Supplementary Materials

We provide additional information and results omitted in our main paper.

- **Detailed setup (Appendix A):** including definitions of other metrics (cf. subsection 4.3); descriptions of datasets (cf subsection 5.1) and OOD methods (cf subsection 5.3).

- **Additional experimental results (Appendix B):** including detailed tables and figures for subsection 6.1, subsection 6.2, subsection 6.3, and subsection 6.4.

A. Detailed Setup

A.1. Metrics

Additional metrics. In addition to the metrics defined in subsection 4.3, we report two other metrics for C-OOD examples:

- **FPR(C-OOD$-$)@TPR(ID$+$)=95**: FPR for accepting C-OOD$-$ data, at TPR$=95\%$ for accepting ID+ data.

- **TPR(C-OOD$+$)@TPR(ID$+$)=95**: TPR for accepting C-OOD$+$ data, at TPR$=95\%$ for accepting ID+ data.

Metrics in other OOD detection frameworks. To further distinguish between our MS-OOD DETECTION framework and the existing frameworks (cf. Figure 2), we describe their evaluation metrics in detail. We use the same notations described in subsection 4.3. Figure 11 contains a comprehensive visualization of the evaluation metrics of all frameworks.

- **Conventional framework (Figure 11 (a))** only considers the existence of ID and S-OOD data during testing. It evaluates the performance by FPR(S-OOD)@TPR(ID)=95: the False Positive Rate of wrongly accepting the S-OOD data when the True Positive Rate of the ID data is 95%. Existing works in OOD detection often abbreviate it as FPR95.

- **SEM framework (Figure 11 (b))** extends the conventional framework by including C-OOD data. It evaluates the performance by the same metric; *i.e.*, FPR(S-OOD) or FPR95. The difference lies in the threshold: SEM treats the C-OOD data as ID (since the model can potentially classify them correctly) and sets the threshold such that 95% of the C-OOD and ID data are accepted. In this framework, if the scoring function is robust to covariate shift (*i.e.*, the score distributions of the C-OOD and ID data are similar), the performance would be similar to the conventional framework. This is, however, not the case given that different models, covariate shifts, and detection methods can result in different score distributions between the ID and C-OOD data (see Figure 16 and Figure 17).

- **G-ODIN framework (Figure 11 (c))** also includes C-OOD data during testing. However, instead of considering C-OOD data as ID, this framework aims to reject C-OOD data similarly to S-OOD data. It reports the FPR95 metric defined in the conventional framework separately for the S-OOD and C-OOD data, and we denote them by FPR(S-OOD) and FPR(C-OOD), respectively, where each is calculated when TPR for the ID data is 95%.

- **SCOD framework (Figure 11 (d))** incorporates selective classification into the conventional framework. That is, in addition to rejecting the S-OOD data, it further considers whether the model can classify the ID data correctly. The evaluation is based on FPR(S-OOD) and FPR(ID$-$) when TPR for the *correctly classified* ID data (*i.e.*, ID$+$) is 95%. The differences to the conventional framework are two-fold: (1) rejection includes the misclassified ID data; (2) the threshold is based on the correctly classified ID data.

Metrics in open-set recognition. Open-set recognition (OSR) tackles semantic shift detection and thus is highly related to conventional OOD detection. We remark that their metrics are relevant. In OOD detection, researchers often compare acceptance of ID data vs. rejection of OOD data [50]. The open-set classification rate (OSCR) curve proposed in [1, 5] shares similar concepts, comparing the classification accuracy of ID data vs. rejection of novel-class data. The key difference is that OOD detection implicitly assumes that all accepted ID data can be correctly classified, and our metric (*i.e.*, TPR(ID$+$)) addresses it, making our metric more similar to the OSCR curve.

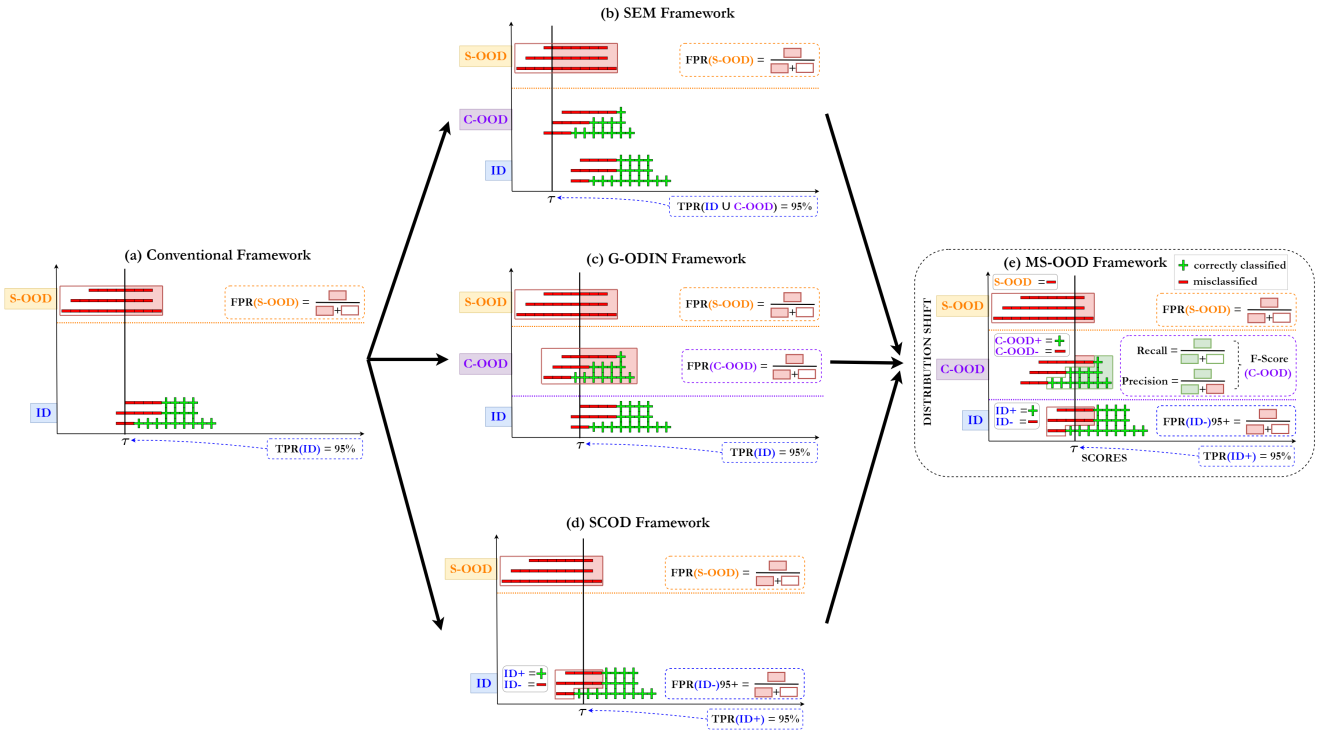

Figure 11: **Comparison of the evaluation metrics among (a) conventional, (b) SEM, (c) G-ODIN, (d) SCOD, and (e) MS-OOD DETECTION frameworks**. Please be referred to Figure 3 for the denotations and subsection 4.3 and subsection A.1 for the definitions.

A.2. Datasets

We describe the datasets used in the main paper in detail.

- **ImageNet-1k** (IN) [4] is a large-scale dataset for image recognition, containing 1,000 classes of real-world photos. It has 1,281,167 training images, 50,000 validation images, and 100,000 test images. We use the validation set as the in-distribution (ID) test data.
- **ImageNet-V2** (IN-V2) [32] is collected similarly to ImageNet-1k with 3 test sets, each with 10,000 images. We use the test version that contains 10 images for each class with a selection frequency of at least 0.7.
- **ImageNet-R** (IN-R) [12] contains 30,000 images from 200 of the 1,000 ImageNet-1k classes. The images are of different styles from the ImageNet-1k, including but not limited to art, cartoons, DeviantArt, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, and video game renditions.
- **ImageNet-S** (IN-S) [42] consists of 50,000 sketch images; 50 images for each of the 1000 ImageNet-1k classes.
- **ImageNet-A** (IN-A) [15] has 7,500 of natural adversarial images from 200 of the 1,000 ImageNet-1k classes. It is constructed based on the ResNet50 model's predictions. Wrongly classified images with high confidence are collected.
- **Street View House Numbers** (SVHN) [28] contains 73,257 training and 26,032 test images of digits 0 to 9. The images are 32 x 32 in resolution. We use only the test images for evaluation.
- **Describable Textures Dataset** (Texture) [3] consists 5,640 textural images found in the wild with image resolution ranging from 300 x 300 to 640 x 640. It is split into training, validation and testing set with equal sizes. We use the training set for the evaluation.
- **Places** [53] features more than 10 million images with more than 400 scene categories. We use the curated version from [19] with 50 classes outside ImageNet-1k classes and randomly sampled 10,000 images.
- **iNaturalist** (iNat) [40] consists of 859,000 images of more than 5,000 fine-grained species of animals and plants. We follow the setting from [19] with 110 classes not found in ImageNet-1k classes and randomly sample 10,000 images.

- **SUN** [48] has 130,159 images of 397 scenes with image resolution higher than 200 x 200. We also follow the same setting in [19] in this dataset, by using only 50 classes outside ImageNet-1k classes randomly sampling 10,000 images.

- **ImageNet-O** (IN-O) [15] contains 7,500 natural adversarial images with classes outside a subset of 200 classes in ImageNet-1k. It is collected similarly to ImageNet-A.

- **Semantic Shift Benchmark** (SSB) [41] is curated from ImageNet-21k [33] by calculating the semantic distance using ImageNet-1k tree-like classes hierarchy. It is divided into 'Easy' (SSB-E) and 'Hard' (SSB-H) splits based on whether the classes are far way (*i.e. dog* with *candle*) or close to ImageNet-1k classes (*i.e. dog* with *wolf*). Each split has 1,000 classes and 50,000 images.

A.3. Detection methods

Following the same notations introduced in subsection 4.2, we briefly provide the mathematical definitions of the OOD detection algorithms used in this paper.

Maximum Softmax Probabilities (MSP) [13] uses the softmax output of a classifier as the scoring function. Let us denote the model's label space by $\mathcal{S} = \{1, 2, \cdots, C\}$, the model's output logit for class $c \in \mathcal{S}$ by $f_c(x)$, and the training data by $D_{\text{tr}} = \{x_1, x_2, \cdots, x_N\}$. The scoring function for MSP is:

$$g_{\text{MSP}}(x, f) = \max_{c \in \mathcal{S}} \frac{e^{f_c(x)}}{\sum_{c'=1}^{C} e^{f_{c'}(x)}}. \tag{2}$$

Maximum Logit Score (MLS) [41] uses only the logits of the classifier:

$$g_{\text{MLS}}(x, f) = \max_{c \in \mathcal{S}} f_c(x). \tag{3}$$

Energy [23] expresses the score by the denominator defined in Equation 2:

$$g_{\text{Energy}}(x, f) = T \cdot \log \sum_{c=1}^{C} e^{f_c(x)/T}, \tag{4}$$

where T denotes the temperature. We follow [23] to set $T = 1$ in the main paper.

Virtual-Logit Matching (ViM) [43] introduces the Residual $\text{res}(x)$ to the energy-based method defined in Equation 4:

$$g_{\text{ViM}}(x, f) = \log \sum_{c=1}^{C} e^{f_c(x)} - \alpha \cdot \text{res}(x). \tag{5}$$

The scaling parameter α can either be treated as a hyperparameter or computed using the formula:

$$\alpha = \frac{\sum_{i=1}^{N} \max_{c \in \mathcal{S}} f_c(x_i)}{\sum_{i=1}^{N} \text{res}(x_i)}. \tag{6}$$

We follow the same setting in [43], using 200,000 uniformly sampled ImageNet training images to compute α. Please be referred to [43] for the detailed derivation of $\text{res}(x_i)$.

Figure 12: **Top 3 image examples for each ID and C-OOD datasets across different detection algorithms using ResNet50.** Red and green border denote *misclassified* and *correctly classified* images consecutively. The row corresponds to detection algorithms (from top to bottom: MSP, MLS, Energy, ViM, GradNorm) and the column corresponds to datasets(from left to right for every three columns: ImageNet, ImageNet-V2, ImageNet-S, ImageNet-R, ImageNet-A).

Table 3: The accuracy (ACC) to classify C-OOD data using different models, and the false positive rate (FPR) to reject C-OOD data using maximum softmax probabilities (MSP).

MODEL	IN-V2		IN-S		IN-R		IN-A	
	ACC↑	FPR↓ (C-OOD)	ACC↑	FPR↓ (C-OOD)	ACC↑	FPR↓ (C-OOD)	ACC↑	FPR↓ (C-OOD)
ResNet18	66.5	94.4	20.2	68.9	33.1	84.5	1.1	87.4
ResNet50	72.4	93.9	24.1	65.7	36.2	71.6	0.0	81.5
ResNet152	75.1	93.7	28.5	66.5	41.3	70.4	6.0	75.3
Robust ResNet50	77.7	93.4	29.9	68.8	42.8	63.7	14.5	74.9
ViT-B-16	77.4	94.2	29.4	60.1	44.0	54.2	20.8	60.3
CLIP-ResNet50	59.5	95.4	35.5	78.4	60.6	92.3	22.8	89.5

GradNorm [18] relies on the gradients w.r.t. the last fully-connected layer of the classifier. Let us define the cross-entropy loss as:

$$\mathcal{L}_{\text{CE}}(f(x), y) = -\log \frac{e^{f_y(x)}}{\sum_{c=1}^{C} e^{f_c(x)}}, \tag{7}$$

where y denotes the ground-truth label of x. The scoring function of GradNorm is defined as:

$$g_{\text{Grad}}(x, f) = \|\frac{1}{C}\sum_{c=1}^{C} \nabla_{\boldsymbol{w}} \mathcal{L}_{\text{CE}}(f(x), c)\|_1, \tag{8}$$

where $\boldsymbol{w}$ denotes the weights (represented as a vector) of the last fully-connected layer. This score represents the average of the derivatives of cross entropy over all classes.

B. Additional Experimental Results

B.1. Qualitative visualization

We show in Figure 13 and Figure 12 several examples of the C-OOD data and ID data. Specifically, we consider ResNet50 with different OOD detection algorithms and we sort the examples by the $g(x, f)$ scores: the higher the score is, the higher chance that it is to be accepted. We then group them into top 3 in Figure 12 and bottom 3 Figure 13 based on the scores

Figure 13: **Bottom 3 image examples for each ID and C-OOD datasets across different detection algorithms using ResNet50.** Denotations follow Figure 12.

Table 4: **Performance across six models, five detection methods, and thirteen datasets under our MS-OOD DETECTION framework.** We use FPR(ID−), F_1-score(C-OOD), and FPR(S-OOD) in percentage to evaluate each data type. The threshold is selected at TPR(ID+)=95. For each model, we also report its classification accuracies (ACC) on the ID and each of the C-OOD datasets. Red text indicates the best-performing detection method for each combination of the neural network models and datasets.

MODEL	METHOD	ID	C-OOD				S-OOD							
		FPR(ID-)↓	F-score(C-OOD)↑				FPR(S-OOD)↓							
		IN	IN-V2	IN-S	IN-R	IN-A	SVHN	Texture	Places	iNat	SUN	IN-O	SSB-E	SSB-H
ResNet18	***ACC(%)***	***69.8***	***66.5***	***20.2***	***33.1***	***1.2***				×				
	MSP	**63.4**	**83.5**	51.5	62.3	1.7	2.5	46.3	51.0	30.7	48.0	97.8	**54.5**	**71.0**
	MaxLogit	71.4	81.7	**52.5**	**64.6**	1.7	1.1	36.1	45.3	29.5	39.5	89.9	55.0	77.4
	Energy	75.1	81.1	50.6	60.6	1.4	1.5	36.1	46.2	33.7	39.8	85.8	57.9	79.6
	ViM	75.1	81.1	50.8	60.6	1.4	0.6	27.0	85.9	83.7	87.2	**64.0**	68.9	79.6
	GradNorm	87.5	79.0	42.9	54.0	**2.4**	**0.1**	**23.4**	**38.9**	**21.2**	**28.6**	85.5	65.1	81.4
ResNet50	***ACC(%)***	***76.1***	***72.4***	***24.1***	***36.2***	***0.0***				×				
	MSP	**61.7**	**86.6**	55.7	**69.8**	**0.0**	4.2	45.1	48.8	29.6	45.4	100.0	50.8	**65.2**
	MaxLogit	70.7	85.3	**58.4**	63.8	**0.0**	2.3	36.3	46.2	26.8	38.8	100.0	**50.0**	69.4
	Energy	74.9	84.9	57.0	60.4	**0.0**	2.9	37.8	48.9	32.5	41.2	99.9	53.4	72.3
	ViM	78.7	83.8	57.8	61.4	**0.0**	**0.2**	**9.1**	72.4	55.5	69.7	**79.2**	63.3	72.4
	GradNorm	89.6	82.5	48.5	57.5	**0.0**	0.9	29.3	**44.8**	**23.7**	**33.8**	94.4	66.3	76.5
ResNet152	***ACC(%)***	***78.3***	***75.1***	***28.5***	***41.3***	***6.0***				×				
	MSP	**60.9**	**87.9**	60.1	**73.7**	12.9	10.5	42.1	48.4	30.3	45.9	97.0	51.7	64.3
	MaxLogit	70.6	86.4	**61.7**	65.2	**13.0**	5.4	29.6	**40.2**	**26.9**	**34.1**	89.6	**48.4**	**64.0**
	Energy	73.5	85.9	60.9	62.6	13.0	7.2	29.5	41.7	30.7	34.7	88.1	50.6	66.0
	ViM	77.8	85.9	61.5	65.3	9.8	**0.1**	**7.7**	64.7	44.0	63.4	**61.2**	60.2	66.1
	GradNorm	91.7	84.0	51.2	60.8	12.2	2.9	30.3	53.7	35.5	41.7	89.9	72.3	77.1
Robust ResNet50	***ACC(%)***	***80.9***	***77.7***	***29.9***	***42.8***	***14.6***				×				
	MSP	**66.5**	**88.4**	**59.4**	**75.4**	**28.2**	20.8	56.2	**57.3**	40.2	**53.1**	98.8	**56.6**	**72.8**
	MaxLogit	72.2	87.8	56.2	68.6	23.5	37.0	64.3	65.5	51.5	62.2	85.0	63.8	77.3
	Energy	94.4	85.6	45.2	20.7	7.2	100.0	95.6	97.2	98.0	97.6	**29.4**	95.5	95.2
	ViM	85.1	86.5	59.3	63.3	20.6	**0.0**	**18.7**	72.3	**25.5**	67.3	57.6	67.4	94.6
	GradNorm	99.2	85.0	45.8	59.8	25.4	100.0	97.8	99.8	100.0	99.8	98.8	99.6	99.1
ViT-B-16	***ACC(%)***	***81.1***	***77.4***	***29.4***	***44.0***	***20.8***				×				
	MSP	**59.4**	**88.9**	63.7	**78.9**	**38.8**	10.1	41.3	**49.7**	32.0	**48.1**	91.3	**53.5**	**69.8**
	MaxLogit	63.9	88.3	**64.0**	69.8	29.8	8.2	38.1	52.5	33.0	49.9	72.6	54.1	72.3
	Energy	75.8	87.0	62.3	61.1	25.5	13.5	46.0	65.2	50.9	62.8	62.2	64.0	81.3
	ViM	76.3	87.0	60.1	59.6	14.3	**0.8**	**31.5**	50.4	**10.7**	49.1	**61.8**	58.7	80.4
	GradNorm	98.1	84.6	46.9	61.3	34.7	96.4	93.1	96.9	95.9	97.1	98.8	98.0	97.1
CLIP ResNet50	***ACC(%)***	***59.8***	***59.5***	***35.5***	***60.6***	***22.8***				×				
	MSP	**72.1**	**77.5**	**63.1**	**80.3**	**41.3**	4.2	**46.6**	**54.6**	**39.3**	**48.7**	85.0	**60.3**	**79.2**
	MaxLogit	86.6	74.5	53.1	76.0	36.4	29.3	82.6	59.1	63.8	53.5	67.4	68.3	80.0
	Energy	90.7	73.9	52.3	73.6	34.9	94.2	92.3	72.8	84.6	72.2	**67.1**	76.7	83.4
	ViM	90.7	73.8	52.3	73.3	34.8	93.4	92.7	74.3	86.3	74.1	68.1	77.0	83.7
	GradNorm	93.9	73.4	51.5	74.9	37.5	**0.0**	46.7	73.2	75.9	58.4	88.5	83.0	93.3

and further indicate if they are correctly classified (*i.e.*, ID+ or C-OOD+) or not (*i.e.*, ID− or C-OOD−). We find that across algorithms the most likely rejected and accepted images are consistent. Furthermore, examples with lower $g(x, f)$ scores Figure 13 are visually much more different (*i.e.* mostly black or white images on ImageNet-S and ImageNet-R),

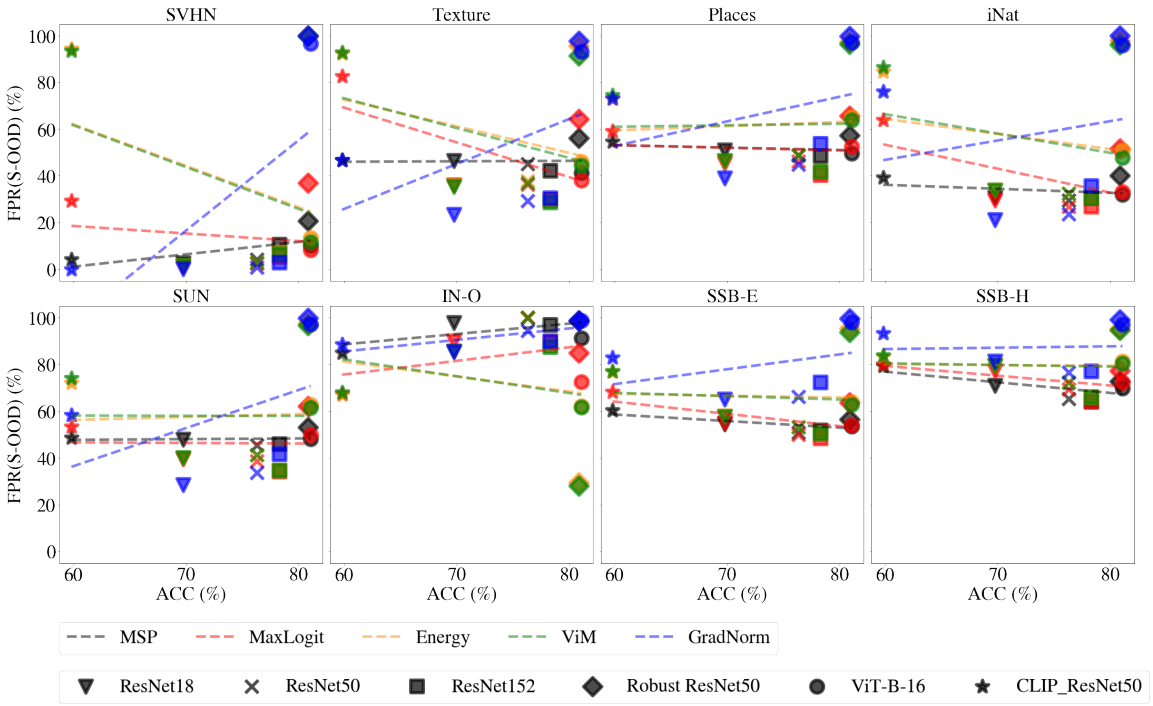

Figure 14: **S-OOD in MS-OOD DETECTION for each individual dataset.** X-axis: ACC of classifying the ID examples; Y-axis: FPR(S-OOD)@TPR(ID+)=95 for wrongly accepting S-OOD examples. Please be referred to Table 4 for detailed values.

suggesting why they are mostly classified incorrectly.

B.2. Additional results for subsection 6.1

We provide the full table of Table 1 in the main paper in Table 3. Specifically, we include other neural network models beyond ResNet50 and CLIP-ResNet50. We see that except for ResNet50 on ImageNet-A, none of the other combinations of models and datasets have zero accuracies in classifying the C-OOD data. Except for the CLIP-ResNet50 which is not fine-tuned on the ImageNet data, we see that a higher ACC often leads to a lower FPR. With that being said, we argue that it may not be ideal to reject all the C-OOD examples because some of them could indeed be correctly classified.

B.3. Tables and additional scatter plots for subsection 6.2, subsection 6.3, and subsection 6.4

We provide in Table 4 the full table used to generate Figure 4, Figure 6, and Figure 8 in the main paper. Specifically, we provide the results for each of the S-OOD datasets. It is worth noting that when we use TPR(ID+)=95 to select the threshold, MSP performs quite well in rejecting S-OOD data. In some datasets or when paired with some neural network models, it can even achieve the best performance (*i.e.*, lowest FPR) compared to other detection methods. Such a superior performance degrades when we use TPR(ID)=95 to select the threshold, as will be discussed in subsection B.4 and shown in Table 5.

We further provide the scatter plots similar to Figure 8 in the main text, but now for each of the eight S-OOD datasets separately in Figure 14. The trends generally follow what we described in subsection 6.4. For instance, robust ResNet50, though achieving a higher ID accuracy, performs quite poorly in rejecting S-OOD data. GradNorm degrades (consistently across datasets) when the ID accuracy increases.

B.4. Additional comparisons for subsection 6.4

We provide in Table 5 the detailed results comparing **FPR(S-OOD)@TPR(ID+)=95** and **FPR(S-OOD)@TPR(ID)=95**. The former is the metric used in MS-OOD DETECTION for S-OOD data; the latter is the metric used in the conventional OOD detection framework. The main difference lies in whether we consider accepting wrongly classified ID data (please note that ID=ID+ $\cup$ ID−); please see subsection 6.4 for some further discussions. For the latter (*i.e.*, **FPR(S-OOD)@TPR(ID)=95**), we also provide the scatter plots similar to Figure 8 in Figure 15.

Overall, we have three key observations. First, the trends in Figure 15 generally follow those in Figure 14. Second, as shown in Table 5, using **FPR(S-OOD)@TPR(ID+)=95** leads to better performance (lower FPR) in rejecting S-OOD data. Third, the

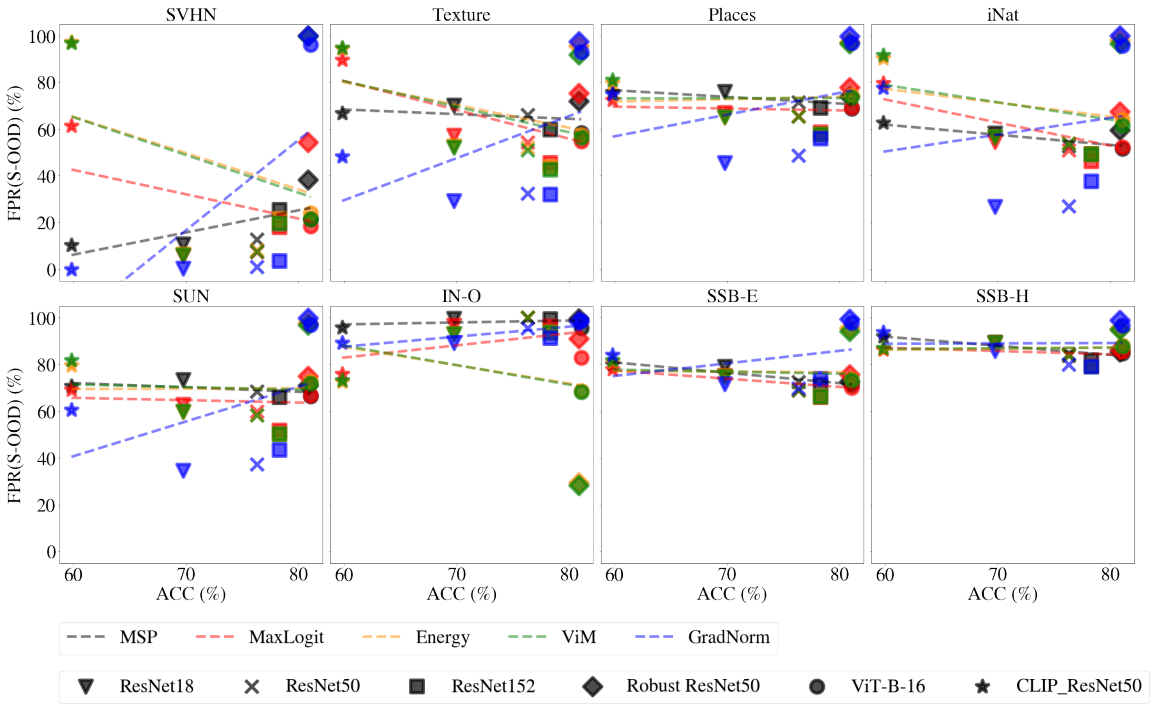

Figure 15: **S-OOD in conventional OOD detection framework for each individual dataset.** X-axis: ACC of classifying the ID examples; Y-axis: FPR(S-OOD)@TPR(ID)=95 for wrongly accepting S-OOD examples. Please be referred to Table 5 for detailed values.

Table 5: **S-OOD performance comparison between conventional framework and MS-OOD DETECTION**. Conventional framework uses TPR(ID)=95 (left column) while MS-OOD DETECTION uses TPR(ID+)=95 (right column). Red indicates the best value for a particular dataset and framework.

MODEL	METHOD	FPR(S-OOD)↓ TPR(ID) vs TPR(ID+)															
		SVHN		Texture		Places		iNat		SUN		IN-O		SSB-E		SSB-H	
ResNet18	MSP	10.7	2.5	70.2	46.3	75.9	51.0	58.2	30.7	73.5	48.0	99.7	97.8	79.0	**54.5**	89.38	**71.0**
	MaxLogit	5.8	1.1	57.2	36.1	66.9	45.3	54.5	29.5	62.9	39.5	96.8	89.9	75.4	55.0	89.24	77.4
	Energy	6.5	1.5	52.8	36.1	65.0	46.2	56.5	33.7	59.8	39.8	93.2	85.8	75.0	57.9	89.26	79.6
	ViM	1.6	0.6	38.8	27.0	92.5	85.9	91.2	83.7	93.1	87.2	**72.7**	**64.0**	80.1	68.9	89.26	79.6
	GradNorm	**0.3**	**0.1**	**29.3**	**23.4**	**45.5**	**38.9**	**26.8**	**21.2**	**34.7**	**28.6**	89.3	85.5	**71.7**	65.1	**85.92**	81.4
ResNet50	MSP	12.9	4.2	66.0	45.1	71.6	48.8	52.8	29.6	68.6	45.4	100.0	100.0	72.6	50.8	84.53	**65.2**
	MaxLogit	7.5	2.3	54.4	36.3	65.7	46.2	50.9	26.8	59.9	38.8	100.0	100.0	**69.0**	**50.0**	83.76	69.4
	Energy	8.2	2.9	52.1	37.8	65.4	48.9	54.0	32.5	58.3	41.2	100.0	99.9	69.2	53.4	83.87	72.3
	ViM	**0.8**	**0.2**	**15.7**	**9.1**	83.5	72.4	71.8	55.5	82.1	69.7	**84.9**	**79.2**	76.2	63.3	83.92	72.4
	GradNorm	1.1	0.9	32.4	29.3	**48.7**	**44.8**	**27.0**	**23.7**	**37.3**	**33.8**	95.6	94.4	69.8	66.3	**80.06**	76.5
ResNet152	MSP	25.5	10.5	59.8	42.1	68.9	48.4	49.3	30.3	66.0	45.9	99.3	97.0	71.3	51.7	81.88	64.3
	MaxLogit	18.0	5.4	45.6	29.6	58.7	**40.2**	46.3	**26.9**	51.9	**34.1**	96.2	89.6	**66.1**	**48.4**	79.4	**64.0**
	Energy	21.5	7.2	43.8	29.5	57.7	41.7	49.4	30.7	50.3	34.7	93.7	88.1	66.4	50.6	**79.01**	66.0
	ViM	**0.3**	**0.1**	**13.0**	**7.7**	78.6	64.7	63.1	44.0	77.8	63.4	**71.5**	**61.2**	73.5	60.2	79.2	66.1
	GradNorm	3.5	2.9	31.9	30.3	**55.9**	53.7	**37.5**	35.5	**43.5**	41.7	91.3	89.9	74.0	72.3	79.12	77.1
Robust ResNet50	MSP	38.4	20.8	71.9	56.2	**74.1**	**57.3**	59.5	40.2	**70.9**	**53.1**	99.6	98.8	**72.8**	**56.6**	**85.44**	**72.8**
	MaxLogit	54.3	37.0	75.4	64.3	77.8	65.5	67.4	51.5	75.0	62.2	91.0	85.0	75.7	63.8	86.48	77.3
	Energy	100.0	100.0	95.7	95.6	97.3	97.2	98.1	98.0	97.7	97.6	**29.5**	**29.4**	95.6	95.5	95.28	95.2
	ViM	**0.1**	**0.0**	**21.9**	**18.7**	77.6	72.3	**31.4**	**25.5**	73.3	67.3	64.7	57.6	73.9	67.4	94.97	94.6
	GradNorm	100.0	100.0	97.4	97.8	99.6	99.8	100.0	100.0	99.8	99.8	98.6	98.8	99.5	99.6	98.89	99.1
ViT-B-16	MSP	21.5	10.1	58.3	41.3	68.7	**49.7**	51.5	32.0	66.6	**48.1**	95.7	91.3	71.3	**53.5**	**84.73**	**69.8**
	MaxLogit	18.5	8.2	54.8	38.1	69.1	52.5	52.3	33.0	66.9	49.9	82.9	72.6	**70.2**	54.1	85.64	72.3
	Energy	24.1	13.5	57.4	46.0	74.3	65.2	64.1	50.9	72.8	62.8	**68.7**	62.2	73.8	64.0	88.16	81.3
	ViM	**2.3**	**0.8**	**43.9**	**31.5**	**61.1**	50.4	**17.8**	**10.7**	**59.5**	49.1	72.9	**61.8**	71.6	58.7	87.73	80.4
	GradNorm	96.1	96.4	92.9	93.1	96.7	96.9	95.6	95.9	96.9	97.1	98.7	98.8	97.9	98.0	96.77	97.1
CLIP ResNet50	MSP	10.3	4.2	66.8	**46.6**	75.9	**54.6**	**62.7**	**39.3**	70.8	**48.7**	95.8	85.0	80.1	**60.3**	91.87	**79.2**
	MaxLogit	61.2	29.3	89.5	82.6	**72.3**	59.1	79.5	63.8	69.1	53.5	75.9	67.4	**77.9**	68.3	**86.25**	80.0
	Energy	97.1	94.2	94.4	92.3	79.5	72.8	90.1	84.6	79.6	72.2	**72.7**	**67.1**	81.2	76.7	86.59	83.4
	ViM	96.7	93.4	94.7	92.7	80.9	74.3	91.6	86.3	81.8	74.1	73.3	68.1	81.7	77.0	87.03	83.7
	GradNorm	**0.0**	**0.0**	**48.3**	46.7	75.0	73.2	77.5	75.9	**60.6**	58.4	89.2	88.5	84.0	83.0	93.75	93.3

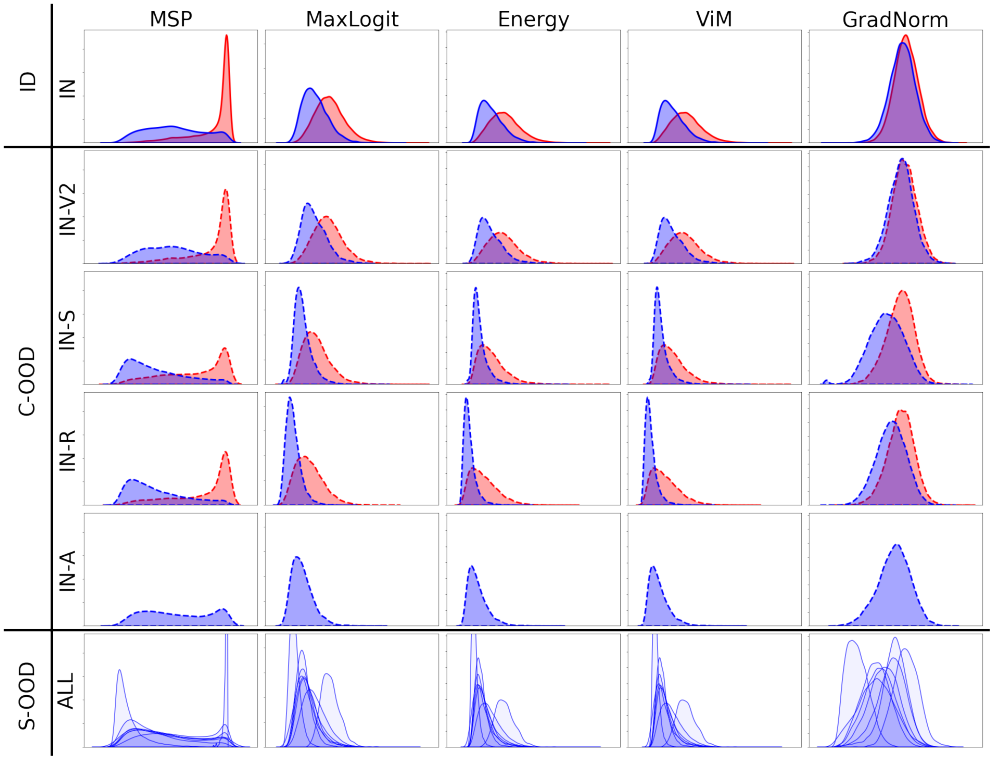

Figure 16: **Histogram of ID+, ID−, C-OOD+, C-OOD− and S-OOD across all datasets at different** $g(x, f)$**.** We use ResNet50 as the model f. Red and Blue region denote *correctly classified* and *misclassified* data respectively. The spike distribution on MSP and S-OOD pair comes from ImageNet-O [15] which is collected specifically to fool a ResNet50 model using MSP.

baseline MSP improves the most after switching the metric from the conventional one (*i.e.*, **FPR(S-OOD)@TPR(ID)=95**) to the new one (*i.e.*, **FPR(S-OOD)@TPR(ID+)=95**). In some datasets or when paired with some neural network models, MSP can even achieve the best performance (*i.e.*, lowest **FPR(S-OOD)@TPR(ID+)=95**) compared to other detection methods. In other words, the poor performance of MSP in the conventional metric mainly results from the need to accommodate the misclassified ID (*i.e.*, ID−) data: these examples normally have much lower MSP scores; to accept them requires a lower threshold τ, hence increasing the number of wrongly accepted S-OOD examples (see subsection B.5 for some further illustrations). In our metric, a detection method does not need to accept ID− examples but rejects them, allowing the use of a higher threshold τ and benefiting MSP the most.

B.5. Additional histogram plots for subsection 6.2, subsection 6.3, and subsection 6.4

We provide additional histogram plots in Figure 16 and Figure 17 using ResNet50 and CLIP-ResNet50, respectively. For each neural network model, we draw the $g(x, f)$ histogram for each detection method on each data type (ID, C-OOD, and S-OOD) and dataset — we normalize separately for ID+, ID−, C-OOD+, C-OOD−, and S-OOD examples to better showcase their distribution differences. We use red color to denote the acceptance cases (*i.e.*, ID+ and C-OOD+); blue color to denote the rejection cases (*i.e.*, ID−, C-OOD−, and S-OOD). As shown, different combinations of datasets, detection methods, and neural network models have quite different histograms. It is worth noting that for ID data (first row in each figure), MSP has the best distinction between ID+ and ID− data: the ID− examples have much lower scores (a long tail to the left). We observe a similar trend on C-OOD: MSP can best distinguish C-OOD+ and C-OOD-, with C-OOD- having lower scores.

B.6. Additional metrics for subsection 6.3

We consider additional metrics (*i.e.*, FPR(C-OOD−) and TPR(C-OOD+)) for C-OOD data in Table 6, using the threshold selected at TPR(ID+)=95. Generally speaking, MSP achieves the best on both ends; *i.e.*, low FPR and high TPR. We also observe that other methods besides MSP have impressive results on FPR(C-OOD−) (*i.e.* rejecting misclassified C-OOD) but come at the cost of also rejecting most correctly classified C-OOD (*i.e.* the TPR(C-OOD+) is lower). These results further illustrate why 1) using TPR(ID)=95, MSP performs poorly on FPR(S-OOD) as it needs a pretty low threshold; 2) using

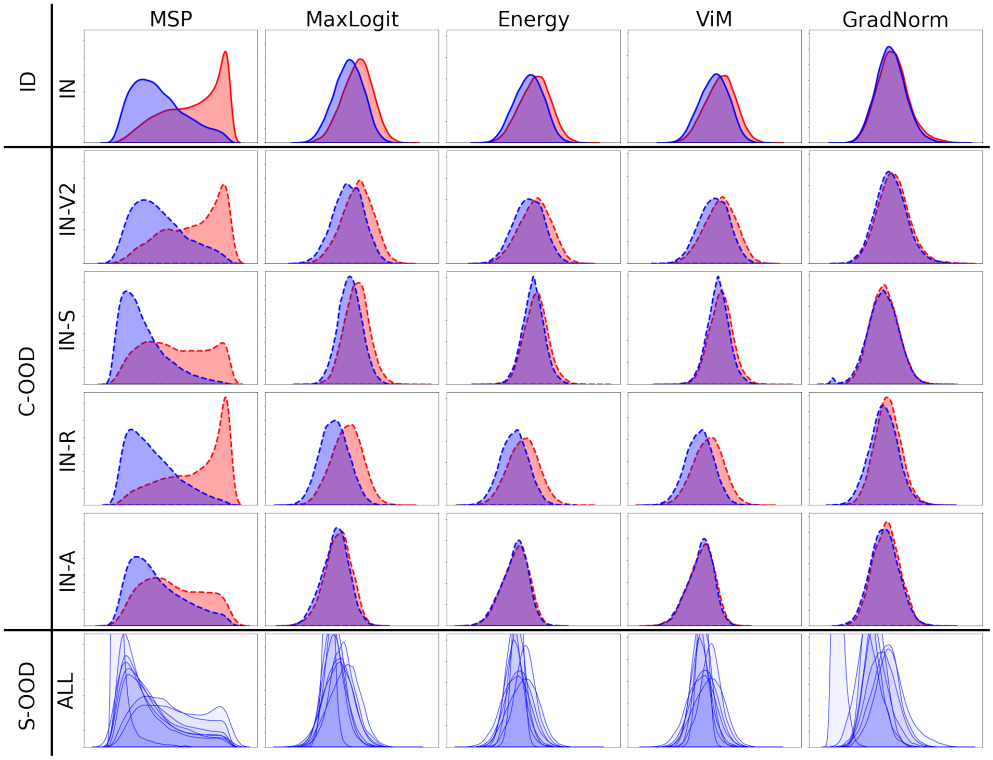

Figure 17: **Histogram of ID+, ID−, C-OOD+, C-OOD− and S-OOD across all datasets at different** $g(x, f)$**.** We use CLIP-ResNet50 as the model f. Red and Blue region denote *correctly classified* and *misclassified* data respectively.

Table 6: **Extension of Table 4 on C-OOD datasets** including FPR(C-OOD−) and TPR(C-OOD+) defined in subsection A.1.

MODEL	METHOD	FPR(C-OOD)↓				TPR(C-OOD)↑			
		IN-V2	IN-S	IN-R	IN-A	IN-V2	IN-S	IN-R	IN-A
ResNet18	***ACC(%)***	***66.5***	***20.2***	***33.1***	***1.2***	***66.5***	***20.2***	***33.1***	***1.2***
	MSP	**62.3**	34.2	48.2	61.9	**94.1**	81.3	**89.4**	46.5
	MaxLogit	70.8	**28.9**	9.2	26.9	93.7	76.2	56.7	20.9
	Energy	74.2	31.0	7.0	23.0	93.7	75.2	49.7	15.1
	ViM	74.1	30.6	**6.9**	**22.9**	93.7	75.0	49.5	15.1
	GradNorm	85.5	51.1	67.8	63.2	93.5	**82.2**	87.8	**68.6**
ResNet50	***ACC(%)***	***72.4***	***24.1***	***36.2***	***0.0***	***72.4***	***24.1***	***36.2***	***0.0***
	MSP	**60.5**	34.1	34.2	60.6	**94.1**	80.1	85.8	**0.0**
	MaxLogit	67.9	24.6	5.0	27.5	93.7	73.2	51.0	**0.0**
	Energy	71.7	25.9	4.2	**24.7**	93.9	72.4	46.4	**0.0**
	ViM	79.7	**17.8**	**4.2**	28.5	94.1	63.5	47.5	**0.0**
	GradNorm	88.5	51.8	64.1	70.9	93.8	**84.3**	**86.0**	**0.0**
ResNet152	***ACC(%)***	***75.1***	***28.5***	***41.3***	***6.0***	***75.1***	***28.5***	***41.3***	***6.0***
	MSP	**59.6**	35.7	33.6	53.9	94.0	81.4	86.2	65.0
	MaxLogit	68.9	26.7	4.9	20.9	93.5	74.4	51.8	29.6
	Energy	72.0	27.7	**4.3**	**19.1**	93.2	74.1	48.4	27.7
	ViM	77.3	**19.3**	4.4	23.2	**94.6**	65.9	51.4	23.7
	GradNorm	90.8	61.8	76.4	80.6	94.2	**87.6**	**90.9**	**88.3**
Robust ResNet50	***ACC(%)***	***77.7***	***29.9***	***42.8***	***14.6***	***77.7***	***29.9***	***42.8***	***14.6***
	MSP	**64.2**	40.0	30.3	57.5	93.7	81.8	85.0	72.0
	MaxLogit	69.5	48.9	6.3	19.8	93.9	83.9	56.6	28.8
	Energy	94.0	95.4	**0.8**	**3.2**	**95.0**	94.6	11.6	4.4
	ViM	85.3	**39.1**	6.3	22.4	94.9	80.8	50.2	26.6
	GradNorm	99.2	100.0	100.0	99.9	94.9	**99.3**	**99.7**	**100.0**
ViT-B-16	***ACC(%)***	***77.4***	***29.4***	***44.0***	***20.8***	***77.4***	***29.4***	***44.0***	***20.8***
	MSP	**60.6**	30.4	19.0	40.3	94.3	80.8	81.0	60.8
	MaxLogit	64.4	29.2	4.3	13.7	94.0	80.2	56.6	26.7
	Energy	75.2	35.3	**2.9**	9.3	94.0	83.7	45.6	19.8
	ViM	76.2	**23.6**	2.9	**9.2**	94.2	67.3	44.0	10.4
	GradNorm	98.5	92.7	94.5	96.9	**94.4**	**98.9**	**97.3**	**98.2**
CLIP ResNet50	***ACC(%)***	***59.5***	***35.5***	***60.6***	***22.8***	***59.5***	***35.5***	***60.6***	***22.8***
	MSP	**73.0**	**45.1**	61.7	67.4	94.8	83.9	**93.9**	85.5
	MaxLogit	86.7	94.3	**57.3**	**64.4**	94.4	98.2	84.1	70.8
	Energy	90.3	98.8	65.4	69.9	94.6	**99.0**	83.0	71.2
	ViM	90.5	98.5	65.6	71.2	94.5	98.8	82.6	72.0
	GradNorm	93.6	84.2	85.9	87.7	**95.0**	87.6	93.4	**91.7**

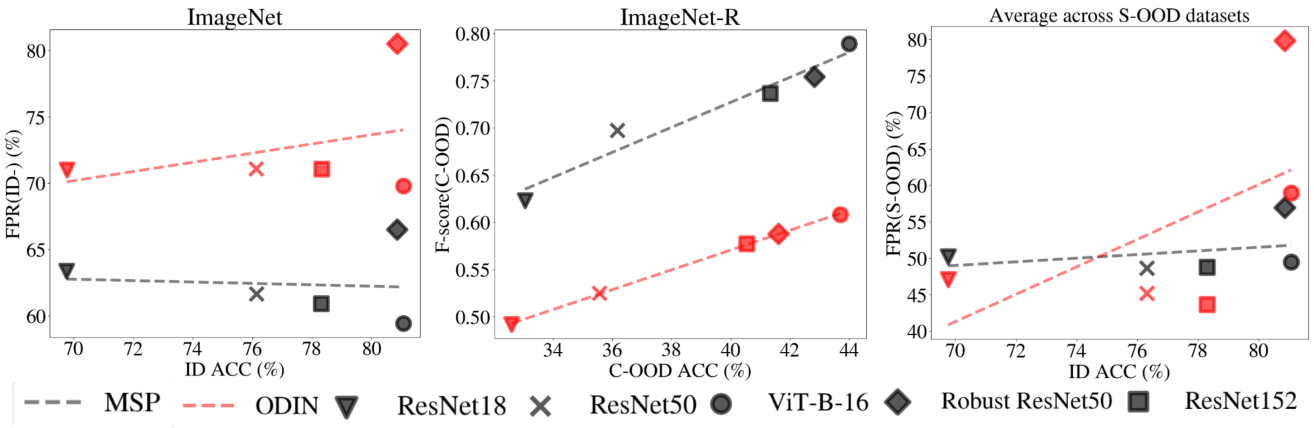

Figure 18: **ODIN performance in MS-OOD DETECTION**. We use ImageNet-R for C-OOD and averaged over S-OOD datasets for S-OOD. Denotations follow Figure 4, Figure 6, and Figure 8

TPR(ID+)=95, MSP can choose a much larger threshold to obtain a much better FPR(S-OOD).

B.7. Additional OOD methods for subsection 6.2, subsection 6.3, and subsection 6.4

We include ODIN in our experiment and set the hyperparameter the same as in [22]. We summarize our results in Figure 18, in which we conduct the same experiments as in Figure 4, Figure 6 (ImageNet-R), and Figure 8. In general, we see similar trends as MSP: on ID− (left) and S-OOD (right), the higher the ID+ accuracy is, the lower the FPR is, except for robust ResNet. For C-OOD, the higher the C-OOD+ accuracy is, the higher the F-score (C-OOD).