Supplementary Material
Heatmap-based Out-of-Distribution Detection

Julia Hornauer
Ulm University, Germany
julia.hornauer@uni-ulm.de

Vasileios Belagiannis
Friedrich-Alexander-University Erlangen-Nürnberg, Germany
vasileios.belagiannis@fau.de

1. Further Results Ablation Study

1.1. OOD Training Data Size

In Fig. 1, Fig. 2 and Fig. 3, the OOD detection results in terms of AUPR-Success, AUPR-Error and FPR at 95% TPR for alternating the OOD training set size with sets of $\{500, 1000, 5000, 10000, 20000, 50000, 80000\}$ samples are shown.

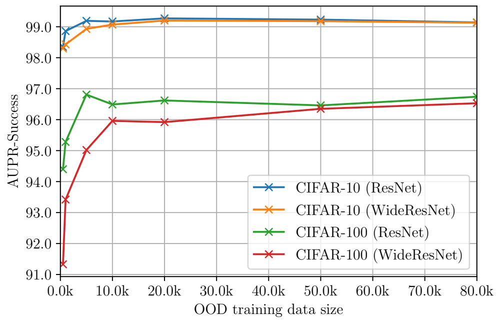


Figure 1: AUPR-Success results when alternating the OOD training set size with sets of $\{500, 1000, 5000, 10000, 20000, 50000, 80000\}$ samples. The in-distribution dataset size is fixed to 50000.

1.2. Number of Classifier Feature Layers

As mentioned, we only rely on the penultimate classifier layer as input to the decoder. Here, we evaluate the usage of more than one feature layer. Therefore, we gradually increase the number of classifier layers forwarded to the decoder. Furthermore, the additional layers are not used as input to the decoder but added to intermediate stages of the decoder. Fig. 4 shows the AUROC from relying only on the penultimate layer as input (1 layer) and adding up to 3 additional feature representations (4 layers). More precisely, we sequentially add layers starting from late classifier stages to early stages. Note that we obtain the intermediate feature representations after the respective ResNet and WideResNet module blocks. Fig. 5, Fig. 6 and Fig. 7 show the performance in terms of AUPR-Success, AUPR-Error and FPR at 95% TPR from relying only on the penultimate layer as input (1 layer) and adding up to 3 additional feature representations (4 layers).

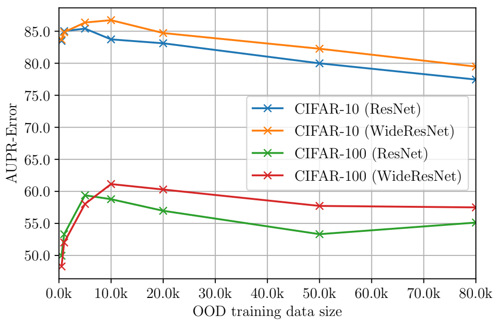


Figure 2: AUPR-Error results when alternating the OOD training set size with sets of $\{500, 1000, 5000, 10000, 20000, 50000, 80000\}$ samples. The in-distribution dataset size is fixed to 50000.

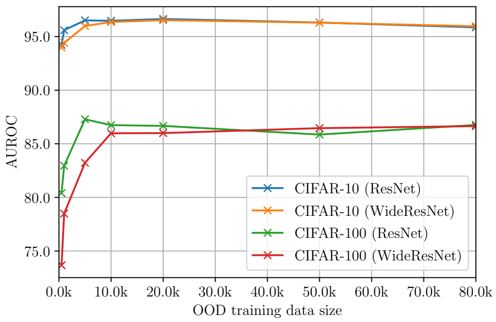


Figure 3: FPR at 95% TPR results when alternating the OOD training set size with sets of $\{500, 1000, 5000, 10000, 20000, 50000, 80000\}$ samples. The in-distribution dataset size is fixed to 50000.

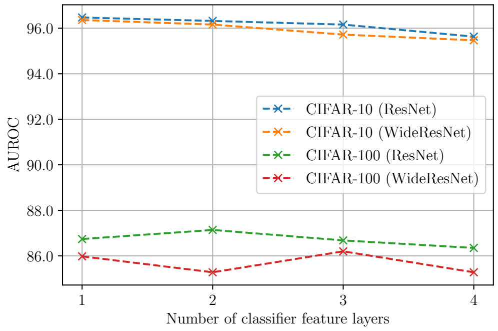


Figure 4: AUROC when using multiple feature representations as decoder input. Variation from relying only on the penultimate layer as input (1 layer) and adding up to 3 additional feature representations (4 layers).

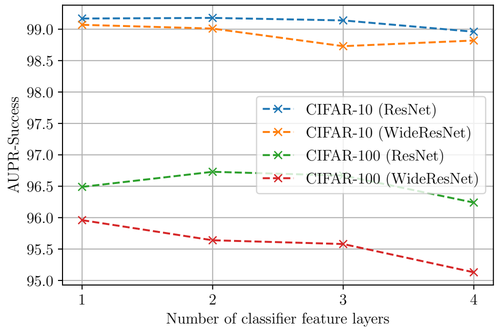


Figure 5: AUPR-Success when using multiple feature representations as decoder input. Variation from relying only on the penultimate layer as input (1 layer) and adding up to 3 additional feature representations (4 layers).

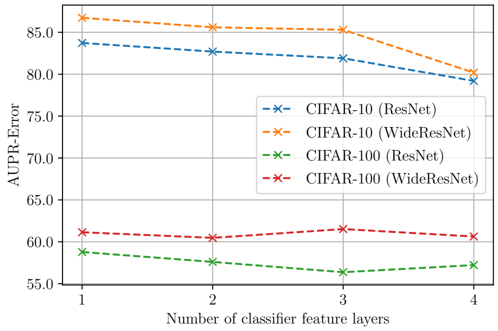


Figure 6: AUPR-Error when using multiple feature representations as decoder input. Variation from relying only on the penultimate layer as input (1 layer) and adding up to 3 additional feature representations (4 layers).

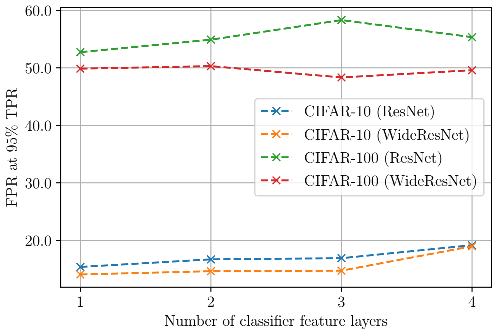


Figure 7: FPR at 95% TPR when using multiple feature representations as decoder input. Variation from relying only on the penultimate layer as input (1 layer) and adding up to 3 additional feature representations (4 layers).

2. Further Lighting Effect Results

In this section, we report the OOD detection performance when augmenting the brightness or the contrast of in-distribution test data. We augment $\frac{1}{5}$ of the in-distribution test data by setting either the brightness factor (B) to a value selected from B $= \{1.5, 1.75, 2.0, 2.25, 2.5\}$ or the contrast factor (C) to a value selected from C $= \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The higher the brightness value or the lower the contrast value, the more augmented in-distribution samples should be detected as out-of-distribution samples because the relevant features are no longer recognizable. The experiment is conducted with ResNet18 pre-trained on CIFAR-10 as in-distribution dataset. To evaluate the ODD detection performance, we use the in-distribution testset and the augmented in-distribution images labeled as OOD. The OOD detection performance in terms of AUROC, AUPR-S, AUPR-E and FPR-95 for different brightness values are reported in Tab. 1. It is clearly visible that all metrics improve with higher brightness values. In Fig. 8 and Fig. 9, we provide additional qualitative examples. Here, a horse and a truck are visualized for different brightness values with corresponding heatmap predictions. The heatmaps clearly show a higher response for images with higher brightness augmentation. Again, with higher brightness value the in-distribution features of the images are less recognizable and therefore the images should be labeled as out-of-distribution. In Tab. 2, the OOD detection performance is provided for different contrast values. As expected, the OOD detection performance increases with reduced image contrast. In Fig. 10 and Fig. 11, qualitative examples of images with reduced contrast are visualized. In both cases, the heatmaps show larger OOD regions for images with lower contrast.

Brightness	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
1.5	56.61	85.39	24.01	88.55
1.75	62.25	86.98	31.61	81.65
2.0	68.45	89.12	41.30	73.50
2.25	73.75	91.04	49.97	65.90
2.5	78.26	92.61	57.07	58.15

Table 1: Evaluation of the influence when images are augmented with increased brightness. Comparison of the OOD detection performance in terms of AUROC, AUPR-Success, AUPR-Error, and FPR at 95% TPR. A subset of the in-distribution test data is augmented with increased brightness values and labeled as out-of-distribution. The experiment is conducted with ResNet pre-trained on CIFAR-10. ↑ indicates that larger values are better, whereas ↓ marks that lower values are better.

Brightness	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
0.5	58.81	86.52	23.62	88.55
0.4	65.97	88.89	31.30	80.25
0.3	76.92	92.55	47.88	63.80
0.2	92.17	97.77	74.28	30.50
0.1	98.62	99.75	90.68	1.45

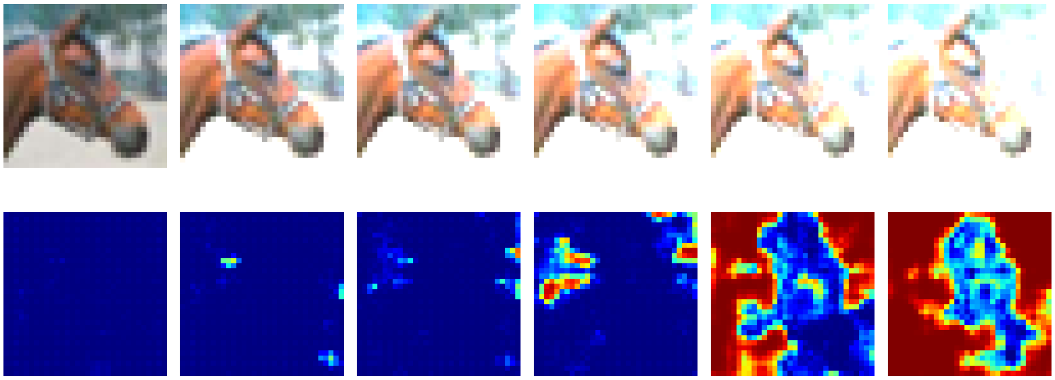
Table 2: Evaluation of the influence when images are augmented with reduced contrast. Comparison of the OOD detection performance in terms of AUROC, AUPR-Success, AUPR-Error, and FPR at 95% TPR. A subset of the in-distribution test data is augmented with reduced contrast values and labeled as out-of-distribution. The experiment is conducted with ResNet pre-trained on CIFAR-10. ↑ indicates that larger values are better, whereas ↓ marks that lower values are better.

3. Additional Out-of-Distribution Detection Results

In Tab. 3 to Tab. 7 the out-of-distribution (OOD) detection results are listed for the individual OOD test sets.

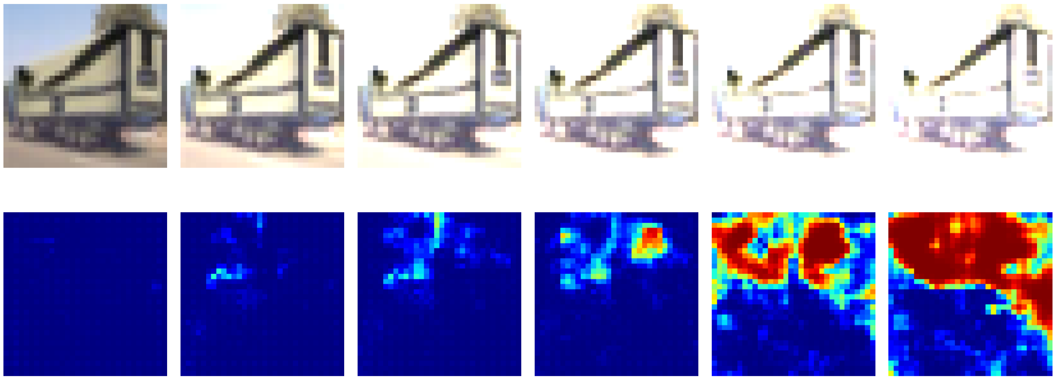
4. Further Visual Illustrations

Fig. 12 shows qualitative results for WideResNet trained on CIFAR-100. The first row presents the original input images, while the estimated heatmaps are in the second row. Fig. 12a shows in-distribution examples of CIFAR-100. In both cases, the heatmaps show no response compared to the original image. Thus, the heatmap entries are close to zero, which in turn means that features extracted with the classifier are representative of in-distribution samples. From Fig. 12b to Fig. 12g, we show the OOD input examples for the classifier trained on CIFAR-100 from all six datasets. The images highlight regions with high out-of-distribution responses. Overall, the OOD examples cover a wide range of OOD types from digits in Fig. 12f to buildings in Fig. 12b or textures in Fig. 12g. The heatmaps contain regions with entries larger than zero, indicating that the images differ from the closest in-distribution sample. The high responses in those different OOD categories again demonstrate the generalization capabilities of our approach.



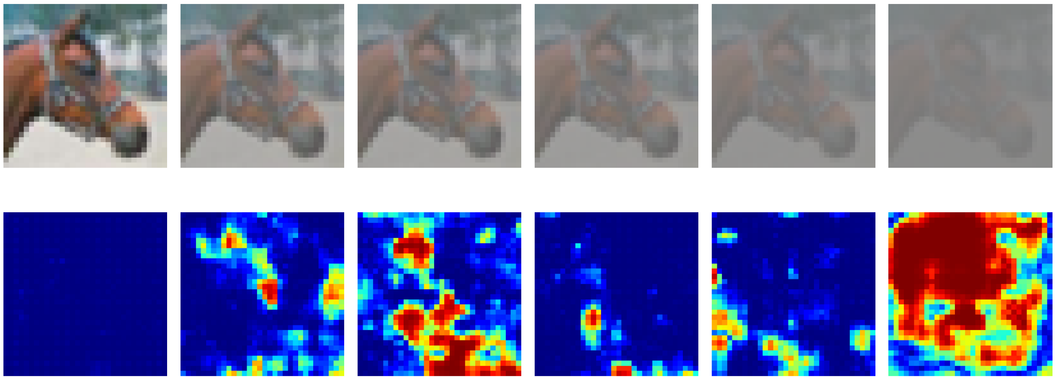
(a) B=1.0 (b) B=1.5 (c) B=1.75 (d) B=2.0 (e) B=2.25 (f) B=2.5

Figure 8: Example images of a horse from the CIFAR-10 testset with corresponding heatmap prediction with different brightness values (B). (a) is the original image without brightness augmentation, whereas from (b) to (f) the brightness value is increased. Blue colors mark in-distribution regions, whereas the red/yellow colors highlight out-of-distribution regions.



(a) B=1.0 (b) B=1.5 (c) B=1.75 (d) B=2.0 (e) B=2.25 (f) B=2.5

Figure 9: Example images of a truck from the CIFAR-10 testset with corresponding heatmap predictions with different brightness values (B). (a) is the original image without brightness augmentation, whereas from (b) to (f) the brightness value is increased. Blue colors mark in-distribution regions, whereas the red/yellow colors highlight out-of-distribution regions.



(a) C=1.0 (b) C=0.5 (c) C=0.4 (d) C=0.3 (e) C=0.2 (f) C=0.1

Figure 10: Example images of a horse from the CIFAR-10 testset with corresponding heatmap predictions with different contrast values (C). (a) is the original image without contrast augmentation, whereas from (b) to (f) the contrast value is reduced. Blue colors mark in-distribution regions, whereas the red/yellow colors highlight out-of-distribution regions.

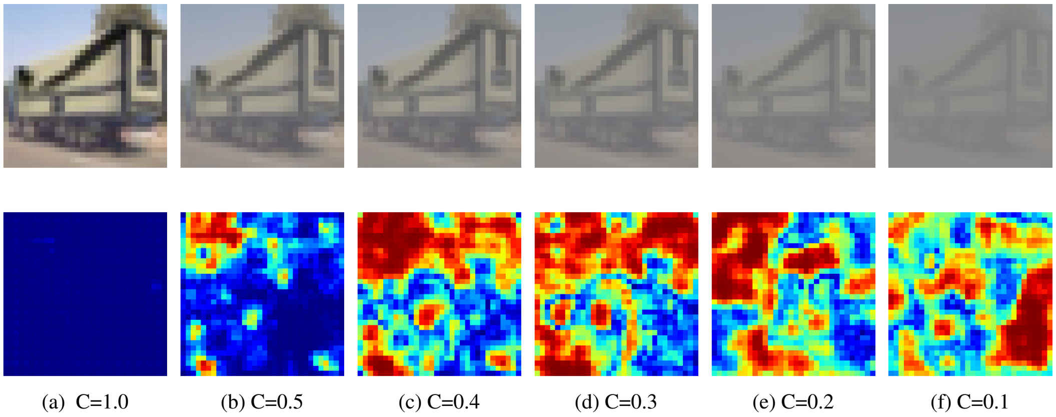


Figure 11: Example images of a truck from the CIFAR-10 testset with corresponding heatmap predictions with different contrast values (C). (a) is the original image without contrast augmentation, whereas from (b) to (f) the contrast value is reduced. Blue colors mark in-distribution regions, whereas the red/yellow colors highlight out-of-distribution regions.

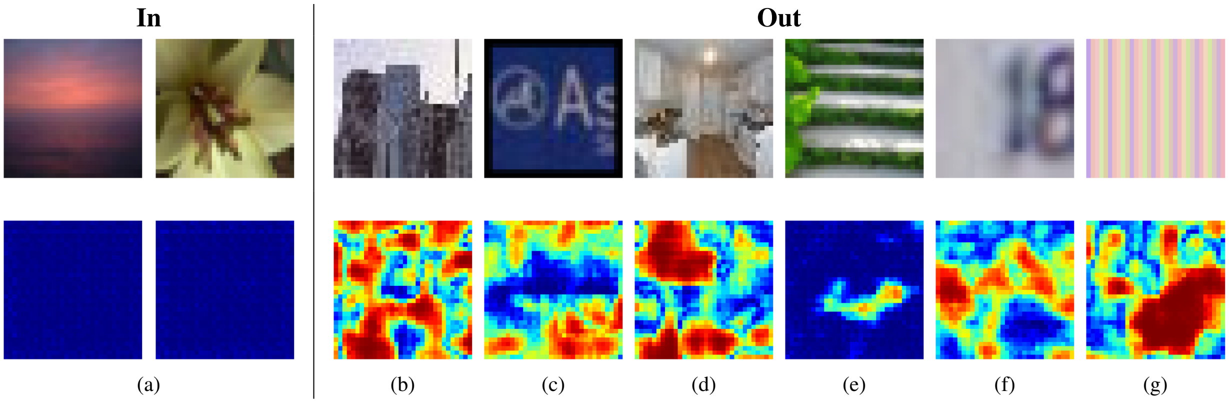


Figure 12: Visual results with WideResNet trained on CIFAR-100 (a) as an in-distribution database. Examples from the out-of-distribution databases iSUN (b), LSUN-Crop (c), LSUN-Resize (d), Places365 (e), SVHN (f) and Textures (g) are illustrated. The original images are displayed in the top row, whereas the heatmaps are in the bottom row. The in-distribution (**In**) heatmaps (a) show no response. The OOD (**Out**) heatmaps (b) - (g) highlight the difference of the images to the training distribution. Blue colors mark in-distribution regions, whereas the red/yellow colors highlight out-of-distribution regions.

Dataset	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
iSUN	MSP [1]	91.05	98.10	63.90	55.05
	ODIN [3]	94.49	98.62	83.39	23.20
	Mahalanobis [2]	93.25	98.54	74.05	38.60
	Energy [4]	93.35	98.48	74.79	33.50
	ReAct [5]	93.87	98.55	75.70	30.45
	Ours	97.72	99.55	84.70	8.24
Textures	MSP [1]	88.41	97.14	58.74	60.45
	ODIN [3]	78.74	93.37	55.35	58.65
	Mahalanobis [2]	93.31	98.43	80.45	30.60
	Energy [4]	86.51	96.24	60.90	52.15
	ReAct [5]	87.12	96.48	62.47	52.00
	Ours	95.75	99.03	82.32	22.44
SVHN	MSP [1]	91.76	98.29	62.57	56.65
	ODIN [3]	84.62	95.75	63.37	50.70
	Mahalanobis [2]	95.80	99.14	79.27	23.15
	Energy [4]	92.11	98.21	69.33	43.85
	ReAct [5]	88.96	97.34	61.80	52.00
	Ours	97.05	99.36	84.78	14.70
Places365	MSP [1]	87.23	96.75	55.54	63.80
	ODIN [3]	81.40	94.57	54.34	57.75
	Mahalanobis [2]	85.28	96.52	51.91	65.50
	Energy [4]	87.64	96.57	63.00	49.55
	ReAct [5]	87.65	96.68	63.47	48.50
	Ours	91.79	97.79	72.07	35.14
LSUN-Crop	MSP [1]	94.10	98.78	75.17	42.25
	ODIN [3]	95.27	98.78	87.48	18.15
	Mahalanobis [2]	93.34	98.59	70.80	38.05
	Energy [4]	96.39	99.17	87.34	18.60
	ReAct [5]	97.86	99.51	92.63	9.90
	Ours	98.72	99.74	92.79	3.77
LSUN-Resize	MSP [1]	91.74	98.27	64.94	53.05
	ODIN [3]	95.43	98.91	85.02	21.65
	Mahalanobis [2]	93.03	98.54	71.34	41.20
	Energy [4]	94.33	98.74	77.40	30.20
	ReAct [5]	94.81	98.77	79.21	26.25
	Ours	97.81	99.56	85.72	7.9

Table 3: Detailed evaluation for ResNet trained with CIFAR-10. Comparison of the OOD detection performance in terms of AUROC, AUPR-Success, AUPR-Error, and FPR at 95% TPR. We compare our approach to methods that do not further optimize the classifier but operate on the pre-trained model. ↑ indicates that larger values are better, whereas ↓ marks that lower values are better.

Dataset	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
iSUN	MSP [1]	91.88	98.36	65.30	55.30
	ODIN [3]	97.67	99.51	90.32	12.35
	Mahalanobis [2]	96.47	99.15	89.36	15.90
	Energy [4]	96.47	99.28	83.80	20.15
	ReAct [5]	65.11	91.73	20.65	97.15
	Ours	98.52	99.69	90.65	4.36
Textures	MSP [1]	89.27	97.57	57.14	64.45
	ODIN [3]	89.49	96.93	71.05	41.00
	Mahalanobis [2]	92.80	98.29	80.43	31.90
	Energy [4]	90.49	97.41	68.04	43.15
	ReAct [5]	43.28	80.96	13.98	98.20
	Ours	93.89	98.52	80.59	28.45
SVHN	MSP [1]	93.22	98.71	62.92	57.35
	ODIN [3]	96.81	99.27	89.39	14.55
	Mahalanobis [2]	95.48	99.03	79.72	22.20
	Energy [4]	96.27	99.21	82.87	19.05
	ReAct [5]	21.91	73.36	10.23	99.95
	Ours	97.98	99.52	91.61	7.55
Places365	MSP [1]	86.94	96.87	53.86	67.45
	ODIN [3]	88.74	96.90	66.49	45.70
	Mahalanobis [2]	76.11	93.95	36.88	80.15
	Energy [4]	90.20	97.38	68.81	42.85
	ReAct [5]	61.18	89.32	20.16	95.05
	Ours	90.16	97.18	71.09	36.95
LSUN-Crop	MSP [1]	94.54	98.95	71.61	47.05
	ODIN [3]	99.19	99.83	96.64	3.60
	Mahalanobis [2]	94.03	98.73	75.92	32.60
	Energy [4]	98.84	99.77	94.52	5.70
	ReAct [5]	51.01	84.50	16.55	96.35
	Ours	98.84	99.76	94.24	3.80
LSUN-Resize	MSP [1]	93.04	98.60	70.01	49.00
	ODIN [3]	98.17	99.62	92.47	9.35
	Mahalanobis [2]	97.31	99.42	90.34	13.60
	Energy [4]	97.21	99.43	87.31	14.65
	ReAct [5]	69.03	92.88	23.58	96.00
	Ours	98.77	99.75	92.23	3.23

Table 4: Detailed evaluation for WideResNet trained with CIFAR-10. Comparison of the OOD detection performance in terms of AUROC, AUPR-Success, AUPR-Error, and FPR at 95% TPR. We compare our approach to methods that do not further optimize the classifier but operate on the pre-trained model. ↑ indicates that larger values are better, whereas ↓ marks that lower values are better.

Dataset	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
iSUN	MSP [1]	80.44	95.41	44.27	72.35
	ODIN [3]	90.34	97.85	66.73	47.15
	Mahalanobis	72.89	92.82	33.85	81.40
	Energy [4]	85.64	96.71	53.23	63.45
	ReAct [5]	86.01	96.84	52.01	64.60
	Ours	89.65	97.69	59.62	52.47
Textures	MSP [1]	75.87	94.01	34.33	81.80
	ODIN [3]	75.46	93.63	36.58	80.30
	Mahalanobis [2]	79.80	94.65	53.24	62.40
	Energy [4]	74.87	93.55	33.39	83.70
	ReAct [5]	80.16	95.26	39.81	79.65
	Ours	78.93	94.33	45.78	69.83
SVHN	MSP [1]	82.60	96.18	42.54	75.55
	ODIN [3]	81.45	95.89	37.81	82.40
	Mahalanobis [2]	82.14	95.95	41.28	75.80
	Energy [4]	84.83	96.82	41.28	81.70
	ReAct [5]	87.13	97.27	47.92	72.00
	Ours	92.02	98.14	67.20	42.24
Places365	MSP [1]	76.01	93.74	35.43	81.35
	ODIN [3]	76.21	93.63	35.45	81.25
	Mahalanobis [2]	64.25	90.11	22.98	91.50
	Energy [4]	76.53	93.79	35.68	82.15
	ReAct [5]	74.72	93.46	32.80	83.50
	Ours	74.97	92.28	39.40	74.48
LSUN-Crop	MSP [1]	80.97	95.62	42.67	74.15
	ODIN [3]	86.10	96.92	50.20	67.55
	Mahalanobis [2]	67.91	91.14	31.16	82.00
	Energy [4]	85.04	96.70	47.39	71.50
	ReAct [5]	91.25	97.98	69.73	42.00
	Ours	94.56	98.63	79.31	26.29
LSUN-Resize	MSP [1]	79.87	95.27	42.79	74.25
	ODIN [3]	90.13	97.81	65.68	49.10
	Mahalanobis [2]	73.79	93.33	32.91	83.65
	Energy [4]	85.51	96.66	52.58	64.20
	ReAct [5]	86.06	9684	52.22	64.95
	Ours	90.31	97.89	61.35	51.10

Table 5: Detailed evaluation for ResNet trained with CIFAR-100. Comparison of the OOD detection performance in terms of AUROC, AUPR-Success, AUPR-Error, and FPR at 95% TPR. We compare our approach to methods that do not further optimize the classifier but operate on the pre-trained model. ↑ indicates that larger values are better, whereas ↓ marks that lower values are better.

Dataset	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
iSUN	MSP [1]	58.28	87.38	20.90	92.70
	ODIN [3]	72.15	92.23	34.54	80.70
	Mahalanobis [2]	90.18	97.63	69.16	41.60
	Energy [4]	67.97	90.91	27.99	86.85
	ReAct [5]	74.71	93.89	31.17	87.60
	Ours	89.75	97.67	62.91	51.40
Textures	MSP [1]	68.97	91.08	28.28	86.80
	ODIN [3]	75.81	93.63	36.06	81.65
	Mahalanobis [2]	83.49	95.47	59.72	54.44
	Energy [4]	74.52	93.24	32.73	85.00
	ReAct [5]	78.42	94.65	40.79	77.30
	Ours	76.13	93.15	44.25	70.84
SVHN	MSP [1]	60.59	90.02	21.48	91.60
	ODIN [3]	89.59	97.88	52.44	68.50
	Mahalanobis [2]	66.06	90.59	26.56	86.00
	Energy [4]	86.22	97.17	44.28	77.35
	ReAct [5]	91.43	98.02	71.79	40.05
	Ours	93.54	98.40	75.13	30.54
Places365	MSP [1]	70.11	92.29	28.46	88.00
	ODIN [3]	74.69	93.46	33.57	83.75
	Mahalanobis [2]	59.47	87.67	20.85	92.40
	Energy [4]	74.65	93.45	33.74	83.20
	ReAct [5]	71.50	92.52	29.39	87.35
	Ours	69.55	89.65	35.77	77.34
LSUN-Crop	MSP [1]	76.67	94.54	37.60	78.90
	ODIN [3]	92.24	98.37	71.06	44.90
	Mahalanobis [2]	52.78	85.96	16.50	96.85
	Energy [4]	90.71	98.04	65.49	50.35
	ReAct [5]	93.84	98.51	83.45	26.15
	Ours	93.74	98.38	76.92	29.91
LSUN-Resize	MSP [1]	57.23	87.00	20.52	92.70
	ODIN [3]	72.08	92.01	36.19	79.65
	Mahalanobis [2]	91.94	98.18	70.05	39.30
	Energy [4]	68.61	90.88	30.16	85.40
	ReAct [5]	74.54	93.84	31.66	86.35
	Ours	93.16	98.53	71.88	39.14

Table 6: Detailed evaluation for WideResNet trained with CIFAR-100. Comparison of the OOD detection performance in terms of AUROC, AUPR-Success, AUPR-Error, and FPR at 95% TPR. We compare our approach to methods that do not further optimize the classifier but operate on the pre-trained model. ↑ indicates that larger values are better, whereas ↓ marks that lower values are better.

Dataset	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
Textures	MSP [1]	69.31	92.26	26.23	89.50
	ODIN [3]	76.08	94.14	36.34	81.80
	Mahalanobis [2]	93.77	98.50	82.76	26.50
	Energy [4]	75.66	94.03	33.81	83.80
	ReAct [5]	81.65	95.33	46.58	69.40
	Ours	76.31	93.80	35.51	81.46
iNaturalist	MSP [1]	73.33	93.65	29.18	88.70
	ODIN [3]	73.90	93.86	28.84	89.90
	Mahalanobis [2]	66.11	90.39	25.09	89.40
	Energy [4]	74.07	93.87	29.74	88.10
	ReAct [5]	88.51	97.40	59.86	55.50
	Ours	83.51	95.87	49.60	64.44
SUN	MSP [1]	73.85	93.46	31.77	85.60
	ODIN [3]	75.75	94.01	32.59	85.30
	Mahalanobis [2]	65.09	90.15	24.25	91.00
	Energy [4]	78.23	94.71	37.68	80.10
	ReAct [5]	86.43	96.76	57.13	58.40
	Ours	96.10	99.09	83.32	18.08

Table 7: Detailed evaluation for ResNet trained with Tiny ImageNet. Comparison of the OOD detection performance in terms of AUROC, AUPR-Success, AUPR-Error, and FPR at 95% TPR. We compare our approach to methods that do not further optimize the classifier but operate on the pre-trained model. ↑ indicates that larger values are better, whereas ↓ marks that lower values are better.

References

[1] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural network. In *ICLR*, 2017.

[2] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *NIPS*, volume 31. Curran Associates, Inc., 2018.

[3] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018.

[4] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *NIPS*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020.

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