T2FNorm: Train-time Feature Normalization for OOD Detection in Image Classification

Supplementary Material

9. Optimality of L2 normalization

In addition to L2 normalization, we investigated various other normalization types, including L1, L3, and L4. While these alternative forms of normalization also enhance performance, L2 emerges as the most effective. As demonstrated in Table 6, all the investigated normalizations outperform the baseline, underscoring the efficacy of normalization in general. Though the separability factor in feature space for L1 normalization is higher, the FPR@95 for L2 normalization is still superior along with a higher separability factor in logit space.

Table 6. L2 normalization proves to be the optimal form of normalization for T2FNorm

L_p	FPR@95%	AUROC	AUPR	$\mathcal{S}$ (Logit)	$\mathcal{S}$ (Feature)
$p=1$	24.0	95.8	95.9	3.5	**8.5**
$p=2$	**19.7**	**96.5**	**96.4**	**6.01**	5.8
$p=3$	20.8	96.4	96.3	5.0	4.7
$p=4$	20.1	96.4	96.3	5.3	4.6

10. Adverse effect of OOD scoring-time normalization

As previously discussed, the utilization of OOD scoring-time normalization gives rise to a detrimental consequence primarily due to the obfuscation of the inherent distinction between ID and OOD samples in terms of their magnitudes. Using the ResNet-18 model trained with CIFAR-10, the results on various OOD datasets are more clearly summarized in Table 7 with three OOD metrics.

11. Feature norm Penalty

Suppressing the norm, which is directly linked to overconfidence [41], can also be feasible by employing the L2 norm of feature as the additional regularization loss. Using the joint optimization of $L_S + \lambda L_{FP}$ (L_S referring to supervised loss, L_{FP} referring feature norm penalty loss), setting $\lambda = 0.01$ to not affect the accuracy, we indeed find little improvement over baseline in FPR@95 metric only. The optimization objective seems to be satisfied with the relatively smaller average norm for both ID and OOD. However, the desired ID/OOD separability is not quite achieved as seen in Table 8.

Table 7. OOD metrics with ResNet-18 model trained in CIFAR-10 datasets for the comparison of OOD scoring-time Normalization Adaptation and OOD scoring-time Normalization Avoidance in the form of Adaptation/Avoidance. We consistently observe the superior gain in OOD detection performance thereby validating the avoidance of normalization at scoring time.

Datasets	FPR@95%	AUROC	AUPR
CIFAR-100	60.8 / **45.3**	88.9 / **91.6**	86.0 / **90.2**
TIN	56.6 / **34.4**	90.6 / **94.2**	87.6 / **93.0**
MNIST	44.1 / **03.1**	94.1 / **99.3**	98.8 / **99.9**
SVHN	50.4 / **09.0**	93.7 / **98.3**	96.4 / **99.3**
Texture	48.4 / **24.8**	92.9 / **95.6**	85.7 / **93.0**
Places365	56.5 / **32.8**	90.8 / **94.3**	96.5 / **98.2**
iSUN	48.2 / **13.9**	93.2 / **97.6**	90.6 / **97.1**
LSUN-c	27.3 / **00.9**	96.2 / **99.8**	95.7 / **99.8**
LSUN-r	45.6 / **13.3**	93.8 / **97.7**	92.1 / **97.5**

Table 8. Comparison of Baseline and Feature norm penalty method

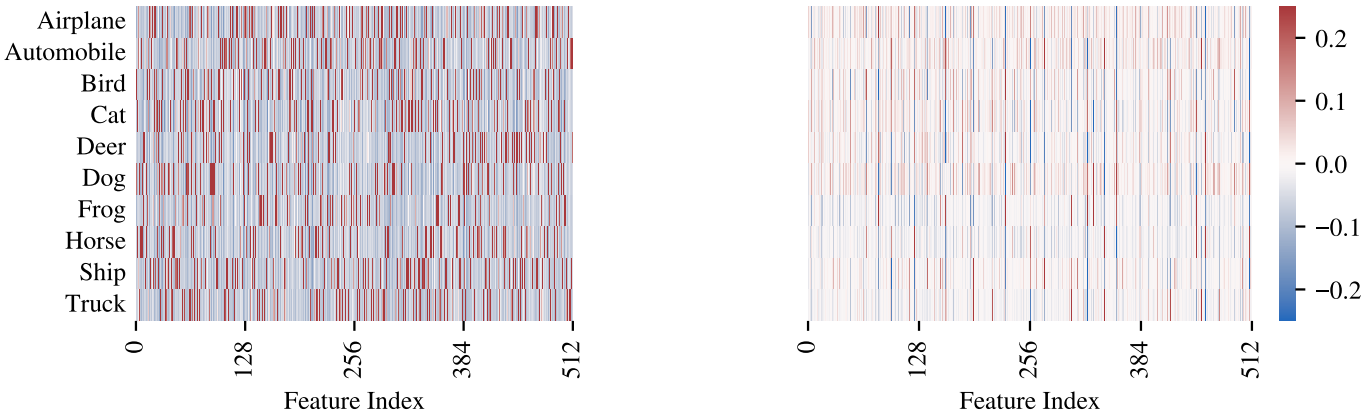
Method	FPR@95%	AUROC	AUPR	ID norm	OOD norm
Baseline	53.4	90.7	90.8	11.8	10.3
Feature norm penalty	50.9	89.1	90.6	3.0	2.8

12. Ablation study on Different Layers

The ResNet-18 architecture is primarily comprised of four residual blocks: Layer1, Layer2, Layer3, and Layer4. We conducted an ablation study to understand the impact of the application of T2FNorm on representation obtained after each of these layers and discovered that pooled representation obtained from Layer4 was the only effective way to boost OOD detection performance as shown in Table 9. This finding aligns with observations documented in [32] where it was noted high-level features are considered to have substantial potential for distinguishing between ID and OOD data as the earlier layers primarily handle low-level features, while the later layers process semantic-level features. Furthermore, as we note in Table 9, implementing normalization in the earlier layers results in a performance that is even poorer than the baseline.

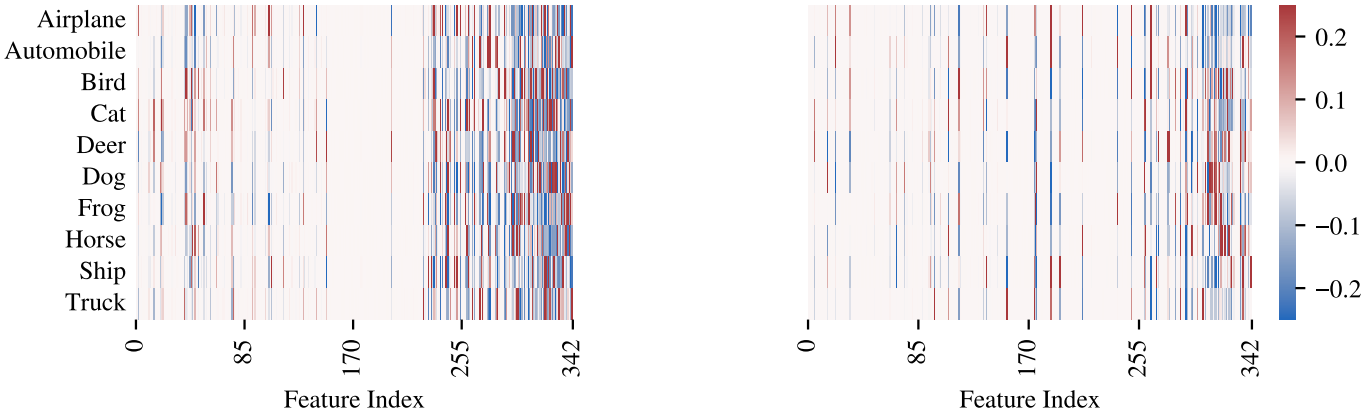
13. FC Weights Visualization

We show weight visualization of all the classes of fully connected layers in Figures 12, 13 and 14 for both T2FNorm as well as LogitNorm across various architectures. It is clearly evident that LogitNorm induces smaller weights on the FC



(a) T2FNorm (b) LogitNorm

Figure 12. FC Weights heatmap in ResNet-18



(a) T2FNorm (b) LogitNorm

Figure 13. FC Weights heatmap in DenseNet

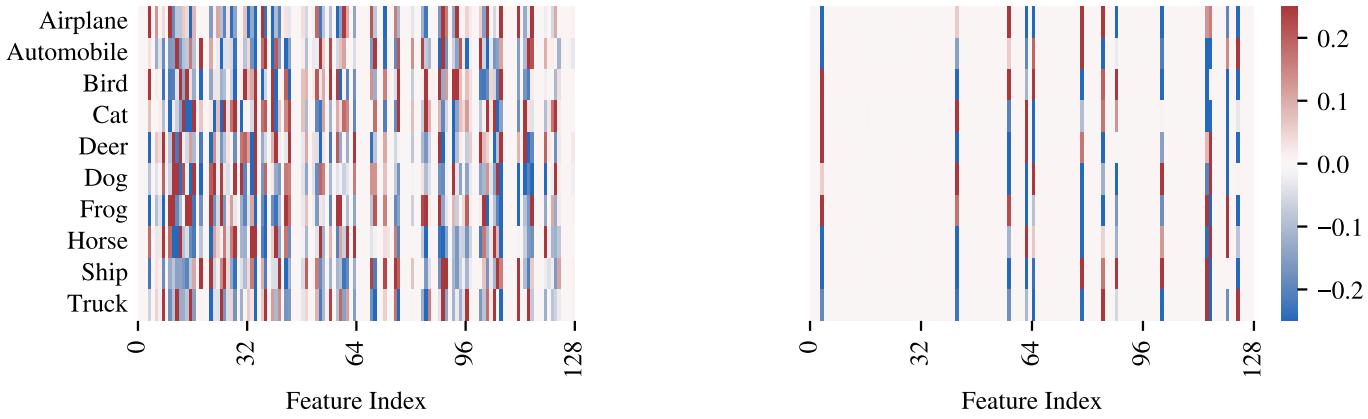
Table 9. Impact of Feature Normalization in the n^{th} Layer

Normalization in n^{th} Layer	FPR@95%	AUROC	AUPR
Layer 1	96.6	43.2	55.8
Layer 2	95.1	51.7	59.4
Layer 3	80.7	76.6	78.2
Layer 4	**19.7**	**96.5**	**96.4**
Baseline	53.43	90.74	90.83

layer in comparison to T2FNorm from the observation of all corresponding 10 classes of CIFAR-10 datasets. The observations show feature importance is sharper for our method in comparison to LogitNorm. We do not deal with the visualization of vanilla cross-entropy baseline as it doesn't address the overconfidence issue.

14. Norm and Separability ratio statistics (for CIFAR-100 as ID)

The statistics of norms in both feature space and logit space along with the separability ratio obtained from the ResNet-18 model trained in CIFAR-100 datasets with various methods are given in Table 10. The observation is similar to the earlier observation, CIFAR-10 as ID. For instance, the separability ratio achieved with our method in the penultimate feature is 2.9 with a significantly lower norm of 0.32 for OOD data in comparison to other methods. We use SVHN as OOD data for the purpose of illustrating the statistics in all settings unless otherwise noted.



(a) T2FNorm

(b) LogitNorm

Figure 14. FC Weights heatmap in WRN-40-2

Table 10. Norm of features and logits for ID and OOD samples. (ID / OOD↓ / S ratio ↑)

Method	Penultimate feature	Logit
Baseline	11.83 / 10.32 / 1.15	18.47 / 14.71 / 1.26
LogitNorm	1.63 / 1.21 / 1.37	1.03 / 0.66 / 1.57
T2FNorm	0.90 / 0.32 / **2.88**	1.31 / 0.47 / **2.80**

15. Progression of ID and OOD norm with epochs (CIFAR 10)

We show (Figure 17, 18) the progression of the average ID norm and average OOD norm in both feature and logit space with epochs during the training of ResNet-18 models. We use SVHN for OOD dataset. We can clearly observe the significant reduction in norm for both LogitNorm and T2FNorm. However, the trend in the progression of the norm in the baseline remains the opposite. Furthermore, the progression of the separability ratio with epochs is shown in Figure 16. It shows the ID/OOD separability in terms of norms in logit space too.

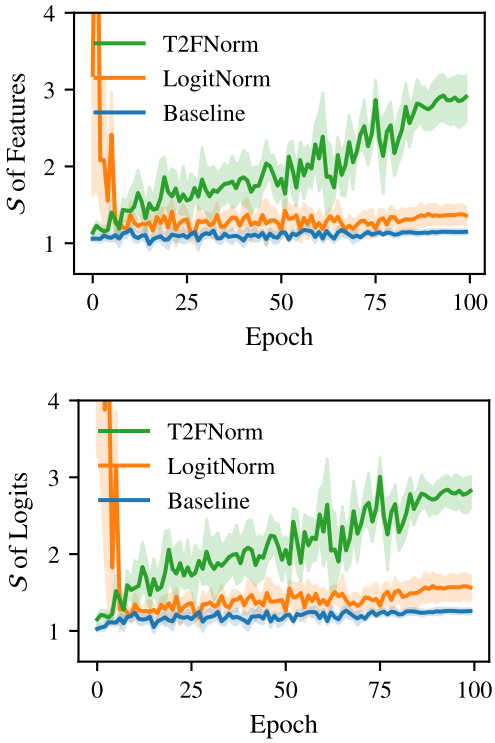


Figure 15. Progression of Separability Ratio $\mathcal{S}$ in feature and logit space over training epochs for CIFAR-100.

15.1. Datasets

15.1.1 ID datasets

CIFAR-10 and CIFAR-100 are two ID datasets used in our experiments.

CIFAR-10 CIFAR-10 is one of the most commonly used datasets for benchmarking computer vision performance, especially for classification tasks. It contains 10 categories of images.

CIFAR-100 CIFAR-100 is a very similar dataset to CIFAR-10 but consists of 100 classes.

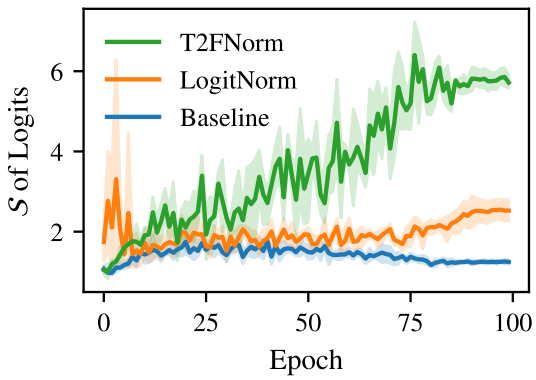


Figure 16. Progression of S at logit space with epochs

15.1.2 OOD datasets

We use a total of 9 OOD datasets: MNIST, iSUN, CIFAR (CIFAR-10 for CIFAR-100 as ID and CIFAR-100 for CIFAR-10 as ID), TinyImagenet (TIN), LSUN-R, LSUN-C, Places365, SVHN, and Texture.

MNIST The MNIST dataset comprises of 70,000 grayscale images, each representing a handwritten digit ranging from 0 to 9 in a resolution of 28x28 pixels. The dataset consists of 60,000 training images and 10,000 testing images.

SVHN SVHN is a real-world digit recognition dataset obtained from house numbers in Google Street View images. It is similar to MNIST images but the difficulty of recognition for machine learning algorithms is a bit harder.

LSUN Variations of LSUN datasets are designed for the purpose of scene understanding in large-scale datasets.

Places365 Places365 is a large-scale scene dataset developed for the purpose of training deep-learning models to understand scenes.

Texture The Textures dataset contains images of various textures. It gives a collection of unique images apart from widely available object or scene images.

TinyImageNet TinyImagenet is a smaller version of the larger ImageNet dataset. It consists of 200 classes. The TinyImagenet dataset was created to make research consisting of rich categories computationally feasible with relatively lesser computing infrastructures.

16. Distribution of Norm

The distribution of feature norm for ID (CIFAR-10) and OOD (SVHN) datasets in each of the three methods (Baseline, LogitNorm, T2FNorm) extracted from ResNet-18 architecture are shown in the Figure 19.In comparison to the baseline, both LogitNorm and T2FNorm have lesser overlap among ID/OOD samples.

17. FPR@95 across various OOD datasets

The FPR@95 metric across various architectures with both CIFAR-10 and CIFAR-100 as ID is shown in the radar plot in Figures 20, 21, 22, 23, 24, and 25.

Observations show that TrainNorm is as competitive as LogitNorm, if not better, in terms of FPR@95 metric.

18. Compatibility with various OOD scoring functions

Table 11 shows the comparison of various methods in terms of FPR@95, AUROC, and AUPR metrics. For instance, using a parameter-free EBO scoring function, our method achieves significantly superior performance in comparison with others.

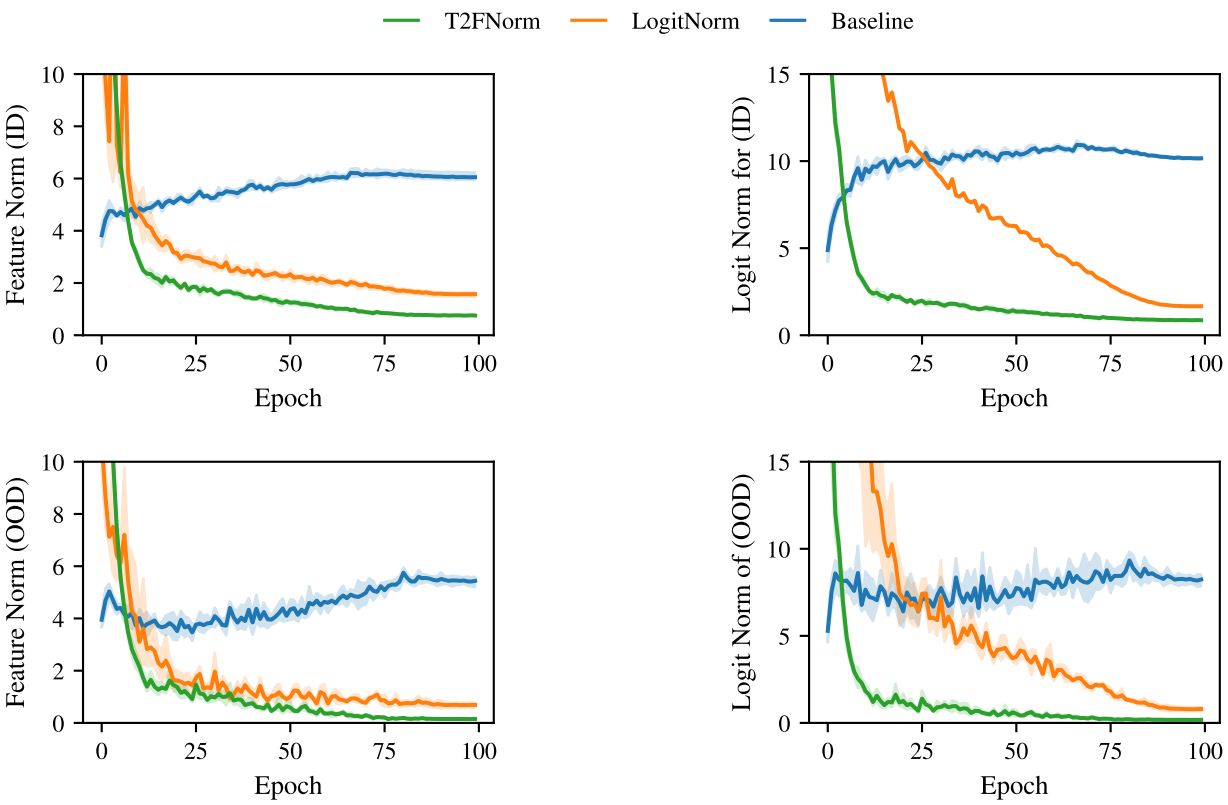


Figure 17. Progression of Norm in feature and logit space over training epochs (CIFAR10).

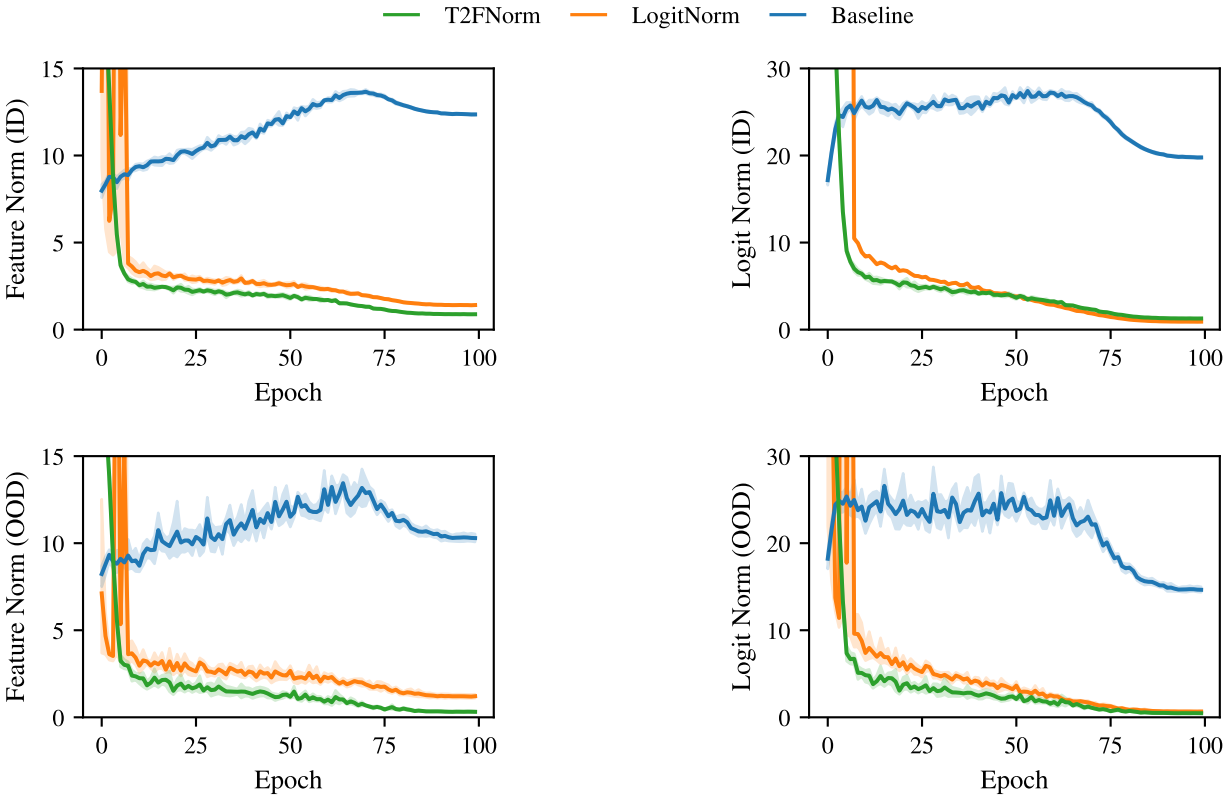
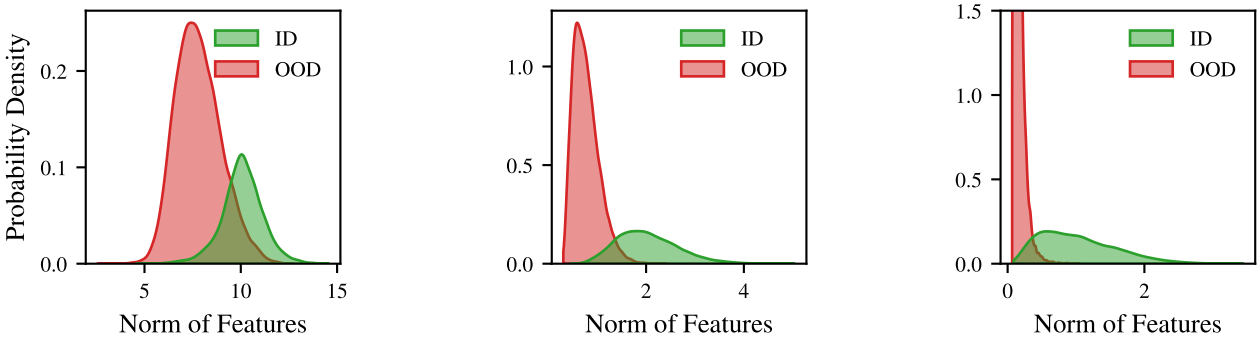


Figure 18. Progression of Norm in feature and logit space over training epochs (CIFAR100).



(a) Baseline (b) LogitNorm (c) T2FNorm

Figure 19. Distribution of norm of feature of ResNet-18 model trained with CIFAR-10

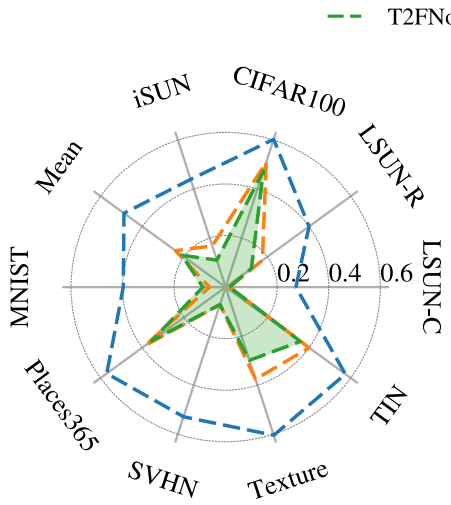


Figure 20. FPR@95 for CIFAR-10 as ID (MSP) (DenseNet)

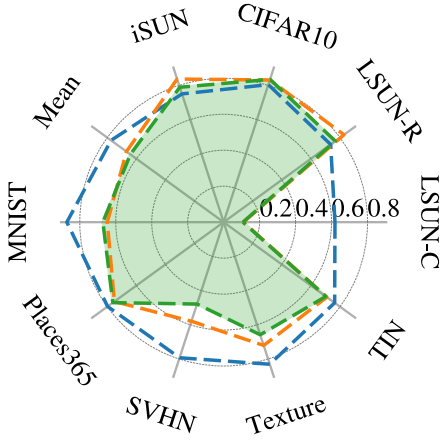


Figure 21. FPR@95 for CIFAR-100 as ID (MSP) (DenseNet)

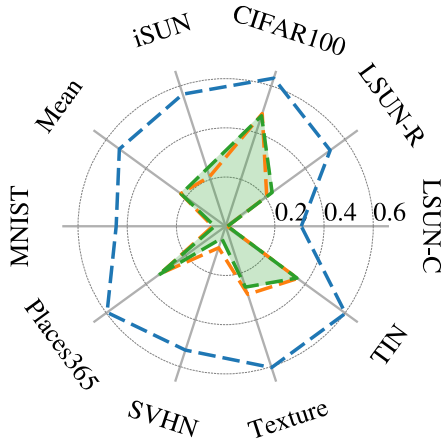


Figure 22. FPR@95 for CIFAR-10 as ID (MSP) (WRN-40-2)

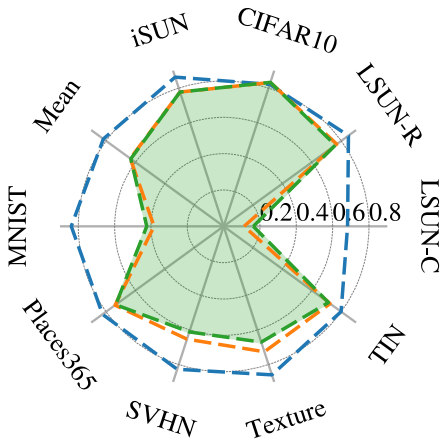


Figure 23. FPR@95 for CIFAR-100 as ID (MSP) (WNR-40-2)

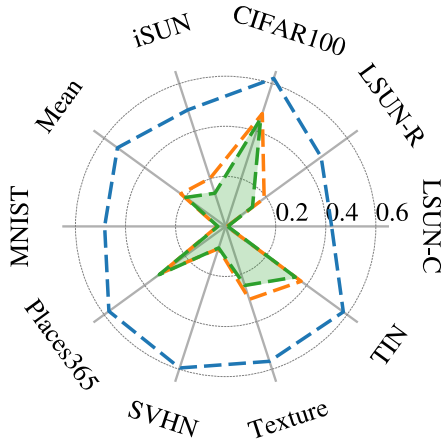


Figure 24. FPR@95 for CIFAR-10 as ID (MSP) (ResNet-18)

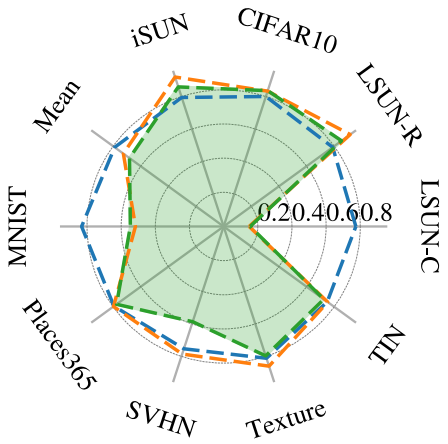


Figure 25. FPR@95 for CIFAR-100 as ID (MSP) (ResNet-18)

Table 11. Mean OOD metrics obtained through various OOD scoring functions in the form of Baseline / LogitNorm / T2FNorm across various architectures.

		CIFAR-10			CIFAR-100		
	Network	FPR@95 ↓	AUROC ↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑
EBO	ResNet	37.7/37.0/**17.9**	91.5/88.9/**96.7**	92.7/89.4/**96.6**	77.6/72.6/**66.6**	81.0/75.0/**83.3**	81.2/75.3/**82.2**
	WRN	35.4/54.9/**22.5**	91.1/85.0/**95.8**	92.1/84.1/**95.7**	78.0/62.6/**60.0**	77.0/81.7/**84.2**	78.3/81.9/**84.4**
	DenseNet	30.3/73.9/**20.0**	93.3/86.3/**96.1**	93.8/83.2/**96.1**	69.2/70.3/**62.3**	82.4/75.7/**83.5**	83.6/77.0/**83.9**
	Mean	34.5/55.3/**20.1**	92.0/86.7/**96.2**	92.9/85.6/**96.2**	75.0/68.5/**63.0**	80.1/77.5/**83.6**	81.1/78.1/**83.5**
GradNorm	ResNet	81.0/25.9/**18.0**	59.3/95.0/**96.6**	69.9/94.9/**96.5**	77.2/71.4/**65.3**	71.0/72.2/**82.0**	76.3/73.6/**81.3**
	WRN	70.4/26.6/**23.4**	63.0/94.6/**95.6**	71.9/94.6/**95.6**	87.8/60.4/**57.2**	52.3/81.3/**84.3**	61.1/81.4/**84.4**
	DenseNet	48.6/31.6/**21.3**	79.5/92.8/**95.5**	84.8/92.9/**95.6**	76.2/69.9/**62.6**	69.2/73.4/**82.2**	73.2/74.4/**82.2**
	Mean	66.7/28.0/**20.9**	67.3/94.2/**95.9**	75.6/94.1/**95.9**	80.4/67.2/**61.7**	64.2/75.6/**82.8**	70.2/76.5/**82.6**
Odin	ResNet	38.7/19.4/**17.2**	87.1/96.4/**96.9**	90.7/96.3/**96.9**	72.5/69.2/**66.5**	82.3/81.6/**84.1**	83.3/81.2/**83.8**
	WRN	43.0/**20.2**/20.8	84.3/**96.2**/96.1	88.4/**96.4**/96.2	73.8/**60.0**/60.1	76.2/84.5/**84.6**	79.3/85.1/**85.3**
	DenseNet	34.7/21.7/**18.1**	90.4/95.6/**96.4**	92.2/95.7/**96.6**	64.9/61.0/**57.5**	82.4/83.8/**85.3**	84.5/84.8/**86.5**
	Mean	38.8/20.4/**18.7**	87.3/96.1/**96.5**	90.4/96.1/**96.6**	70.4/63.4/**61.4**	80.3/83.3/**84.7**	82.4/83.7/**85.2**
TempScale	ResNet	45.6/22.5/**19.7**	91.3/95.9/**96.5**	91.7/95.5/**96.3**	73.4/66.1/**61.6**	81.6/81.9/**84.6**	81.4/80.1/**82.7**
	WRN	45.1/**22.5**/22.8	90.5/**95.9**/95.8	91.0/**95.7**/95.7	70.5/58.2/**57.9**	79.2/84.9/**85.1**	79.2/84.0/**84.7**
	DenseNet	39.3/23.2/**20.1**	92.6/95.5/**96.2**	92.7/95.2/**96.2**	66.6/58.7/**56.5**	82.4/84.6/**85.9**	82.5/83.8/**85.6**
	Mean	43.3/22.7/**20.9**	91.5/95.8/**96.2**	91.8/95.5/**96.0**	70.1/61.0/**58.7**	81.1/83.8/**85.2**	81.0/82.7/**84.3**