct [18]	90.59	46.94	88.76	50.91	93.89	32.70	75.76	76.76	87.25	51.83
GEN (Ours) + Residual [21]	**92.23**	42.05	93.01	31.69	**94.36**	33.85	82.58	69.24	**90.55**	**44.21**

Table 4. *Performance of Methods Requiring ID Data. BiT-S-R101x1*, *DeiT*, *RepVGG*, *ResNet-50-D*, *Swin*, and *ViT-B/16*are included along with the averaged performance across models. The ID dataset is ImageNet-1K, the OOD datasets are OpenImage-O, Textures, iNaturalist and ImageNet-O. Units for AUROC and FPR95 are percentages. The best performing method is in bold, the second best is underlined.

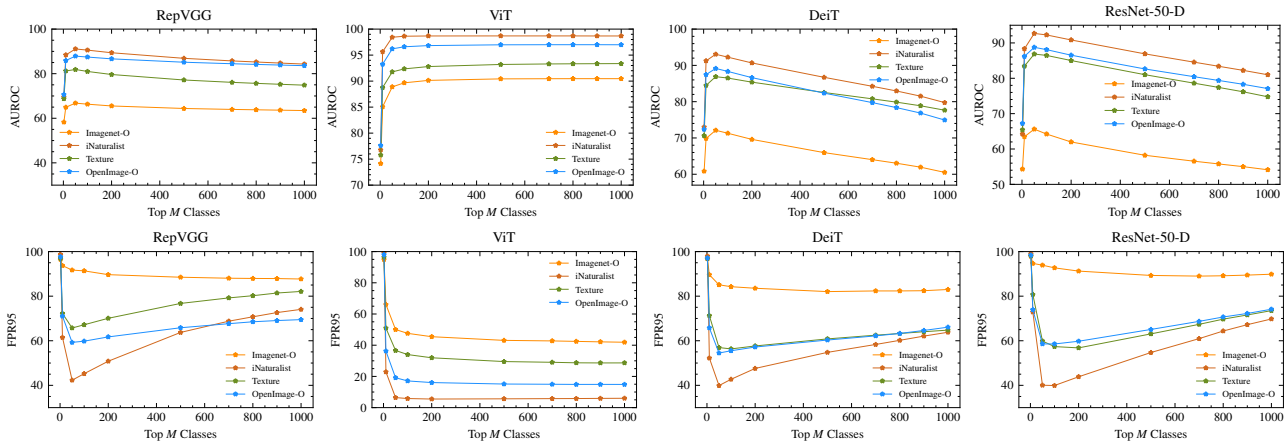

Figure 2. *OOD Detection Performance of GEN Score with Varying M.* Reported are (top) AUROC and (bottom) FPR95 values.

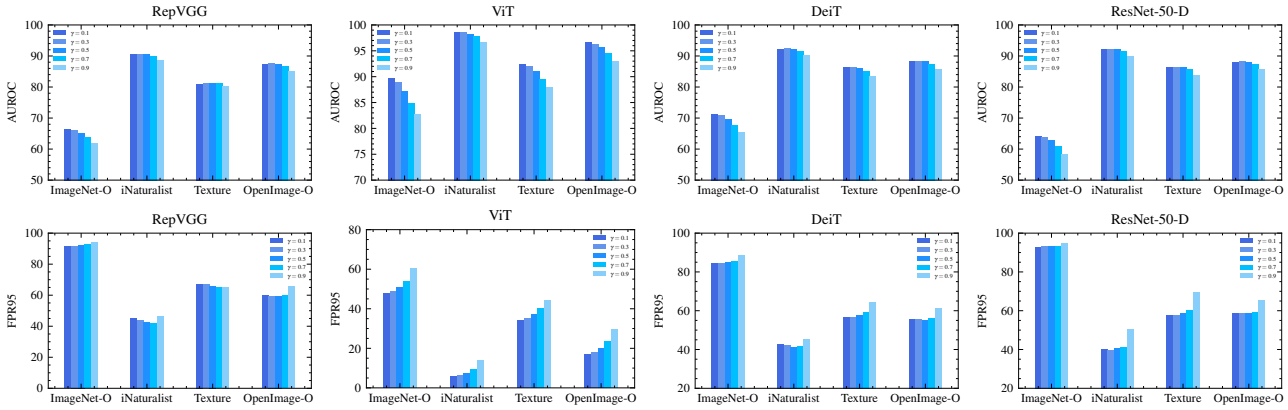

Figure 3. *OOD Detection Performance of GEN Score with Varying γ.* Reported are (top) AUROC and (bottom) FPR95 values.

Swin [15] and BiT [10] architectures (see Fig. 4 in the main paper), and here we demonstrate a similar behavior for the Energy [14] score. The original Energy [14] method simply uses all logits to calculate the score. We instead utilize a subset of M largest logits. The results for $M = \{1, 2, 5, 10, 20, 30, 50, 100, 200, 500, 700, 1000\}$ are shown in Fig. 4. It shows that AUROC decreases and FPR95 increases for most of the classifiers except for ViT [4] when the number of incorporated logits is increased. That is to say, using more logits indeed degrades the OOD detection performance of most architectures (with the exception of ViT [4]).

G. Sensitivity to Temperature Scaling

A pretrained network might also be adjusted to yield better calibrated predictions. Since calibration methods rely on some training data, which cannot be assumed to be available, we investigate into the sensitivity of post-hoc OOD scores w.r.t. applying a classifier calibration. In particular, we simulate the effects of the simple and popular temperature scaling approach [16], which scales the logits by an inverse temperature $1/T$. Once the right temperature T is determined (using validation data), it can be absorbed into the layer generated the logits (and therefore the original logits might become inaccessible). We simulate temperatures $T \in \{0.2, 0.5, 1, 2, 5\}$ and illustrate the sensitivity of AUROC and FPR95 values for post-hoc OOD detection scores in Table 5.

The MaxLogit [6] score is agnostic to temperature scaling by construction. It can be seen that GEN is relatively insensitive to temperature scaling in terms of AUROC values, but shows some sensitivity in the FPR95 results. Energy [14] is slightly less sensitive than GEN in terms of FPR95 score, but more sensitive in terms of AUROC score, and MSP [7] overall is more sensitive. GradNorm [9] shows the highest sensitivity to temperature scaling. Note that all methods (except the invariant MaxLogit [6] score) are relatively sensitive in their FPR95 results.

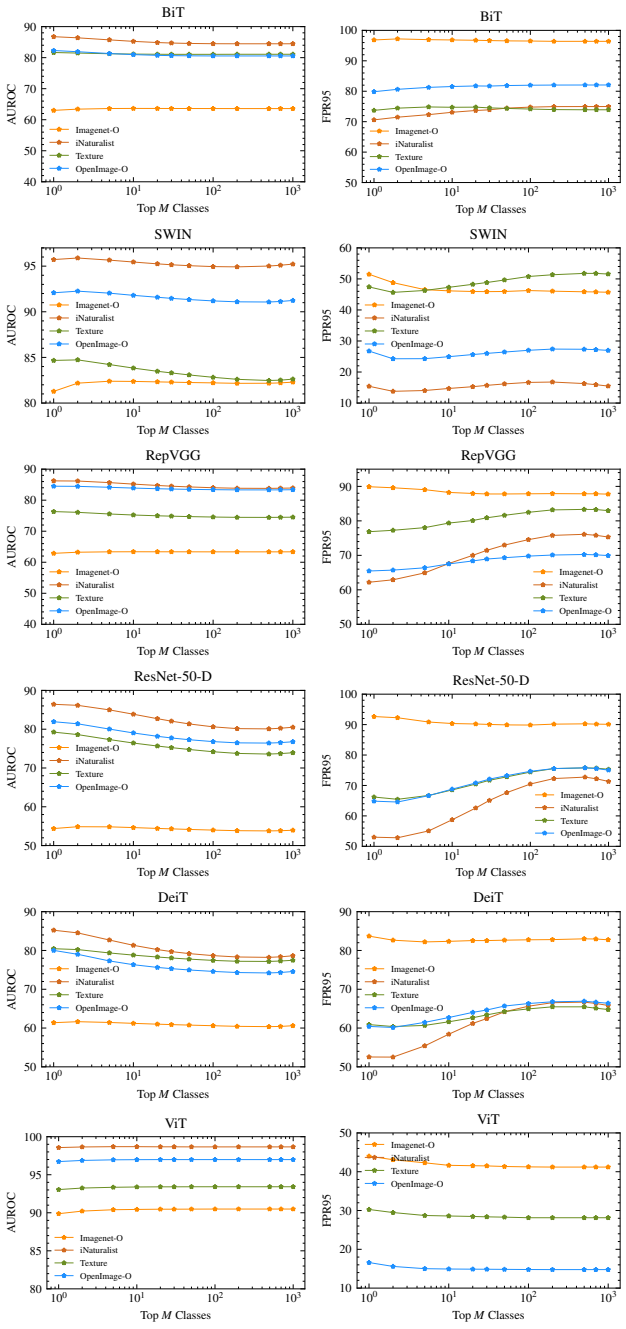

Figure 4. *OOD Detection Performance of Energy Score with Varying M.* Reported are (left) AUROC and (right) FPR95 values.

Method	Average Performance AUROC ↑	FPR95 ↓
MSP [7] Temp-0.2	77.99	78.75
MSP [7] Temp-0.5	81.85	69.41
MSP [7] Temp-1	87.58	43.36
MSP [7] Temp-2	**88.51**	37.33
MSP [7] Temp-5	88.03	**37.24**
MaxLogit [6] Temp-0.2	88.42	35.34
MaxLogit [6] Temp-0.5	88.42	35.34
MaxLogit [6] Temp-1	88.42	35.34
MaxLogit [6] Temp-2	88.42	35.34
MaxLogit [6] Temp-5	88.42	35.34
Energy [14] Temp-0.2	88.48	35.03
Energy [14] Temp-0.5	**88.70**	**33.82**
Energy [14] Temp-1	87.81	34.96
Energy [14] Temp-2	62.34	61.23
Energy [14] Temp-5	62.43	61.67
GradNorm [9] Temp-0.2	13.47	99.84
GradNorm [9] Temp-0.5	15.70	98.89
GradNorm [9] Temp-1	**41.68**	**84.45**
GradNorm [9] Temp-2	19.25	99.50
GradNorm [9] Temp-5	14.37	99.85
GEN (Ours) Temp-0.2	89.53	38.94
GEN (Ours) Temp-0.5	90.82	34.74
GEN (Ours) Temp-1	**91.46**	**32.27**
GEN (Ours) Temp-2	87.24	61.46
GEN (Ours) Temp-5	84.23	69.88

Table 5. *Sensitivity to Temperature Scaling.* The reported is the average performance across 6 classifiers —*BiT-S-R101x1*, *DeiT*, *RepVGG*, *ResNet-50-D*, *Swin*, and *ViT-B/16*— and 4 datasets — OpenImage-O, Textures, iNaturalist, and ImageNet-O.

ARCH + OOD Method	iNaturalist A↑	iNaturalist F↓	Texture A↑	Texture F↓	OpenImage-O A↑	OpenImage-O F↓	ImageNet-O A↑	ImageNet-O F↓
BiT-S-R101x1								
FeatureNorm	74.67	77.50	74.30	65.95	53.97	87.64	50.54	93.30
ProbsDistance	86.66	73.96	81.27	77.05	82.51	82.49	65.64	96.95
GradNorm [9]	86.13	**58.34**	**83.12**	**55.72**	70.68	79.34	53.73	**91.90**
GEN (Ours)	**88.67**	68.32	81.48	77.93	**83.77**	80.43	**66.09**	97.30
Swin								
FeatureNorm	4.05	100.00	15.65	99.61	11.32	99.90	22.55	99.90
ProbsDistance	94.64	20.78	86.33	45.43	92.45	26.78	82.91	**47.85**
GradNorm [9]	33.79	88.81	37.12	93.02	45.52	77.94	50.27	78.05
GEN (Ours)	**97.25**	**11.55**	**89.43**	**40.95**	**94.70**	**22.60**	**84.45**	54.00

Table 6. *Feature vs. Probability Space.* Using feature norms in most cases degrades the performance hence making GradNorm [9] unstable especially on the largest OpenImage-O [11] dataset.

H. Comparison with GradNorm [9]

We conduct extra experiments on BiT [10] and Swin [15] to test our approach. First, we compare our method to the recent post-hoc method GradNorm [9], which claims that using joint information from feature space and probability space is helpful for OOD detection. Based on our experimental observations, it is not always true, and the performance depends on the model architecture. Second,

It is claimed in GradNorm [9] that using joint information from feature space and probability space could achieve better OOD results. There, feature information is represented as feature norm $\|\mathbf{z}\|_1$, and probability information is compressed as the total variation (*i.e.* l_1-distance) between uniform distribution and predictive distribution $\|\mathbf{p} - \mathbf{u}\|_1$. We further investigate whether this conclusion holds for

other architectures and OOD datasets. We reproduce and extend Table 5 of GradNorm [9] for all six architectures and six OOD datasets. The results for BiT [10] and Swin [15] are shown in Table 6, and results for other architectures can be found in supplementary material. It can be seen that feature norms $\|\mathbf{z}\|_1$ are not always distinctive for OOD detection and could cause occasional bad performance of GradNorm [9]. Besides, our score which only uses information from probability space outperforms the score using probability distance and GradNorm [9] in most datasets.

I. Analysis of GradNorm [9]: Dependence on the Checkpoint

We compare the performance of OOD detection methods for BiT-S-R101x1 architecture with two different weights (checkpoints). The first one is the official checkpoint of BiT [10] used by ViM [21], and the second one is the fine-tuned set of weights provided by GradNorm [9]. The results in Table 7 are averaged AUROC and FPR95 on four OOD datasets. One can notice that GradNorm [9] performs worse when official checkpoint is used. However the downstream performance—ImageNet classification, see Table 2 in the main paper—is worse for the fine-tuned checkpoint from GradNorm [9] indicating a certain bias in GradNorm [9] checkpoint. Moreover, GEN consistently outperforms GradNorm [9] on the two most challenging datasets, OpenImage-O and ImageNet-O.

Arch + Method	**iNaturalist**		**Texture**		**OpenImage-O**		**ImageNet-O**	
	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
BiT-S-R101x1								
MSP [7]	87.90	64.53	79.76	77.1	83.05	**76.21**	57.16	96.90
MaxLogit [6]	86.78	70.52	81.65	73.59	82.33	79.75	62.99	96.90
GradNorm [9]	86.13	**58.34**	**83.12**	**55.72**	70.68	79.34	53.73	**91.90**
GEN (Ours)	**88.67**	68.32	81.48	77.93	**83.77**	80.43	**66.09**	97.30
BiT-S-R101x1 [9]								
MSP [7]	87.57	63.94	76.87	81.51	80.18	80.56	55.55	97.65
MaxLogit [6]	89.38	62.71	78.53	79.81	80.35	81.93	59.26	97.70
GradNorm [9]	**90.45**	**49.41**	**83.30**	**58.74**	73.59	**79.13**	54.43	**93.45**
GEN (Ours)	89.03	68.20	77.85	86.45	**81.34**	83.31	**63.26**	97.45

Table 7. *OOD Detection Performance Depends on Checkpoint.* OOD detection results for BiT-S-R101x1 with official checkpoint [10] and the one provided by GradNorm [9]. The performance of GradNorm [9] gets worse for the official weights.

References

[1] Koby Bibas, Meir Feder, and Tal Hassner. Single layer predictive normalized maximum likelihood for out-of-distribution detection. In *NeurIPS*, 2021. 2, 4

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 1

[3] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, 2021. 1, 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2, 5

[5] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, 2019. 1, 2

[6] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022. 2, 3, 4, 5, 6, 7

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 2, 3, 5, 6, 7

[8] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 1

[9] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021. 1, 2, 3, 5, 6, 7

[10] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020. 1, 5, 6, 7

[11] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 6

[12] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 2, 4

[13] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 2, 3

[14] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 2, 3, 5, 6

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1, 5, 6, 7

[16] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 5

[17] 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. *IJCV*, 115(3):211–252, 2015. 1

[18] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021. 1, 2, 4

[19] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1, 2

[20] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1

[21] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022. 1, 2, 4, 7

[22] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 2

[23] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 2