Defending CLIP via Noise-Induced Feature Dynamics for Training-Free, Zero-shot Adversarial Robustness Supplementary Material

Debarshi Brahma Soma Biswas
Indian Institute of Science (IISc), Bengaluru
{debarshib, somabiswas}@iisc.ac.in

1. Full Results on Adversarial Robustness

Robustness under $\epsilon = 4/255$. We report the full results for the 15 datasets in Table 2, under 10-step PGD [2] attack with $\epsilon = 4/255$. We compare against both adversarially finetuned and test-time defense methods. The attack budgets used during finetuning is denoted using subscripts. The training-based methods finetuned with $\epsilon = 4/255$ show better robustness than training with a lower ϵ. Among test-time methods, TTE, Anti-Adv and HD provides little robustness, while TTC improves over them. However, TTC uses higher counterattack optimization steps (steps = 5) [4], which results in a compromise in the clean accuracy [4]. This is also often unrealistic, since the user does not have prior knowledge of the attack strength. Our proposed approach utilizes the same hyperparameters and demonstrate consistently improved robustness across all datasets, resulting in a significantly improved average robust accuracy ($+27.28\%$ over TTC), while also improving on the average clean accuracy ($+0.81\%$ over TTC).

An adversary with the knowledge of model weights can craft an adversarial attack by the following optimization.

$$x_{adv} = \arg \max_{\|x_{adv} - x\|_p \leq \epsilon} \mathcal{L}(f_\theta(x), y_{true}) \tag{1}$$

where, $\mathcal{L}$ denotes the loss function and y_{true} is the ground truth label of image x. As the attack budget ϵ increases, the image perturbation becomes larger, making the change more perceptible to human eyes. With sufficiently large ϵ values, the image can exhibit significant degradation with little resemblance to its original form, resulting in arbitrary wrong predictions. However, the higher budget also allows finding an adversarial example which is farther away from the input image, resulting in a higher chance of misclassification. Hence for larger ϵ values, the optimization can move the example farther away from the input along highly sensitive feature space directions, often moving it across decision boundaries. In such regions, first-order effects of local behavior alone may be insufficient, and second-order effects can be more pronounced. As discussed in Sec. 4.2, our proposed approach leverages the relative change in drifts across two increasing noise levels, effectively capturing this second-order effect, which may provide a more discriminative signal for adversarial detection and mitigation.

	AOM	TTC	Ours
Running Time (secs)	0.040	0.100	0.102

Table 1. Comparison of per-sample time requirements for training-free defense methods on a single NVIDIA RTX 2080-Ti GPU.

Robustness under CW attack. The full results for all the 15 datasets under CW [1] attack with $\epsilon = 1/255$ is reported in Table 3. Although the adversarial finetuning methods achieve better robustness than test-time defenses, they show degradation in clean accuracies. While TTC shows comparatively improved performance, our proposed approach achieves consistent improvements across all datasets, demonstrating significantly better average robustness, with an absolute gain of 10.66% over TTC.

Robustness under PGD-100 attack. In addition to PGD-10 and CW attack in the main paper, we also perform experiments on the 100-step PGD attack ($\epsilon = 1/255$) and report the robust accuracies across nine datasets in Table 4. While TTC improves the performance over zero-shot CLIP, our proposed Defend-CLIP significantly outperforms across all the datasets, achieving an average gain of 9.31% over TTC.

2. Effect of Other Hyperparameters

We observed a tradeoff between robustness and accuracy, which is dependent on the threshold r_{th} (Sec. 6.3). The accuracy plots with respect to r_{th} are reported for CIFAR10, CIFAR100, TinyImageNet and the Average across all 15 datasets in the main paper. We further illustrate the plots for the remaining datasets in Fig. 1. From Fig. 2 histograms (main paper), it can be observed that although the relative change in feature drift $r(x)$ exhibits discriminative distribu-

Dataset		CLIP	Adversarial Finetuning							Test-time Defence					Δ
			CLIP-FT	$TeCoA_1$	$TeCoA_4$	$PMG\text{-}AFT_1$	$PMG\text{-}AFT_4$	$FARE_1$	$FARE_4$	TTE	Anti-Adv	HD	TTC	Ours	
TinyImageNet	Rob.	0.00	2.19	4.87	10.12	4.39	9.59	0.29	1.24	1.77	0.09	0.01	6.75	**39.70**	+39.70
	Acc.	57.64	77.06	70.86	63.84	66.85	59.77	73.63	70.69	56.73	52.62	51.07	51.85	**58.74**	+1.10
CIFAR10	Rob.	0.43	2.75	7.69	11.70	10.20	15.59	1.94	5.42	3.47	0.32	1.67	28.51	**77.77**	+77.34
	Acc.	85.12	84.90	64.61	65.15	70.69	71.45	74.44	78.46	**84.74**	83.44	78.23	81.18	83.77	-1.35
CIFAR100	Rob.	0.05	0.67	6.54	9.25	7.60	10.80	2.64	4.54	1.37	0.22	0.00	9.06	**45.34**	+45.29
	Acc.	57.14	59.51	35.96	36.30	40.32	41.51	46.67	47.38	**58.61**	53.96	52.86	56.34	56.06	-1.08
STL10	Rob.	0.16	3.75	24.80	31.83	28.49	35.40	9.99	17.59	32.56	2.25	3.39	52.40	**85.96**	+85.80
	Acc.	96.40	94.49	87.40	81.69	88.56	84.35	91.72	89.11	**96.26**	95.47	89.50	95.83	95.13	-1.27
DTD	Rob.	0.11	0.00	4.20	5.16	4.31	5.21	0.90	2.50	7.16	0.37	0.16	11.40	**25.53**	+25.42
	Acc.	40.64	36.49	25.16	20.11	21.76	17.29	32.07	28.03	**41.35**	38.55	34.89	35.69	36.01	-4.63
OxfordPets	Rob.	0.00	1.66	0.90	3.71	1.74	5.10	0.19	0.30	3.18	0.10	0.00	24.64	**74.73**	+74.73
	Acc.	87.44	84.14	62.12	53.94	65.88	56.66	79.37	70.10	**88.13**	80.53	80.91	64.70	81.00	-6.44
Flowers102	Rob.	0.00	0.13	1.87	3.81	2.57	4.26	0.03	0.62	3.52	0.05	0.00	13.60	**43.97**	+43.97
	Acc.	65.46	53.37	36.80	27.78	37.00	28.88	47.98	41.01	**65.20**	62.80	58.22	63.24	61.88	-3.58
FGVCAircraft	Rob.	0.00	0.00	0.03	0.12	0.03	0.06	0.00	0.03	0.43	0.00	0.00	6.40	**10.02**	+10.02
	Acc.	20.10	14.04	5.31	3.51	5.55	3.24	10.86	7.77	**20.18**	15.64	16.36	15.99	16.14	-3.96
StanfordCars	Rob.	0.00	0.00	0.15	0.41	0.15	0.40	0.01	0.04	1.46	0.00	0.00	12.84	**24.66**	+24.66
	Acc.	52.02	42.11	20.91	15.18	25.44	16.79	38.68	32.09	**52.73**	36.14	44.28	41.52	47.56	-4.46
SUN397	Rob.	0.00	0.02	1.30	2.31	1.90	3.24	0.13	0.57	5.95	0.11	0.00	13.43	**41.44**	+41.44
	Acc.	58.50	55.73	36.69	28.16	37.98	29.93	52.42	43.57	**59.12**	55.99	53.17	46.68	55.20	-3.30
Caltech101	Rob.	0.59	4.81	15.75	21.00	19.48	25.03	5.15	10.13	30.19	3.14	1.27	36.66	**74.40**	+73.81
	Acc.	85.66	83.63	71.68	64.41	75.45	69.06	80.95	76.58	**85.84**	83.99	82.33	86.15	85.80	+0.14
Caltech256	Rob.	0.12	1.41	8.29	11.76	10.65	13.68	2.18	5.09	23.23	1.44	0.34	27.25	**64.22**	+64.10
	Acc.	81.72	78.53	61.14	52.05	62.24	53.32	73.32	67.22	**82.48**	79.40	79.12	76.59	80.07	-1.65
Food101	Rob.	0.00	0.04	0.56	1.35	1.03	2.12	0.06	0.24	5.31	0.07	0.01	17.89	**49.26**	+49.26
	Acc.	83.88	64.86	29.98	21.90	36.61	27.97	55.31	41.98	**83.96**	75.95	80.30	80.00	72.67	-11.21
EuroSAT	Rob.	0.00	0.00	9.77	10.71	9.61	10.36	0.00	7.34	0.11	0.03	0.20	13.57	**37.86**	+49.2
	Acc.	42.59	27.64	16.58	17.53	18.53	19.19	21.88	18.22	**44.38**	36.81	39.08	53.24	31.11	-11.48
PCAM	Rob.	0.00	0.00	20.54	44.13	12.59	36.38	0.64	3.74	0.22	0.25	12.04	**47.39**	36.15	+42.93
	Acc.	52.02	47.21	49.96	49.98	50.03	49.80	52.54	50.17	**50.92**	52.61	50.38	52.73	52.61	+0.59
Average	Rob.	0.10	1.16	7.15	11.16	7.65	11.81	1.59	3.95	7.99	0.56	1.27	21.45	**48.73**	+48.63
	Acc.	64.42	60.24	45.01	40.10	46.86	41.95	55.46	50.82	**64.71**	60.26	59.38	60.11	60.92	-3.50

Table 2. The classification accuracies (%) on adversarial images under 10-step PGD with attack budget $\epsilon = 4/255$ (Rob.) along with clean accuracies (Acc.) across all the 15 datasets. We compare with state-of-the-art adversarial finetuning methods as well as test-time training-free methods. The column Δ shows the gains with respect to the zero-shot CLIP performance. Our approach achieves significant robustness gains across all datasets, outperforming prior works on average.

tions for clean and adversarial examples (adversarial samples peak towards higher $r(x)$ values while clean samples form a leftward peak), there still exists some overlap which induces this tradeoff. Consequently, as the chosen threshold r_{th} gradually increases, fewer adversarial samples fall above the threshold and are detected, resulting in a lower robust accuracy and a higher clean accuracy. Conversely, a lower r_{th} detects more adversarial examples, but also increases the number of false positives, reducing the clean accuracy. This tradeoff pattern is consistently visible across all the datasets (Fig. 1).

In this paper, we define the anchor feature f_{anchor} (Sec. 4 and 5) as the mean feature obtained by adding N Gaussian noise samples $\delta^{(j)} \sim \mathcal{N}(0, \sigma^2 I)$ $(j = 1, 2, \ldots, N)$ to the original image and averaging the resultant features.

$$f_{anchor}(x) = \frac{1}{N} \sum_{j=1}^{N} f_\theta(x + \delta^{(j)}) \tag{2}$$

The effect of the number of Gaussian noise samples across six datasets, DTD, CIFAR10, CIFAR100, STL10, FGVCAircraft, TinyImageNet is illustrated in Fig. 2 for $N = 1, 5, 10, 20$. As observed, adding a single sampled noise can have random effects on the feature estimate, which results in consistently lower robust and clean accuracies. As N increases, the performance slightly improves, but saturates after $N = 10$, indicating that averaging over more number of samples provides a stable anchor feature.

Time Requirements We compare the per-sample time requirements of our method with the recent training-free test-time defenses, AOM [3] and TTC [4] in Table 1. AOM involves simple fixed Gaussian noise addition and feature extrapolation, resulting in very less inference time per image. On the other hand, TTC involves optimizing an added counterattack perturbation and takes more time compared to AOM. Our proposed approach involves computing the relative drift change across two noise levels for adversar-

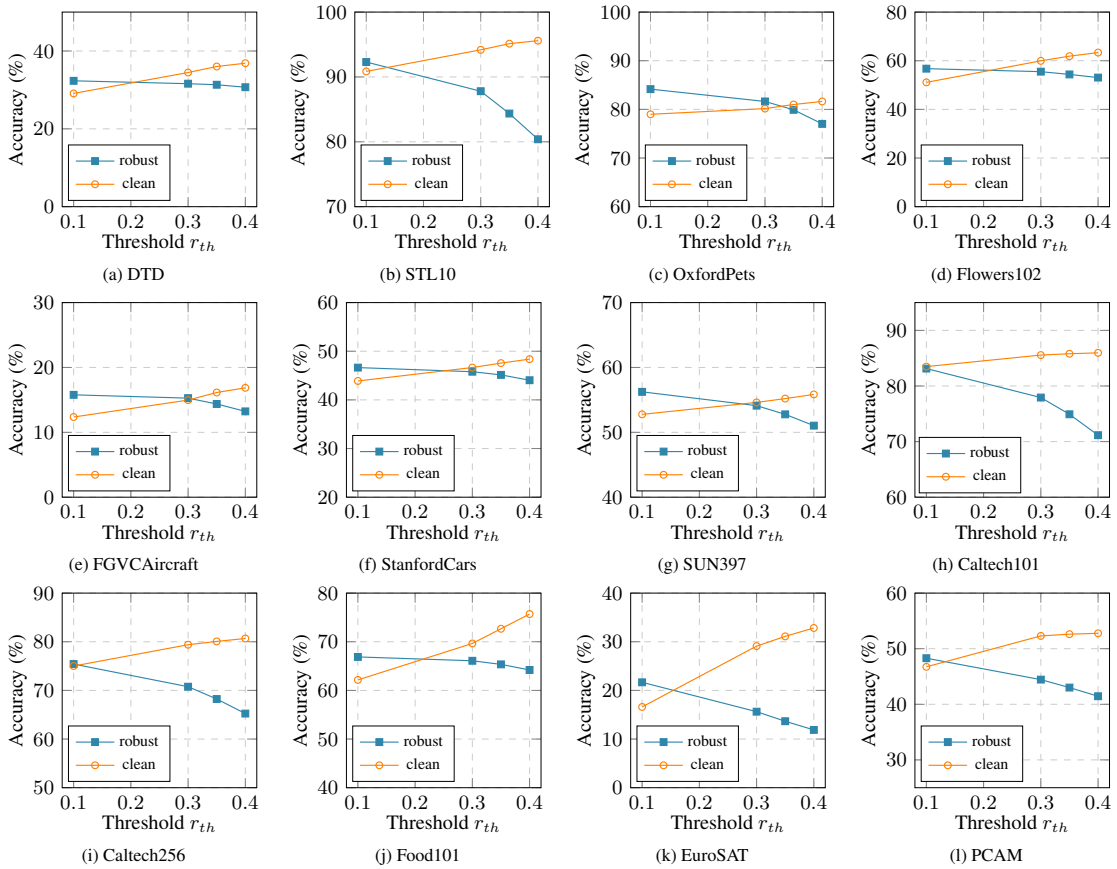

Figure 1. Effect of varying threshold r_{th} on clean and robust accuracies for the remaining datasets, under PGD-10 attack with $\epsilon = 1/255$. We observe a consistent accuracy-robustness tradeoff pattern with changing threshold.

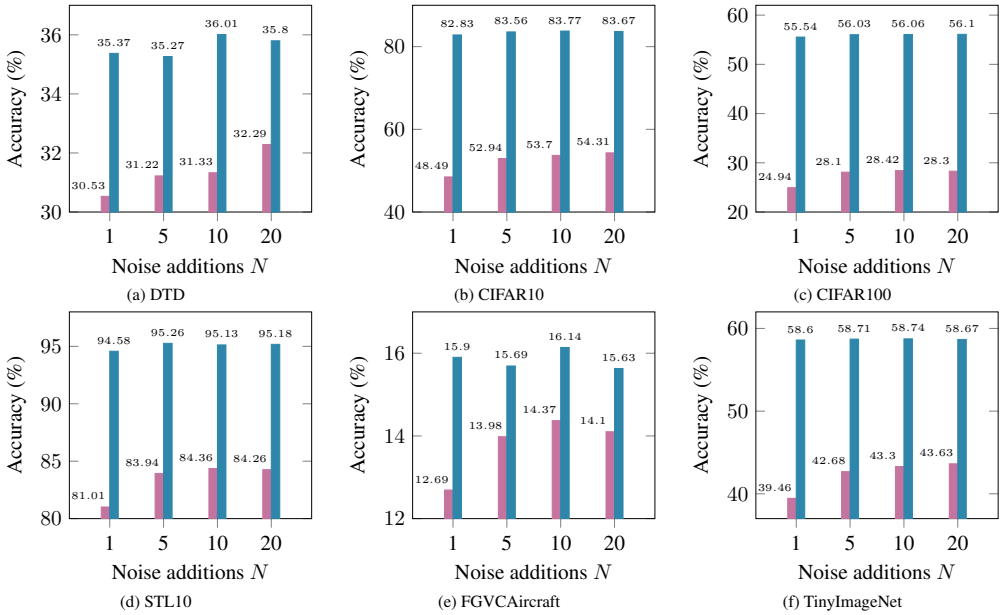

Figure 2. Effect of the number of noise additions (N) on the clean (blue) and robust (purple) accuracies, for $N = 1, 5, 10, 20$.

Dataset		CLIP	Adversarial Finetuning				Test-time Defence					Δ
			CLIP-FT	TeCoA	PMG-AFT	FARE	TTE	Anti-Adv	HD	TTC	Ours	
TinyImageNet	Rob.	0.36	1.06	48.00	43.79	27.71	19.40	5.48	3.70	19.75	**42.31**	+41.95
	Acc.	57.64	77.06	70.86	66.85	73.63	**56.73**	52.76	52.49	51.85	58.74	+1.10
CIFAR10	Rob.	0.87	0.94	33.27	39.50	20.6	40.01	12.53	14.79	29.04	**52.86**	+51.99
	Acc.	85.12	84.90	64.61	70.69	74.44	**84.74**	83.52	78.64	81.18	83.77	-1.35
CIFAR100	Rob.	0.29	0.39	18.27	20.83	11.67	18.73	6.56	3.04	14.38	**27.70**	+27.41
	Acc.	57.14	59.51	35.96	40.32	46.67	**58.61**	53.95	53.50	56.34	56.06	-1.08
STL10	Rob.	12.23	9.95	69.73	72.39	59.60	78.64	38.66	37.73	76.40	**84.49**	+72.26
	Acc.	96.40	94.49	87.40	88.56	91.72	**96.26**	95.45	89.54	95.85	95.13	-1.27
DTD	Rob.	2.87	2.77	16.28	13.72	14.36	22.62	6.06	10.11	27.39	**30.96**	+28.09
	Acc.	40.64	36.49	25.16	21.76	32.07	**41.35**	38.92	35.25	36.98	36.01	-4.63
OxfordPets	Rob.	1.64	1.14	37.91	39.28	33.85	51.12	22.99	13.84	57.15	**79.15**	+77.51
	Acc.	87.44	84.14	62.12	65.88	79.37	**88.13**	80.62	80.64	83.35	81.00	-6.44
Flowers102	Rob.	1.35	0.80	21.13	21.34	17.25	34.97	8.06	8.51	36.84	**53.47**	+52.12
	Acc.	65.46	53.37	36.80	37.00	47.98	**65.20**	62.66	57.79	64.16	61.88	-3.58
FGVCAircraft	Rob.	0.00	0.00	2.25	1.86	1.35	5.15	0.83	0.97	12.41	**13.59**	+13.59
	Acc.	20.10	14.04	5.31	5.55	10.86	**20.18**	15.88	16.18	18.00	16.14	-3.96
StanfordCars	Rob.	2.38	2.04	8.74	10.53	9.14	21.19	4.76	5.11	30.38	**45.39**	+43.01
	Acc.	52.02	42.11	20.91	25.44	38.68	**52.73**	36.21	43.60	48.16	47.56	-4.46
SUN397	Rob.	1.75	1.48	18.36	20.39	15.73	29.37	8.85	7.90	39.44	**52.80**	+51.05
	Acc.	58.50	55.73	36.69	37.98	52.42	**59.12**	56.00	54.07	55.13	55.20	-3.30
Caltech101	Rob.	20.88	15.95	56.23	61.58	54.86	69.44	41.47	36.26	66.17	**75.24**	+54.36
	Acc.	85.66	83.63	71.68	75.45	80.95	85.84	84.02	83.00	**86.53**	85.80	+0.14
Caltech256	Rob.	9.69	7.24	42.63	44.55	39.58	59.81	27.17	24.54	58.79	**68.34**	+58.65
	Acc.	81.72	78.53	61.14	62.24	73.32	**82.48**	79.38	79.38	79.66	80.07	-1.65
Food101	Rob.	1.09	0.55	12.87	16.57	12.93	44.61	15.03	9.77	54.65	**64.91**	+63.82
	Acc.	83.88	64.86	29.98	36.61	55.31	**83.96**	75.81	81.02	82.18	72.67	-11.21
EuroSAT	Rob.	0.03	0.03	11.66	11.94	10.66	6.44	2.57	3.47	12.69	**13.93**	+13.90
	Acc.	42.59	27.64	16.58	18.53	21.88	44.38	36.78	40.12	**53.24**	31.11	-11.48
PCAM	Rob.	0.10	1.10	48.29	46.36	16.41	10.70	5.07	46.92	**52.86**	43.05	+42.95
	Acc.	52.02	47.21	49.96	50.03	52.54	50.92	52.49	50.35	**52.73**	52.61	+0.59
Average	Rob.	3.70	3.03	29.71	30.97	22.19	34.15	13.74	15.11	39.22	**49.88**	+46.18
	Acc.	64.42	60.25	45.01	46.86	55.46	**64.71**	60.28	59.70	63.02	60.92	-3.50

Table 3. The classification accuracies (%) on adversarial images under 10-step CW attack with $\epsilon = 1/255$ (Rob.) along with clean accuracies (Acc.) across 15 datasets. We compare our approach with state-of-the-art adversarial finetuning methods and test-time training-free methods. The rightmost column Δ shows the gains with respect to the zero-shot CLIP performance. We show significant robustness gains from prior works, with a modest reduction in average clean accuracy.

Method	DTD	CIFAR 10	CIFAR 100	Food101	OxfordPets	Caltech101	Aircraft	StanfordCars	EuroSAT	Average
CLIP	2.55	0.28	0.08	0.73	0.71	11.03	0.00	0.04	0.02	1.72
TTC	28.19	24.95	13.30	42.48	60.32	64.75	13.35	35.07	8.00	32.27
Ours	**31.70**	**46.31**	**24.63**	**54.43**	**77.43**	**69.62**	**11.85**	**45.03**	**13.26**	**41.58**

Table 4. Classification accuracies (%) on adversarial images under 100-step PGD attack with $\epsilon = 1/255$. Defend-CLIP shows significant robustness gains over prior works.

ial detection followed by mitigating the adversarial samples. This results in comparable time requirements as TTC, while demonstrating significant robustness gains. Overall, the per-sample time requirements for all the methods are extremely low and suitable for practical use-cases.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 1

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learn-

ing models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[3] Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19921–19930, 2025. 2

[4] Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15172–15182, 2025. 1, 2