On Evaluating Stateful Defence Models against Query-Based Black-Box Attacks

Supplementary Material

1. Code

The code is provided at https://github.com/ziyadt1995/evaluating_stateful_defences

2. Algorithm

The pseudocode for DazzlePatch Attack is provided in Algorithm 1.

Algorithm 1 Proposed DazzlePatch Attack

Require: original input x, host image bank $\mathcal{B}$, black-box model f, steps T, budget ϵ

$x_{\text{adv}} \leftarrow x$, $x_{\text{best}} \leftarrow x$, $loss_{best} \leftarrow 10000$
for $t = 1$ to $T - 1$ **do**
  sample $x_b \sim \mathcal{B}$
  $x_{\text{adv}} \leftarrow \text{Perturb}(x_{\text{adv}},\ \epsilon)$
  $x_{\text{q}} \leftarrow \text{CropPatch}\big(x_{\text{adv}}, x_b\big)$
  $loss \leftarrow f(x_{\text{q}})$
  if $loss > loss_{\text{best}}$ **then**
    $loss_{\text{best}} \leftarrow loss$
    $x_{\text{best}} \leftarrow x_{\text{adv}}$
  end if
end for
return x_{best}

3. Datasets and Models

We replicate each defence in its original setting to ensure fidelity. Because datasets differ, we avoid cross-dataset claims and instead compare methods only within a shared setting. Consequently, we apply our evaluation protocol to assess the robustness of three stateful defences: GWAD, QPA, and AdvQDet. GWAD is evaluated on the CIFAR-10 dataset, while QPA and AdvQDet are evaluated on ImageNet. For ImageNet-based evaluations, we randomly sample 200 images from the validation set provided by Huang and Zhang [3], where each selected sample corresponds to a distinct label; for GWAD experiments, instead, we use 1000 samples from the CIFAR-10 test dataset. The classifier correctly classified the samples evaluated. The classifiers used in our work are based on ResNet architecture [2] as in the original works [4, 5, 7]: ResNet-152 for QPA, ResNet-18 for GWAD, and ResNet-50 for AdvQDet.

4. Additional GWAD Experiments

We report the results on GWAD under Square Attack [1] on HoDS (Histogram of Delta Similarity) size of 16. If we decrease the HoDS to 16 then the detection performance becomes significantly better. However, the attack performance is significantly high i.e. above 99% under ℓ_∞ budget of 0.05.

Table 1. Square Attack results on GWAD defence (CIFAR-10) with HoDS size 16.

Metric (%)	$\epsilon = 0.01$	$\epsilon = 0.05$
ASR under ℓ_∞	59.00	99.50
ASR under $\ell_2^{normalised} = 0.1$	58.90	1.10
Avg. hit rate	42.36	36.26

5. OARS Experiments

To evaluate performance of OARS on SDMs and reconfirm its failure against strong SDMs, we perform experiments on 200 ImageNet samples with AdvQDet and QPA as defences. We run the OARS-square attack for 700 iterations. On AdvQDet, the first run of the OARS attack results in 0% ASR with Avg. hit rate of 99.57%, and after 4 random restarts we still get 0% ASR. Meanwhile, OARS attack on QPA results in an ASR of 4.0% with Avg. hit rate of 92.26%. With 4 additional restarts, the ASR jumps to 7.0% with Avg. hit rate of 93.76%. The results show that OARS is not capable of bypassing strong SDMs.

6. Additional AdvQDet Experiments

GWAD has FP rate of 0% on CIFAR-10 while QPA has a FP rate of 0.06% on ImageNet as reported in the original works [4, 5]. FP rate of AdvQDet is not reported. Therefore, we evaluate have evaluated it in our work.

6.1. False Positive Rate

AdvQDet suffers from a high false positive rate on certain datasets. Here, we explore its vulnerability on further datasets. We test all datasets here under a cosine similarity threshold of 0.90.

ImageNet. We find that for certain ImageNet classes, AdvQDet [7] exhibits a notably high false positive rate, as shown in Tab. 2. However, this behavior is not consistent across all labels.

Chinese Traffic Signs. We also test the false positive rate of AdvQDet on Chinese Traffic Sign dataset [8]. The false positive rate is high on most labels. The results are given in Tab. 3.

Skin Cancer. Similarly, on the skin cancer dataset [6], we find that for Bascal Cell Carcinoma (BCC) label, the FP rate is 97.47% with non-cancerous labels. Skin cancer

Table 2. AdvQDet false positive analysis on ImageNet.

Label	FP (%)	Samples
Great white shark	58.0	50
Grey whale	30.0	50
Lemon	6.0	50
Leatherback turtle	12.0	50
Bull frog	12.0	50
Killer whale	18.0	50
Water snake	16.0	50

Table 3. False Positive Analysis on Chinese Traffic Dataset

Traffic Sign	FP (%)	Samples tested
Speed limit 5	41.38	29
Speed limit 40	0	1
Stop sign	6.66	15
No left turn	32.76	58
No straight ahead	56.66	30
Roundabout	0	16
No motor vehicles	11.53	52
Speed limit 60	84.75	59

dataset has images that are extremely visually similar. The visual similarity exposes the FP rate of AdvQDet.

AdvQDet claims zero-shot generalisation in the original work. However, we show that on datasets that have visual similarity between labels, it falsely flags benign images as attacks. Nevertheless, retraining the prompt token on these datasets may mitigate the FP rate.

6.2. Complete DazzlePatch Results

We have provided per-run DazzlePatch attack results on AdvQDet in Tab. 4. The Avg. hit rate given is the average number of hits across all samples in a run.

Using multiple runs under the same ℓ_∞ distance holds promise. However, in these experiments, we have also increased the perturbation budget to $\ell_\infty = 0.07$ and $\ell_\infty = 0.09$ and observe a significant benefit in the ASR.

7. Complete QPA Experiments

We report the per-run experimental results of DazzlePatch Attack with patch size 170 on QPA defence [4]. The results are given in Tab. 5.

Under a perturbation budget of $\ell_\infty = 0.05$, the performance of DazzlePatch stagnates at the 5th run, and increasing the perturbation budget significantly improves the ASR. However, varying the patch sizes in DazzlePatch attack may also improve ASR, as in the case of AdvQDet.

8. Future Work

Our work shows that stateful defence models [4, 5, 7] that claim 100% robust accuracy can be defeated under their assumed threat models. Moreover, this work opens new challenges. One of them is configuring optimal settings for the DazzlePatch attack under a smaller, stricter budget. To this end, novel attacks must be proposed to enhance the *dehosting robustness* phase of the DazzlePatch attack. Similarly, it remains to be seen whether stateful defence models like QPA and AdvQDet can be further robustified against adaptive attacks by retraining the models or introducing additional defences in the pipeline.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. *arXiv preprint arXiv:1911.07140*, 2019. 1

[4] Shaofei Li, Ziqi Zhang, Haomin Jia, Yao Guo, Xiangqun Chen, and Ding Li. Query provenance analysis: Efficient and robust defense against query-based black-box attacks. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 1641–1656. IEEE, 2025. 1, 2, 3

[5] Jeonghwan Park, Niall McLaughlin, and Ihsen Alouani. Mind the gap: Detecting black-box adversarial attacks in the making through query update analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10235–10243, 2025. 1, 2

[6] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 1

[7] Xin Wang, Kai Chen, Xingjun Ma, Zhineng Chen, Jingjing Chen, and Yu-Gang Jiang. Advqdet: Detecting query-based adversarial attacks with adversarial contrastive prompt tuning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6212–6221, 2024. 1, 2

[8] Dmitry Yemelyanov. Chinese traffic signs. https://www.kaggle.com/datasets/dmitryyemelyanov/chinese-traffic-signs. Accessed: 2025-11-21. 1

Table 4. Additional per-run DazzlePatch attack experiments on AdvQDet with 200 ImageNet samples

Run No.	Patch Size	ϵ	Run Misclassifications	Avg. hit rate (%)	Total Misclassifications	Running ASR (%)
1	160	0.05	42	40.55	42	21.00
2	160	0.05	9	41.79	51	25.50
3	160	0.05	9	42.86	60	30.00
4	160	0.05	7	43.86	67	33.50
5	160	0.05	2	44.32	69	34.50
6	170	0.05	2	65.93	71	35.50
7	170	0.05	3	66.32	74	37.00
8	170	0.05	1	66.60	75	37.50
9	170	0.05	2	67.18	77	38.50
10	170	0.05	1	67.33	78	39.00
11	160	0.07	14	39.70	92	46.00
12	160	0.07	8	38.88	100	50.00
13	160	0.07	4	40.06	104	48.00
14	170	0.07	2	65.25	106	52.00
15	170	0.07	0	65.11	106	52.00
16	170	0.07	2	64.89	108	54.00
17	160	0.09	20	37.82	128	64.00
18	160	0.09	0	36.51	128	64.00
19	160	0.09	6	36.42	134	67.00
20	170	0.09	5	63.79	139	69.50
21	170	0.09	3	64.43	142	71.00
22	170	0.09	9	65.83	151	75.50

Table 5. Additional per-run DazzlePatch attack experiments on QPA [4] with 200 ImageNet samples

Run No.	ϵ	Avg. hit rate (%)	Running ASR (%)	Run Misclassifications	Total Misclassifications
1	0.05	2.63	34.50	69	69
2	0.05	3.86	46.0	23	92
3	0.05	3.67	52.50	13	105
4	0.05	3.85	54.00	3	108
5	0.05	3.41	54.50	1	109
6	0.07	1.22	66.00	23	132
7	0.07	3.42	71.00	10	142
8	0.07	2.44	75.00	8	150
9	0.07	2.35	76.00	2	152
10	0.07	1.99	77.00	2	154
11	0.09	2.28	81.00	8	162
12	0.09	1.63	83.00	4	166
13	0.09	2.07	86.50	7	173
14	0.09	2.35	87.00	1	174
15	0.09	1.86	87.50	1	175