

## A APPENDIX

This supplementary material presents additional details of Section 3.2, 4.1 and 4.2. It is organized as follows:

- **Detailed results for CIFAR-10 and CIFAR-100 benchmarks** We present detailed results on CIFAR-10 and CIFAR-100 benchmarks and compare with other competitive methods in Sec. 4.2.2.
- **Description of OOD datasets** We provide additional information about the OOD datasets used in our experiments in Sec. 4.1.
- **GradRect with different  $p$  value of  $L_p$  norm** We conduct an experimental study to evaluate the role of  $p$  value in  $L_p$  norm in Sec 3.2.

## B DETAILED RESULTS FOR CIFAR-10 AND CIFAR-100 BENCHMARKS

We report the detailed results on CIFAR-10 benchmark in Table 7 and CIFAR-100 benchmark in Table 8. For both tables, the results except for GradNorm and GradRect are in line with (Ahn et al., 2023).

Table 7: Comparison in OOD detection on CIFAR-10 benchmark. Backbone is DenseNet pretrained on CIFAR-10. Baseline methods include post-hoc methods (MSP Hendrycks & Gimpel (2016), ODIN Liang et al. (2018), Mahalanobis Lee et al. (2018), Free Energy Liu et al. (2020), ReAct Sun et al. (2021), DICE Sun & Li (2022) and GradNorm Huang et al. (2021)). All values in this table are percentages.  $\downarrow$  (or  $\uparrow$ ) indicates smaller (or larger) values are preferred.

Method	SVHN		LSUN-c		LSUN-r		iSUN		Textures		Places365		Average	
	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$
MSP Hendrycks & Gimpel (2016)	47.24	93.48	33.57	95.54	42.10	94.51	42.31	94.52	64.15	88.15	63.02	88.57	48.73	92.46
ODIN Liang et al. (2018)	25.29	94.57	4.70	98.86	3.09	99.02	3.98	98.90	57.50	82.38	52.85	88.55	24.57	93.71
Mahalanobis Lee et al. (2018)	6.42	98.31	56.55	86.96	9.14	97.09	9.78	97.25	21.51	92.15	85.14	63.15	31.42	89.15
Energy Liu et al. (2020)	40.61	93.99	3.81	99.15	9.28	98.12	10.07	98.07	56.12	86.43	39.40	91.64	26.55	94.57
ReAct Sun et al. (2021)	41.64	93.87	5.96	98.84	11.46	97.87	12.72	97.72	43.58	90.37	43.31	91.01	26.45	94.95
DICE Sun & Li (2022)	25.99	95.90	0.26	99.92	3.91	99.20	4.36	99.14	41.90	88.18	48.59	89.13	20.83	95.24
GradNorm Huang et al. (2021)	17.60	95.14	0.90	99.78	4.80	98.87	5.20	98.80	55.70	88.08	43.60	89.84	21.30	95.08
GradRect	13.00	96.17	0.30	99.88	4.90	99.12	7.30	98.17	50.00	89.24	43.20	89.88	19.78	95.41

Table 8: Comparison in OOD detection on CIFAR-100 benchmark. Backbone is DenseNet pretrained on CIFAR-100. Baseline methods include post-hoc methods (MSP Hendrycks & Gimpel (2016), ODIN Liang et al. (2018), Mahalanobis Lee et al. (2018), Free Energy Liu et al. (2020), ReAct Sun et al. (2021), DICE Sun & Li (2022) and GradNorm Huang et al. (2021)). All values in this table are percentages.  $\downarrow$  (or  $\uparrow$ ) indicates smaller (or larger) values are preferred.

Method	SVHN		LSUN-c		LSUN-r		iSUN		Textures		Places365		Average	
	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$
MSP Hendrycks & Gimpel (2016)	81.70	75.40	60.49	85.60	85.24	69.18	85.99	70.17	84.79	71.48	82.55	74.31	80.13	74.36
ODIN Liang et al. (2018)	41.35	92.65	10.54	97.93	65.22	84.22	67.05	83.84	82.34	71.48	82.32	76.84	58.14	84.49
Mahalanobis Lee et al. (2018)	22.44	95.67	68.90	86.30	23.07	94.20	31.38	93.21	62.39	79.39	92.66	61.39	55.37	82.73
Energy Liu et al. (2020)	87.46	81.85	14.72	97.43	70.65	80.14	74.54	78.95	84.15	71.03	79.20	77.72	68.45	81.19
ReAct Sun et al. (2021)	83.81	81.41	25.55	94.92	60.08	87.88	65.27	86.55	77.78	78.95	82.65	74.04	62.27	84.47
DICE Sun & Li (2022)	54.65	88.84	0.93	99.74	49.40	91.04	48.72	90.08	65.04	76.42	79.58	77.26	49.72	87.23
GradNorm Huang et al. (2021)	31.40	89.83	0.40	99.88	54.60	86.43	54.60	86.41	94.00	76.39	63.40	79.97	49.73	86.48
GradRect	30.20	89.63	0.60	99.85	33.80	91.93	29.40	91.58	94.20	74.76	65.00	77.87	42.20	87.60

## C DESCRIPTION OF OOD DATASETS

To ensure proper performance evaluation, we need to make sure that the OOD datasets don't contain samples with ID category. In experiments, for CIFAR-10 and CIFAR-100 benchmarks, we leverage commonly-used six OOD datasets for evaluation. Specifically, we employ Texture Cimpoi et al. (2014), SVHN Netzer et al. (2011), Places365 Zhou et al. (2017), LSUN-Crop Yu et al. (2015), LSUN-Resize Yu et al. (2015), and iSUN Xu et al. (2015). The detailed description of OOD datasets is listed below:

The Street View House Numbers (SVHN) dataset consists of images depicting house numbers. The dataset has ten categories corresponding to the digits 0-9. Places365 includes a collection of large-scale scene images classified into 365 distinct categories. The test set of this dataset consists of

900 images per category. LSUN consists of 10,000 images depicting scenes. Both LSUN-Crop and LSUN-Resize are the cropped and resized version of LSUN dataset respectively. iSUN is a large-scale eye-tracking dataset.

For the ImageNet case, we select four testsets from subsets of iNaturalist [Van Horn et al. \(2018\)](#), SUN [Xu et al. \(2015\)](#), Places365 [Zhou et al. \(2017\)](#), and Texture [Cimpoi et al. \(2014\)](#). The detailed description of them is listed below:

iNaturalist comprises a comprehensive collection of 859,000 images panning more than 5,000 distinct species of plants and animals. SUN and Places are scene datasets. Texture comprises a vast collection of 5,640 real-world texture images across 47 distinct categories. To ensure the validity of the evaluation process, in this research study, we leverage OOD datasets craft by [Huang & Li \(2021\)](#) with categories disjoint from the ImageNet-1k dataset.

## D GRADRECT WITH DIFFERENT P VALUE OF $L_p$ NORM

As stated in Sec 3.1, we choose  $p=2$  by default for  $L_p$  norm in GradRect score. For other choices of  $p$ , we conduct experiments with different  $p$  value of  $L_p$  norm in GradRect score and report the results using  $L_{1\sim 4}$ , the fraction norm  $L_{p=0.5}$  and  $L_\infty$  in Table 9.

Among all different  $p$  values considered, the  $L_1$  norm achieves the best performance due to the characteristic of  $L_1$  norm. Compared to other higher-order norms unfairly emphasizing dimensions while disregarding others,  $L_1$  norm can treat all dimensions in the gradient space equally when capturing information. Note that the extreme case  $L_\infty$  has the worst performance for the reason only the largest element is considered and all other information is ignored.

Table 9: Effect of different  $p$  value of  $L_p$  norm in GradRect score. The experiments are conducted on the ImageNet-1k benchmark.

$L_p$ norm	FPR95 ↓	AUROC ↑
$p = 0.5$	53.65	85.13
$p = 1$	46.70	89.10
$p = 2$	47.08	88.88
$p = 3$	61.86	81.34
$p = \infty$	79.57	75.31