Supplementary for:
“DAD: Data-free Adversarial Defense at Test Time”

Gaurav Kumar Nayak* Ruchit Rawal* Anirban Chakraborty

Department of Computational and Data Sciences
Indian Institute of Science, Bangalore, India
{gauravnayak, ruchitrawal, anirban}@iisc.ac.in

1. Qualitative Analysis of Low Frequency Component of adversarial data at different Radius

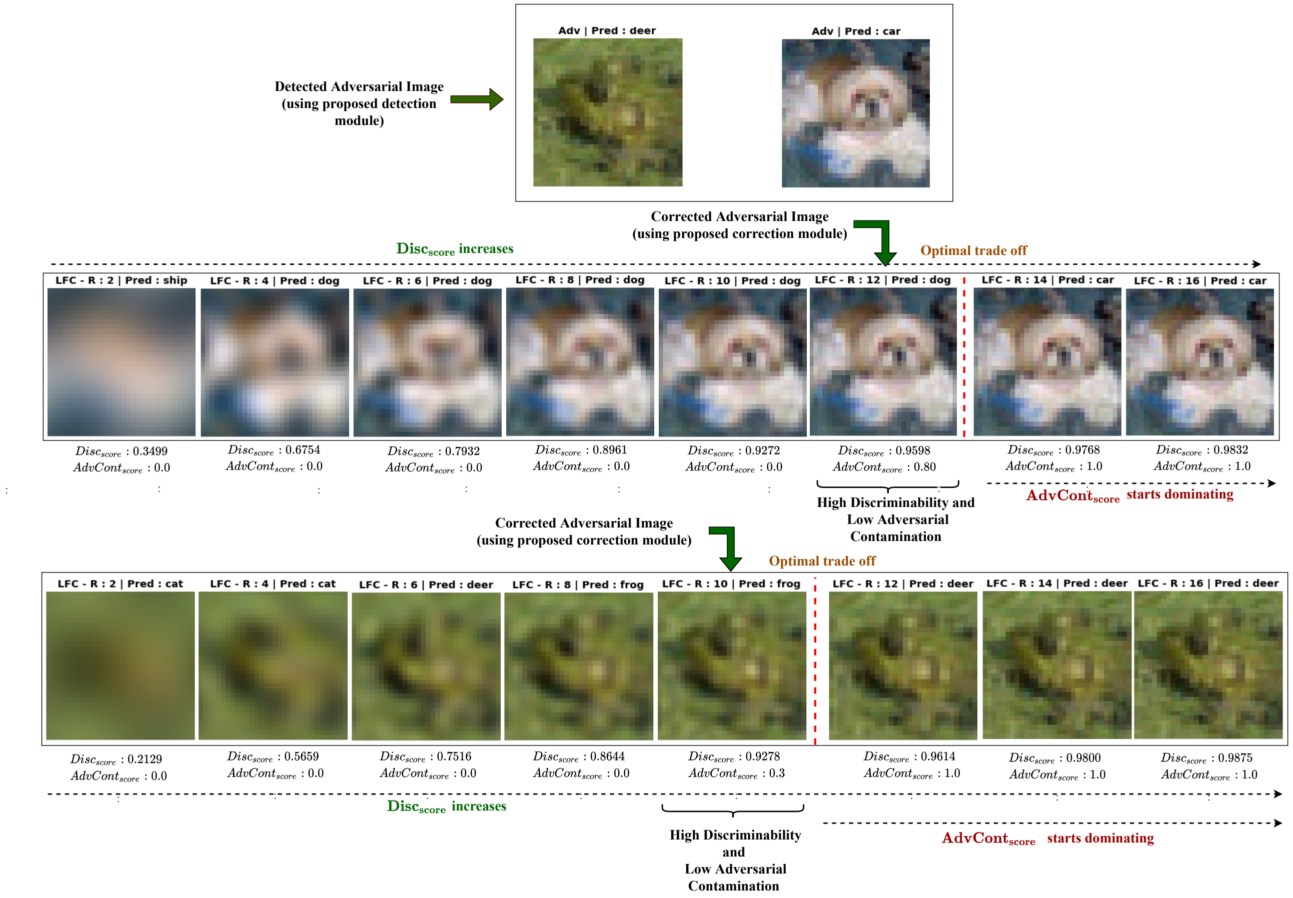
Detected Adversarial Image (using proposed detection module)
Adv | Pred : deer
Adv | Pred : car
Corrected Adversarial Image (using proposed correction module)
$\text{Disc}_{\text{score}}$ increases
Optimal trade off
LFC - R : 2 | Pred : ship
LFC - R : 4 | Pred : dog
LFC - R : 6 | Pred : dog
LFC - R : 8 | Pred : dog
LFC - R : 10 | Pred : dog
LFC - R : 12 | Pred : dog
LFC - R : 14 | Pred : car
LFC - R : 16 | Pred : car
$Disc_{score}$: 0.3499 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.6754 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.7932 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.8961 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.9272 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.9598 $AdvCont_{score}$: 0.80
$Disc_{score}$: 0.9768 $AdvCont_{score}$: 1.0
$Disc_{score}$: 0.9832 $AdvCont_{score}$: 1.0
High Discriminability and Low Adversarial Contamination
$\text{AdvCont}_{\text{score}}$ starts dominating
Corrected Adversarial Image (using proposed correction module)
Optimal trade off
LFC - R : 2 | Pred : cat
LFC - R : 4 | Pred : cat
LFC - R : 6 | Pred : deer
LFC - R : 8 | Pred : frog
LFC - R : 10 | Pred : frog
LFC - R : 12 | Pred : deer
LFC - R : 14 | Pred : deer
LFC - R : 16 | Pred : deer
$Disc_{score}$: 0.2129 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.5659 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.7516 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.8644 $AdvCont_{score}$: 0.0
$Disc_{score}$: 0.9278 $AdvCont_{score}$: 0.3
$Disc_{score}$: 0.9614 $AdvCont_{score}$: 1.0
$Disc_{score}$: 0.9800 $AdvCont_{score}$: 1.0
$Disc_{score}$: 0.9875 $AdvCont_{score}$: 1.0
$\text{Disc}_{\text{score}}$ increases
High Discriminability and Low Adversarial Contamination
$\text{AdvCont}_{\text{score}}$ starts dominating

Figure 1. Visually demonstrating the trade off between discriminability and adversarial contamination. Our correction module suitably handles it through the proposed Algorithm 1 in the main draft by selecting LFC in the spatial domain at optimal radius (r^*) having high $Disc_{score}$ and low $AdvCont_{score}$ i.e. max radius at which $Disc_{score}$ dominates over $AdvCont_{score}$. At radius ($r^* + 1$), the adversarial contamination starts strongly influencing the predictions of pretrained model (T_m) and consequently T_m's prediction for all subsequent radius remains same as the detected adversarial image's prediction. As shown, $r^* + 1$ is 14 and 12 in the top and bottom row respectively. Thus, we select the radius(r^*) as 12 and 10 in these cases and the corresponding LFC in the spatial domain when passed to the model (T_m) yields correct predictions.

[*] denotes equal contribution.

2. Performance of Proposed Detection Module using different Arbitrary Datasets

As described in Sec. 4.1 in the main draft, the target detector model F_t is trained by adapting the source-detector model F_s. The model F_s comprises of S_m and L_{advdet} where S_m is trained on an arbitrary dataset $D_{arbitrary}$. Hence, to assess the effect of the choice of $D_{arbitrary}$ on the detection module and consequently our proposed method DAD (combined detection and correction), we conduct experiments on two distinct source datasets for each target dataset (i.e. CIFAR-10 and FMNIST) in addition to the results provided on TinyImageNet [1] (as source dataset) in the Table 1 in the main draft. We perform experiments with FMNIST and MNIST as $D_{arbitrary}$ for CIFAR-10, while MNIST and CIFAR-10 as $D_{arbitrary}$ for FMNIST. Similar to the results presented in Table 1 in the main draft, we observe from Table 1 (shown below) that we achieve impressive detection accuracy across both the target datasets for each corresponding source dataset ($D_{arbitrary}$).

Target Dataset	Source Dataset (Arbitrary Data)	Detection Accuracy	Clean Accuracy	Adversarial Accuracy
CIFAR-10	MNIST	93.03	99.98	86.08
	FMNIST	93.54	99.85	87.23
FMNIST	MNIST	88.51	99.03	77.99
	CIFAR-10	84.41	84.07	84.75

Table 1. Results (in %) of our proposed detection module comprising clean and adversarial detection accuracy along with overall detection accuracy on Auto Attack, are reported for different target datasets (D_{test}, i.e. CIFAR-10 and FMNIST). For each target dataset, we also vary the source dataset ($D_{arbitrary}$) which is completely different and arbitrary to the target dataset in terms of dissimilar semnatics and non-overlapping categories.

3. Combined (Detection and Correction) Performance on other attacks

In order to evaluate the efficacy of our proposed approach (DAD) across a wide variety of attacks, we extend the analysis on combined performance presented in Sec. 6 and Fig. 3 of the main draft on the state-of-the-art Auto Attack to other popular attacks such as PGD and IFGSM. We observed from Table 2 that we achieve a respectable adversarial accuracy of more than 35% on CIFAR-10 and more than 21% on FMNIST across architectures (Resnet-18 and Resnet-34) on both the attacks while maintaining reasonable clean accuracy.

Dataset	Model	PGD		IFGSM	
		Clean Accuracy	Adversarial Accuracy	Clean Accuracy	Adversarial Accuracy
CIFAR-10	resnet18	89.01	36.38	85.49	35.44
	resnet34	88.28	37.92	88.61	31.94
FMNIST	resnet18	90.09	21.29	88.17	23.15
	resnet34	90.45	21.79	90.57	22.21

Table 2. Performance of our proposed method (DAD) on PGD and IFGSM adversarial attacks where we report the overall clean and adversarial accuracy (in %) across different architectures i.e. Resnet-18 and Resnet-34 for CIFAR-10 and FMNIST.

4. Combined (Detection and Correction) Performance on MNIST

In this section, we evaluate our proposed combined module (detection module followed by correction module) solution strategy, i.e., DAD on the MNIST dataset [2], apart from FMNIST and CIFAR presented in Sec. 6 (Figure 3) of the main draft, to further validate our framework's performance.

We provide combined results on three distinct choices of $D_{arbitrary}$ i.e. CIFAR-10, FMNIST, and TinyImageNet for our target dataset MNIST (D_{test}). We observed (in Table 3) that we achieve a significant boost in the adversarial accuracy across all three choices, without compromising much on the clean accuracy. These results verify that DAD can achieve good performance on a wide range of target datasets (D_{test}) for different choices of the source datasets ($D_{arbitrary}$). Please note the clean accuracy and adversarial accuracy of T_m (resnet18) without our framework was 99.29% and 0.00% respectively against state-of-the-art Auto Attack.

Source (arbitrary) Dataset	Model	Auto Attack	
		Clean Accuracy	Adversarial Accuracy
CIFAR-10	resnet18	90.46	34.77
FMNIST	resnet18	96.12	31.15
TinyImageNet	resnet18	90.29	33.19

Table 3. Results (in %) on MNIST using our proposed DAD framework containing detection and correction modules. In the detection module, the target detection model (F_t) is obtained by adapting the source detection model (F_s) using source-free UDA. The model S_m is appended with detection layers to form F_s. So, we also report the performances for different choices of dataset $D_{arbitrary}$ (i.e. CIFAR-10, FMNIST and TinyImagenet) on which S_m is trained.

5. Attack Parameters for various arbitrary datasets

Source/arbitrary datasets	ϵ	ϵ_{step}	Number of iterations
TinyImageNet	8/255	2/255	20
MNIST	0.3	0.01	100
CIFAR-10	8/255	2/255	20
FMNIST	0.2	0.02	100

Table 4. The parameter values for PGD attack taken across different datasets for creating the adversarial dataset ($A_{arbitrary}$).

6. Distribution of selected radius across samples

As motivated in the Sec. 4.2 of the main draft, our correction Algorithm 1 estimates a suitable radius (r^*) entirely at the test time for each incoming sample without assuming any prior knowledge either about the training dataset or the adversarial attack. We demonstrated the importance of selecting r^* optimally in Table 3 of the main draft by comparing our correction algorithm's performance with a random baseline (R.B.) (wherein a random radius is selected for each sample). We observed that although R.B.'s performance varied across datasets (being higher for CIFAR-10 than FMNIST), it was comfortably outperformed by our correction algorithm.

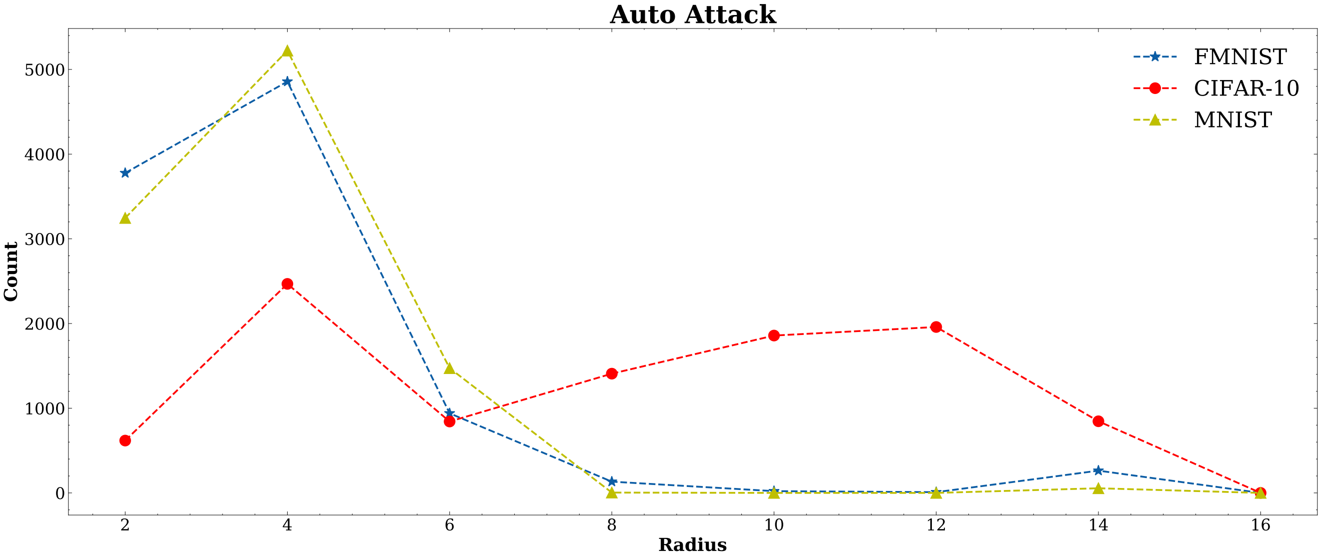

Figure 2. Distribution of radius selected by our proposed correction module on **Auto Attack** adversarial samples across different target datasets i.e. FMNIST, CIFAR-10 and MNIST.

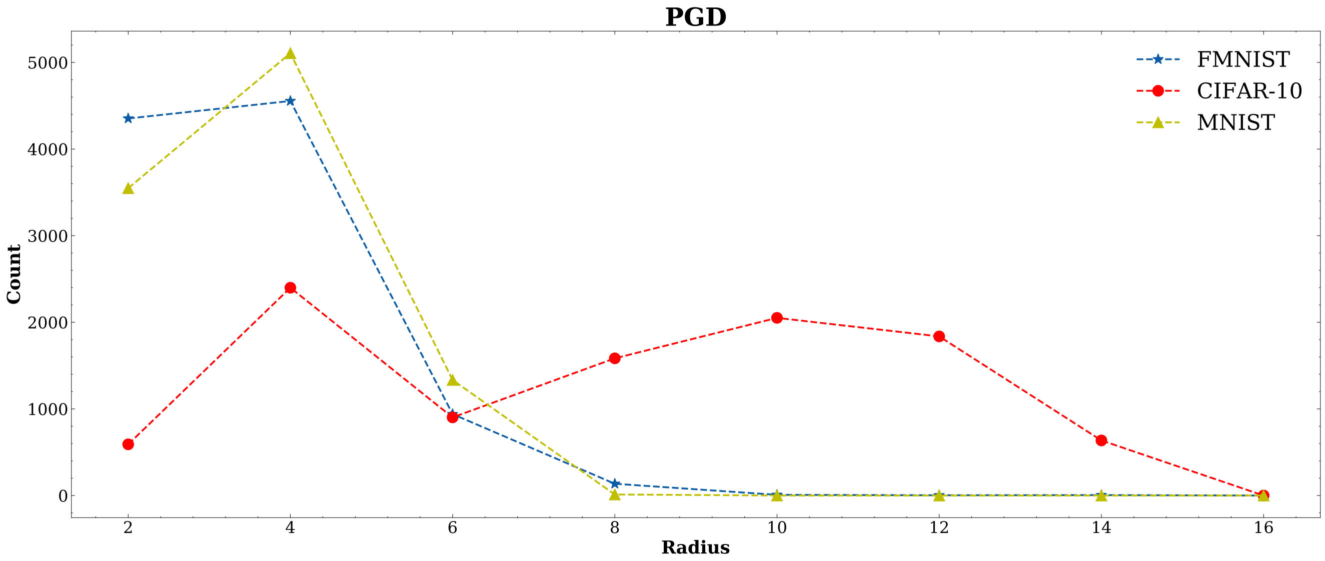

Figure 3. Distribution of radius selected by our proposed correction module on **PGD** Attack adversarial samples across different target datasets i.e. FMNIST, CIFAR-10 and MNIST.

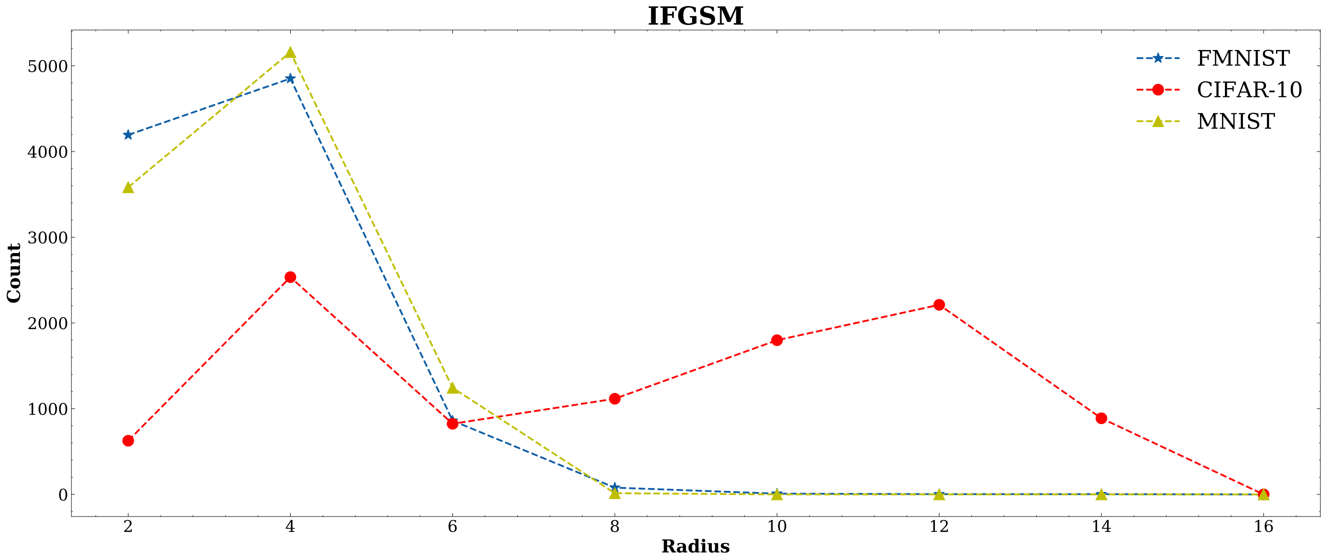

Figure 4. Distribution of radius selected by our proposed correction module on **IFGSM** Attack adversarial samples across different target datasets i.e. FMNIST, CIFAR-10 and MNIST.

In order to investigate the selection of r^* from another perspective, we plot the frequency distribution of r^* for various attacks on Resnet-18 (T_m) trained on multiple datasets as shown in Figures 2, 3 and 4. We observe that for MNIST and FMNIST datasets r^* is majorly selected at a lower radius, whereas for the CIFAR-10 dataset the frequency distribution loosely resembles the uniform distribution. This explains the decent performance of R.B. on CIFAR-10 in Table 3 of the main draft. More importantly, the figures indicate that r^* can significantly vary across datasets (MNIST/FMNIST vs CIFAR-10). Moreover, r^* can even vary across samples of a particular dataset (For *e.g.* in CIFAR-10 samples, different radius are almost equi-proportionally selected). Our algorithm is able to accurately estimate r^* on a sample-by-sample basis without any prior knowledge about the training dataset or the corresponding adversarial attack, as evident by the impressive performance shown in Table 2 of the main draft.

We further compare our performance of a) carefully choosing a constant radius (for all the samples) with b) sample-level radius selection (finer-granularity control) i.e. r^*, through our proposed algorithm. We select the constant radius as $r = 4$ since it's the most frequently selected radius by our algorithm across different types of attacks and datasets (as shown in

Figures 2, 3 and 4). We observe that the sample-level selection strategy often provides significant improvements (shown in Table 5). For e.g. in CIFAR-10 we observe $\approx 16 - 18\%$ improvement by our algorithm over choosing a constant radius ($r = 4$). However, we do notice on the MNIST dataset we observe slightly lower performance. Please note that our proposed algorithm obtains non-trivial improvement in adversarial accuracy even when choosing a constant radius or selecting radius at the sample level. We prefer the sample-level radius selection approach, as we often obtain a large gain in performance across a number of architecture, attack, and datasets configurations.

Dataset	Attack	Performance (in %) ($r = 4$)	Performance (in %) (r^* selected at sample level)
CIFAR-10	pgd	22.08	39.39
	ifgsm	22.26	38.49
	auto attack	22.37	40.25
FMNIST	pgd	27.16	32.22
	ifgsm	27.84	32.38
	auto attack	30.74	35.80
MNIST	pgd	47.59	44.6
	ifgsm	48.63	44.76
	auto attack	48.76	45.81

Table 5. Performance comparison of a) choosing a constant radius $r = 4$ v/s b) sample-level selection strategy i.e. r^*

References

[1] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[2] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.