A Robust Out-of-Distribution Detection Framework via Synergistic Smoothing

Supplementary Material

The appendix is organised as follows:

- In Section A, we provide results for our proposed post-processing methods on CIFAR-100 and ImageNet.
- In Section B, we provide ablation studies for hyperparameters λ, number of samples, and the noise magnitude.
- In Section C, we analyse our post-processor using other OOD baseline scores (MSP, EBO, and fDBD).
- In Section D, we discuss the robustness of ROSS against adaptive and black-box attacks.
- In Section E, we provide an analysis of inference latency and computational overhead.
- In Section F, we expand on the theoretical differences between ROSS and directional perturbation methods.
- Finally, in Section G, we detail methodological clarifications regarding score direction and validation splits.

A. Benchmark Results on Large-Scale Datasets

In this section, we present additional results for our proposed post-processing methods for out-of-distribution (OOD) detection. In addition to the evaluation on CIFAR-10 presented in the main text, we also present results for CIFAR-100 and ImageNet.

On the CIFAR-100 benchmark, the stability measure (σ_{med}) alone provides poor separation between in-distribution (ID) and out-of-distribution (OOD) data. At the same time, the median-smoothed GEN score (S_{med}) is far more effective. A naive combination, represented by the $S_{\text{med}}/\sigma_{\text{med}}$ score, significantly degrades detection performance compared to using the median score by itself.

For the ImageNet benchmark, we see a similar trend. The σ_{med} and $S_{\text{med}}/\sigma_{\text{med}}$ scores provide poor ID-OOD separation. Both S_{med} and S_{ROSS} provide the best results, with S_{ROSS} narrowly winning out.

These findings are consistent with insights from PRO [6], which found that for large-scale models, the score instability of ID inputs can become indistinguishable from that of OOD inputs. Despite this challenge, our proposed method, ROSS, successfully integrates the stability metric without this penalty, resulting in almost no degradation compared to the strong baseline median scores.

B. Ablation Studies

In this section, we present a series of ablation studies to examine the impact of key components and hyperparameters in our proposed ROSS framework.

B.1. Sensitivity to Hyperparameter λ

We present experiments with different values for the stability weighting parameter λ. The analysis is performed on models trained on CIFAR-10, CIFAR-100 and ImageNet.

The selection of the hyperparameter λ in the ROSS score is a trade-off between the effect of the median score and stability. For the CIFAR-10 model, a λ value of 0.01 or 0.02 achieves the best overall average performance. For CIFAR-100, a smaller value of $\lambda = 0.005$ provides the best average results, primarily due to the median score providing a stronger separation in these models. However, varying values of λ have little effect on the final results. For ImageNet, we see that a larger λ of 0.1 produces the best results. As with CIFAR-100, the choice of λ does not change the results drastically.

B.2. Number of Samples

We evaluate the robustness of the ROSS-fDBD score against PGD-min and PGD-max adversarial attacks on a CIFAR-10 model. We analyse the impact of the number of samples (N) used to compute the score and the magnitude of internal noise (σ).

While smaller sample sizes ($N = 5, 10$) occasionally yield better scores, these results are attributed to the high variance and noise inherent in limited sampling. As N increases, the metrics converge, reflecting a more accurate and stable estimation of the model's true robustness against adversarial perturbations. Consequently, we adopt $N = 25$ as it represents an optimal trade-off, providing the stability observed at higher sample counts ($N = 100$) while maintaining manageable runtime costs.

B.3. Noise Magnitude: The Robustness-Accuracy Trade-off

A key advantage of ROSS is that it allows for an explicit trade-off between performance on clean data and adversarial robustness, controlled by the noise magnitude hyperparameter σ. As shown in the data presented in Table 10 and visualised for PGD-Max attacks in Figure 3, a small noise value ($\sigma = 0.025$) achieves the highest AUROC on unattacked inputs (91.15%) but collapses under strong attacks, dropping to just 18.94% AUROC against a PGD-max attack with $\epsilon = 8/255$. A larger noise magnitude ($\sigma = 0.25$) boosts adversarial robustness, maintaining an AUROC of 50.46% under the same strong attack, but it degrades performance on clean data (58.84% AUROC). In our main body of experiments, we use $\sigma = 0.1$, which offers a practical balance, sacrificing a minimal amount of perfor-

mance on clean inputs to achieve a substantial improvement in adversarial robustness over the lower-noise setting.

C. Performance with Alternative Base Scores

We extend the analysis of ROSS beyond the GEN base score. This section presents results for ROSS and the baseline score when applied to three other widely-used OOD detection scores: Maximum Softmax Probability (MSP), Energy-based OOD detection (EBO), and Fast Decision Boundary OOD detector (fDBD). For each base score, we again evaluate the performance of the median score (S_{med}), the stability score (σ_{med}), their naive ratio, and our final proposed score (S_{ROSS}) on a model trained on CIFAR-10 (Tables 11-13).

While median smoothing (S_{med}) performs competitively, marginally outperforming ROSS on the fDBD benchmark, they exhibit sensitivity to the underlying score distribution. This instability is most evident on the GEN score (Table 1), where ROSS achieves an average FPR95 of 45.77%, significantly outperforming S_{med} (53.79%) by a margin of 8%. Similarly, on EBO, ROSS improves upon S_{med} by over 7% (Table 12). ROSS mitigates this sensitivity by adapting to the noise profile of each score. Although S_{med} achieves a slight edge in isolated cases, ROSS delivers consistently optimal or near-optimal metrics across all tested base scores (MSP, EBO, fDBD, and GEN). This cross-score stability establishes ROSS as the more robust, agnostic choice for OOD detection.

D. Robustness Against Adaptive Attacks

To thoroughly evaluate the robustness of ROSS and verify that our results are not over-optimistic, we consider adaptive attacks such as Expectation over Transformation (EOT) [2]. In our standard evaluation, we directly target the base scoring function. This provides the most precise gradient signal, as the goal is to shift the score distribution sufficiently to alter the robust statistic [9].

As demonstrated by Gao et al. [12], EOT offers no advantage over standard Projected Gradient Descent (PGD) attacks unless the noise level is substantial ($\sigma \geq 0.5$). For ROSS, which operates effectively at a relatively low noise magnitude ($\sigma = 0.1$), standard attacks are often optimal. This is because base gradients provide precise descent directions, whereas EOT relies on high-variance estimates [31]. Furthermore, it is critical to note that smoothing within ROSS does not result in gradient masking; rather, it is a derived statistical property of the landscape.

Empirically, when we ran adaptive EOT attacks (using CIFAR-10 as ID, $N = 5$, $\epsilon = 4/255$) explicitly targeting ROSS, they proved weaker than standard PGD applied to the base post-processor (resulting in approximately +3% AUROC and -7% FPR95 compared to PGD). We also investigated black-box attack methods and found that, within the standard adversarial threat model ($\epsilon \leq 8/255$), black-box attacks were unable to reliably degrade the ROSS score. Successful degradation required perturbation magnitudes significantly higher than the standard radius.

E. Inference Latency Analysis

While ROSS requires N forward passes ($N = 25$) to compute the stability statistic, it is important to note that these evaluations are entirely independent and highly parallelisable. On modern GPU architectures, the N perturbed inputs are naturally processed in a single batch, avoiding sequential bottlenecks.

We measured the wall-clock inference time on an NVIDIA RTX 4060 Ti using 100 CIFAR-10 images. The base MSP score requires approximately 9.6 ms/img, while ROSS ($N = 25$) takes approximately 11.43 ms/img. Due to parallelisation, this represents a minor latency increase of only $\sim 1.2\times$, rather than a naive $25\times$ scaling. Furthermore, as shown in our ablation studies (Table 9), ROSS remains highly robust even at $N = 5$ (where AUROC drops by $< 0.5\%$ compared to $N = 25$). Therefore, for users with capable GPUs, the latency difference introduced by ROSS is negligible for practical deployments.

F. Directional Perturbation vs. Non-Directional Stability

A fundamental distinction between ROSS and prior perturbation-based methods (such as ODIN and PRO) lies in directionality versus stability. Existing methods rely heavily on directional gradient optimisation—specifically ascending or descending scores to push representations apart. This creates an asymmetric vulnerability: optimising an input for one specific direction (e.g., minimising the score) frequently leaves the model acutely vulnerable to the opposite attack direction (e.g., maximising the score). For instance, as shown in Table 3 of the main text, PRO-fDBD degrades severely under PGD-max attacks.

In contrast, ROSS is fundamentally non-directional. We do not computationally optimise the input toward a specific score threshold. Instead, we measure the statistical stability of the local score landscape using the median and median absolute deviation (σ_{med}). This approach provides symmetric robustness, remaining highly effective against both PGD-min and PGD-max attacks, and entirely avoids the reliance on fragile gradient tracking during inference.

G. Further Evaluation Details

Score Direction: Throughout our evaluations, we strictly adhere to the standard OOD detection convention: High Score = ID (In-Distribution) and Low Score = OOD. Consequently, PGD-min attacks attempt to aggressively lower the

score of ID samples (to misclassify them as OOD), whereas PGD-max attempts to artificially raise the score of OOD samples (to misclassify them as ID). As detailed in Algorithm 1 and Equation 2 of the main text, the stability bonus is deliberately added (increasing the final score) only if the sample is already high-confidence (ID-like).

Validation Split and Thresholding: The S_{95} threshold is calculated using the standard validation split containing 10% held-out ID data, ensuring consistency with established frameworks like OpenOOD. It strictly follows the standard False Positive Rate at 95% True Positive Rate (FPR95) threshold.

Recalibration: Recalibration of the S_{95} threshold is only necessary if the definition of the ID task changes significantly (e.g., severe domain shift). Under standard conditions, S_{95} is computed efficiently just once post-training.

Table 4. Analysis of various post-processors using the base **GEN** score for a model trained on **CIFAR-100** (ID) and evaluated against various OOD benchmarks. Performance is reported as FPR95 (%) ↓ / AUROC (%) ↑. Best metric is in **bold** and second best is underlined.

Post-processor	near-OOD		far-OOD				Avg.
	CIFAR-10	TIN	MNIST	SVHN	Texture	Places365	
S_{med}	72.65/**69.48**	**67.18/74.08**	60.88/71.92	**56.75/75.33**	**86.36**/56.88	**72.87/71.41**	**69.45**/69.85
σ_{med}	97.35/41.15	98.56/35.58	99.29/18.15	97.44/35.57	99.33/38.18	98.16/37.48	98.36/34.35
$S_{\text{med}}/\sigma_{\text{med}}$	84.11/66.21	80.28/71.27	**44.65/83.66**	72.80/72.40	89.11/**63.66**	83.24/68.97	75.70/**71.03**
S_{ROSS}	**72.64**/69.48	67.21/74.08	60.96/71.87	56.77/75.32	86.42/56.84	72.88/71.41	69.48/69.83

Table 5. Analysis of various post-processors using the base **GEN** score for a ResNet-50 model trained on **ImageNet** (ID) and evaluated against several OOD benchmarks. Performance is reported as FPR95 (%) ↓ / AUROC (%) ↑. Best results are in **bold**; second best are underlined.

Post-processor	near-OOD		far-OOD			Avg.
	SSB-hard	NINCO	iNaturalist	Textures	OpenImage-O	
S_{med}	77.40/71.63	52.46/82.60	29.89/91.24	48.51/**87.02**	36.69/88.41	48.99/84.18
σ_{med}	98.01/48.51	99.07/48.98	99.70/32.79	99.94/27.94	99.65/33.96	99.27/38.44
$S_{\text{med}}/\sigma_{\text{med}}$	91.35/60.08	92.46/63.20	82.96/78.37	84.78/79.58	85.08/76.61	87.53/71.57
S_{ROSS}	**76.89**/**71.67**	**51.65**/**82.77**	**29.61**/**91.29**	**48.48**/86.90	**36.49**/**88.43**	**48.62**/**84.21**

Table 6. Analysis of the ROSS-GEN post-processor on **CIFAR-10** with different lambda (λ) values. Performance is reported as FPR95 (%) ↓ / AUROC (%) ↑. Best metric is in **bold** and second best is underlined.

λ	CIFAR-100	TIN	MNIST	SVHN	Texture	Places365	Avg.
0.005	58.75/83.20	49.90/85.80	49.83/84.68	57.42/76.89	52.04/84.72	42.14/88.53	51.68/83.97
0.01	55.74/83.46	48.51/86.02	48.33/84.87	54.86/77.42	51.23/**84.80**	41.39/88.69	50.01/84.21
0.02	53.35/83.71	46.66/86.21	46.69/85.03	51.57/77.96	**49.98**/84.75	40.32/88.82	48.09/84.41
0.05	49.90/84.09	44.26/86.53	44.28/**85.23**	46.89/78.80	50.33/84.59	38.98/89.05	45.77/84.72
0.1	**48.43/84.26**	**43.53/86.65**	**43.70**/85.18	**44.88/79.21**	51.13/84.26	**38.44/89.15**	**45.02/84.79**

Table 7. Analysis of the ROSS-GEN post-processor on **CIFAR-100** with different lambda (λ) values. Performance is reported as FPR95 (%) ↓ / AUROC (%) ↑. Best metric is in **bold** and second best is underlined.

λ	CIFAR-10	TIN	MNIST	SVHN	Texture	Places365	Avg.
0.005	72.68/69.55	**66.79/74.15**	**60.74/71.95**	**56.70**/75.37	86.66/56.91	72.81/**71.46**	**69.40/69.90**
0.01	72.64/69.48	67.21/74.08	60.96/71.87	56.77/75.32	86.42/56.84	72.88/71.41	69.48/69.83
0.02	72.68/69.55	**66.79**/74.14	60.85/71.87	**56.70**/75.37	86.80/56.84	72.83/**71.46**	69.44/69.87
0.05	**72.33**/69.99	67.23/74.12	69.00/64.53	57.94/**76.91**	**85.13/58.42**	**72.36**/71.19	70.67/69.19
0.1	72.35/**70.10**	67.28/**74.16**	68.98/63.80	58.02/76.87	85.21/58.10	72.45/71.24	70.72/69.05

Table 8. Analysis of the ROSS-GEN post-processor on **ImageNet** with different lambda (λ) values. Performance is reported as FPR95 (%) ↓ / AUROC (%) ↑. Best metric is in **bold** and second best is underlined.

λ	SSB-Hard	NINCO	iNaturalist	Textures	OpenImage-O	Avg.
0.005	77.24/71.63	52.35/82.60	29.77/91.24	**48.34/87.03**	37.05/88.41	48.95/84.18
0.01	77.18/71.64	52.25/82.62	29.52/91.25	48.42/86.99	36.78/88.42	48.83/84.18
0.02	77.07/71.65	51.88/82.65	29.42/91.25	48.43/86.98	36.81/**88.43**	48.72/84.19
0.05	76.89/71.67	51.65/82.77	29.61/91.29	48.48/86.90	36.49/**88.43**	48.62/**84.21**
0.1	**76.75/71.68**	**50.12/82.89**	**29.11/91.31**	48.70/86.78	**36.34/88.44**	**48.20/84.22**

Table 9. OOD detection robustness vs. number of samples (N) for noise $\sigma = 0.1$ and $\lambda = 0.05$. Results are averaged over all benchmarks for a **CIFAR-10** model under PGD attack and reported as FPR95 (%) ↓ / AUROC (%) ↑.

Samples (N)	$\epsilon = 2/255$		$\epsilon = 4/255$		$\epsilon = 8/255$	
	PGD-Min	**PGD-Max**	**PGD-Min**	**PGD-Max**	**PGD-Min**	**PGD-Max**
5	59.48/78.34	60.09/78.79	71.52/71.63	69.94/72.39	88.27/56.98	87.98/54.83
10	58.33/78.58	58.64/79.03	70.68/71.86	68.45/72.64	88.03/57.09	87.76/54.79
25	58.42/78.75	58.80/78.78	70.83/71.95	70.35/71.62	88.30/56.85	88.08/54.32
50	58.07/78.76	57.58/79.27	70.53/71.95	67.42/72.83	88.17/56.85	87.66/54.79
100	58.05/78.79	57.59/79.32	70.61/71.95	66.62/73.27	88.24/56.78	87.75/54.75

Table 10. OOD detection robustness vs. noise magnitude (σ) for sample size $N = 25$. Results are averaged over all benchmarks for a **CIFAR-10** model under PGD attack using GEN and reported as FPR95 (%) ↓ / AUROC (%) ↑.

Noise Magnitude	**No Attack**	$\epsilon = 2/255$		$\epsilon = 4/255$		$\epsilon = 8/255$	
	($\epsilon = 0$)	Min	Max	Min	Max	Min	Max
0.025	32.19/91.15	58.27/79.34	65.61/77.65	77.81/65.98	84.74/55.65	94.35/44.82	95.68/18.94
0.05	36.07/89.55	56.45/80.69	53.17/79.34	74.02/71.11	71.12/66.65	92.25/54.90	85.41/43.85
0.1	45.77/84.72	58.42/78.75	58.80/78.78	70.83/71.95	70.35/71.62	88.30/56.85	88.08/54.32
0.25	82.11/58.84	82.95/58.87	82.96/57.71	83.90/57.41	83.69/57.21	85.60/55.80	83.90/50.46

Table 11. Analysis of proposed post-processors using base **MSP** score. The model is trained on **CIFAR-10** (ID) and evaluated against various OOD benchmarks. Performance is reported as FPR95 (%) ↓ / AUROC (%) ↑. Best metric is in **bold** and second best is underlined

Post-processor	**near-OOD**		**far-OOD**				**Avg.**
	CIFAR-100	**TIN**	**MNIST**	**SVHN**	**Texture**	**Places365**	
S_{med}	53.80/83.09	46.56/85.30	48.38/**84.24**	48.58/80.74	**45.75/84.90**	40.46/**87.23**	47.25/**84.25**
σ_{med}	**46.94**/82.10	43.17/83.93	45.62/80.29	**40.14/82.16**	48.00/80.04	40.00/85.14	**43.98**/82.28
$S_{\text{med}}/\sigma_{\text{med}}$	46.95/82.90	**43.16**/84.90	45.63/81.74	40.16/82.04	48.00/81.50	40.00/86.46	43.98/83.26
S_{ROSS}	46.95/**83.48**	43.17/**85.48**	**45.60**/83.46	40.19/81.97	48.00/83.56	**39.99**/87.17	43.98/84.19

Table 12. Analysis of proposed post-processors using the base **EBO** score. The model is trained on **CIFAR-10** (ID) and evaluated against various OOD benchmarks. Performance is reported as FPR95 (%) ↓ / AUROC (%) ↑. Best metric is in **bold** and second best is underlined

Post-processor	**near-OOD**		**far-OOD**				**Avg.**
	CIFAR-100	**TIN**	**MNIST**	**SVHN**	**Texture**	**Places365**	
S_{med}	64.98/85.72	54.54/88.30	31.63/92.29	63.97/80.97	52.81/88.30	50.65/89.38	53.10/87.49
σ_{med}	65.00/82.47	55.88/85.28	52.44/84.27	64.57/73.86	57.61/**84.16**	45.96/88.32	56.91/83.06
$S_{\text{med}}/\sigma_{\text{med}}$	52.83/76.59	48.90/78.14	54.55/69.28	45.59/**78.89**	80.30/65.39	46.98/79.20	54.86/74.58
S_{ROSS}	**48.63/84.17**	**43.23/86.66**	**44.14/84.83**	**45.53**/78.34	**55.81**/83.21	**38.34/89.33**	**45.95/84.42**

Table 13. Analysis of proposed post-processors using the base **fDBD** score. The model is trained on **CIFAR-10** (ID) and evaluated against various OOD benchmarks. Performance is reported as FPR95 (%) ↓ / AUROC (%) ↑. Best metric is in **bold** and second best is underlined

Post-processor	**near-OOD**		**far-OOD**				**Avg.**
	CIFAR-100	**TIN**	**MNIST**	**SVHN**	**Texture**	**Places365**	
S_{med}	49.19/84.72	42.44/87.32	**36.40/87.25**	**36.49/84.05**	**41.84/87.20**	**34.31/90.22**	**40.11/86.79**
σ_{med}	47.61/77.73	44.19/79.00	48.41/70.57	42.09/77.63	63.39/69.09	41.83/79.60	47.92/75.60
$S_{\text{med}}/\sigma_{\text{med}}$	46.03/81.79	42.00/83.99	45.23/76.98	39.20/79.95	56.45/75.76	38.56/86.02	44.58/80.75
S_{ROSS}	**45.91/85.10**	**40.57/87.58**	38.65/86.72	36.51/83.87	45.41/86.41	34.79/90.19	40.31/86.64