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

Ziad Tariq Muhammad Ali[1], Raja Muhammad Atif Azad[1], Muhammad Ajmal Azad[1], Iain Rice[1]
Umar Daraz[1], Ali Shariq Imran[2], James Holyhead[3]

[1] Birmingham City University, [2] Norwegian University of Science and Technology, [3] Covatic

ziad.ali@mail.bcu.ac.uk, ali.imran@ntnu.no, james@covatic.com
{atif.azad, muhammadajmal.azad, iain.rice, umar.daraz}@bcu.ac.uk

## Abstract

*Stateful Defence Models (SDMs) aim to detect the process of adversarial example generation during the query stage. Although they are not designed to counter zero-query attacks, they have shown varying levels of success against query-based black-box attacks. Recently, several SDMs have claimed* $100\%$ *robustness against query-based attacks, which is an extraordinary assertion requiring a thorough evaluation. In this work, we show that such defenses exhibit both shared and system-specific weaknesses. Exposing the vulnerabilities requires following a standard set of evaluation strategies, which we propose in our paper. Furthermore, we show that these vulnerabilities are amplified under DazzlePatch, a novel patch attack that uniquely replaces the borders of the input during the query phase to minimize detection while perturbing the central patch using standard query-based attacks. To ensure compliance with the* $\ell_\infty$ *threat model, the attack restores the original borders in the final iteration, yielding a valid adversarial example within the permissible perturbation budget. Our results demonstrate a substantial reduction in detection rates and a corresponding increase in attack success rates across multiple SDMs. We then show that incorporating input randomisation, such as Random-Resized Cropping (RRC), significantly enhances SDM robustness, reducing attack success rates by up to* $26.5\%$*. These findings suggest that while current SDMs are vulnerable to tailored adaptive attacks, integrating them with additional defense mechanisms may offer improved resilience.*

## 1. Introduction

Black-box attacks are a class of attacks that attempt to compromise the performance of a machine learning system by making changes in the input space without relying on machine learning (ML) model related information [3, 6]. In contrast, white-box methods may rely on ML model parameters, architecture or training dataset to launch an attack [38]. In a black-box setting, the adversary may be limited to only accessing model outputs, such as predicted probabilities or class labels. In more transparent scenarios, adversaries might even access the underlying training dataset [9]. Black-box attacks can further be categorised into transferability and query-based black-box attacks [12]. Transferability (zero-query) attacks attempt to transfer an attack by generating the attack on a model available to an attacker, called the surrogate model, and transfer it to a target model [36].

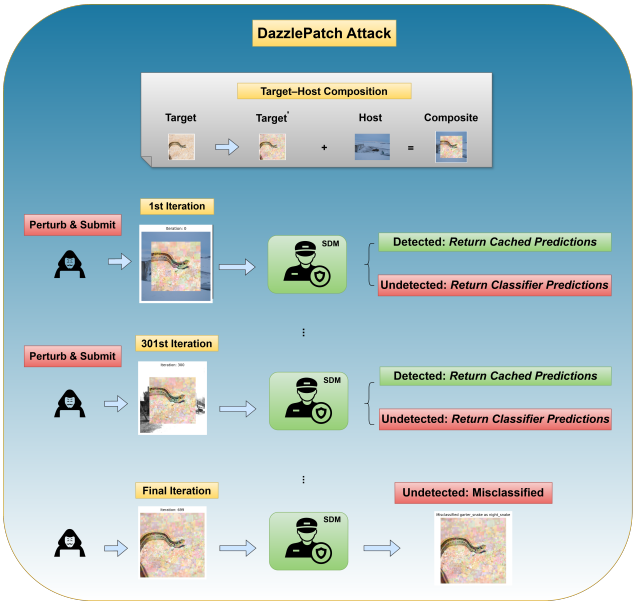


Figure 1. Proposed DazzlePatch replaces the border of a target image to avoid detection from SDMs while allowing optimisation on central patch. The attack state is updated exclusively based on undetected queries. In the final iteration, the target′ image is fully restored with its original borders to satisfy $\ell_p$ constraint.

To mitigate query-based black-box attacks, various

methods have been proposed. The defences can be categorised into two main classes. One method attempts to mitigate them through input or output randomisation strategies [2, 38]; inputs can be randomised by adding noise [38] or transforming them stochastically [4]. The other class of methods, called Stateful Defence Models (SDMs) [12], aim to detect attacks during the querying stage. SDMs maintain a history of queries (or inputs) sent to the model and try to recognise attacks as variations of past queries. Each of the two methods offer trade-offs in success and computational expense.

Although previously shown vulnerable to the so-called adaptive attacks [21], recently the SDMs have presented some remarkable results in defence to an advanced adaptive attack called Oracle-Guided Adaptive Rejection Sampling (OARS) attack [21]. OARS is a state-of-the-art black-box attack that exposed the vulnerabilities of previously proposed SDMs considered to be secure [12, 14, 20, 30]. However, recently proposed SDMs [32, 37, 43] have claimed $0\%$ attack success rate (ASR) against OARS, which is an extraordinary claim that merits further investigation.

Therefore, in this work, we investigate the recently proposed SDMs that claim $0\%$ attack success rate; these are AdvQDet [43], GWAD, [37], and QPA [32]. We find that GWAD [37] and QPA [32] are vulnerable to attacks under the $\ell_\infty$ distance metric. AdvQDet, although resistant to $\ell_\infty$ based attacks, can still be defeated by a new attack presented in this paper termed **DazzlePatch attack**, which achieves an ASR of $75.50\%$ with $\ell_\infty$ perturbation budget of $0.09$ under 22 random restarts (attack seed initializations). When restricted to smaller budgets of $0.05$ and $0.07$, the attack still reaches $39.00\%$ and $54.00\%$ ASRs, respectively. In the original work, AdvQDet was evaluated with a perturbation budget of $0.05$. Similarly, we also show the effectiveness of this attack on QPA and thereby present the attack as a new benchmark for testing stateful defences. Our evaluation thus suggests that stateful detection can be bypassed substantially. However, we then combine SDMs with input randomisation and notice a significant improvement in defence. Therefore, the results indicate that while vanilla SDMs are vulnerable, they can be enhanced with additional defences.

The main contributions of this work are below.

- We evaluate recently proposed Stateful Defence Models (SDMs) and demonstrate that contrary to prior claims of $100\%$ robustness these models remain highly vulnerable under adaptive adversarial conditions.
- We propose a novel patch-based attack as an effective benchmark for defence models and test it against the selected SDMs. The attack takes a fixed-size central patch from the input image and overlays it onto a separate host image, resulting in an output where the inserted patch is clearly framed by the border of the host image. The host image is sourced from a separate dataset, and is selected without repetition. This attack significantly lowers detection rates and improves attack success rates against AdvQDet [43] and QPA [32].
- We show that incorporating input randomisation techniques [4, 38] can enhance the robustness of stateful detectors against query-based black-box attacks.
- In the light of our findings, we introduce a suite of evaluation strategies to comprehensively assess robustness of SDMs and recommend their adoption for future evaluations.

## 2. Related Work

Black-box attacks can be categorised into transferability and query-based black-box attacks [45]. Both categories of attacks can be utilised to compromise any defence type, including adversarial training, SDMs, or input randomisation. However, adversarial training [33] has stood the test of time and has been the most successful defence against white-box and black-box attacks. Nevertheless, results indicate that robustness across popular benchmark datasets must be improved [8, 15]. We discuss this progress below.

### 2.1. Transferability Attacks

Transferability attacks craft adversarial examples on surrogate models and deploy them against target models. Although transferability-based attacks were mitigated by [42] by training an ensemble of models, this defence was subsequently shown to be vulnerable to query-based attacks that utilise probability scores [5, 25] or hard-labels [10, 11].

Although SDMs were not originally proposed to defend against transferability attacks [12], the robustness of SDMs under transferability attacks has not yet been systematically evaluated and remains an open line of work.

### 2.2. Query-based Attacks

Although transferability based attacks hold potential, they too must rely on querying the model several times to increase chances of a successful attack on an SDM. Query-based attacks are usually stronger than transferability-based attacks because the latter relies on surrogate and target model similarity [45]. Therefore, our work focuses on query-based black-box attacks.

Commonly effective defences against these attacks are input/output randomisation [2, 4, 38] or stateful detection [12, 30, 32, 37, 43]. Randomisation based techniques perturb the inputs or outputs of the models to thwart the attack; they have enjoyed varying levels of success defending against query-based black-box attacks. Stateful Defence Models (SDMs), first proposed by [12], aim to detect adversarial examples during query-based black-box attacks by tracking the history of past queries. The investigation in [12] tested the SDM against an adaptive query-

based attack – specifically, a query-blinding attack – where the adversary adapts its strategy to the defence by applying random input transformations to obscure the attack pattern. The attacker's objective is to evade the detector repeatedly while ensuring that predictions for the original and transformed inputs remain consistent, ultimately inducing misclassification. The results showed that the SDM remained resilient even under this adaptive setting. The authors evaluated their defence under the $\ell_\infty$ distance metric, a widely adopted and more stringent perturbation constraint than $\ell_2$ under which achieving robustness is generally more challenging [15]. They further assume that robustness under $\ell_\infty$ will naturally extend to other $\ell_p$ norms [12].

Following their work, Blacklight [30], Perceptual Image Hashing (PIHA) [14] and IIOT-SDA [20] proposed defences that further enhanced the performance of stateful detection defences.

However, all the proposed SDM defences were later shown to be vulnerable to the OARS attack [12]. OARS changes the proposal distribution of the attack. For example, in the case of Square Attack [5] instead of sampling a single square, it samples multiple squares that lead to the least detection. OARS [21] exposed multiple defences and showed that the stateful defences were "not yet secure".

OARS thus intensified the arms race and attracted subsequent works [32, 37, 43] on developing stateful detectors that are shown to be invulnerable to the OARS attack, claiming a $0\%$ attack success rate. This is an extraordinary claim that warrants a careful investigation. We briefly explain the working mechanics of OARS-resistant SDMs and discuss their limitations.

#### 2.2.1. Gradient's Watch Adversarial Detection (GWAD)

GWAD [37] detects attacks by training a classifier on the *Histogram of delta similarities (HoDS)*. The classifier detects attack by classifying a histogram of cosine similarity of differences ($\delta$) between consecutive queries, capturing the signature of adversarial behavior. Essentially, it analyses the query update patterns, that is HoDS, rather than the inputs themselves.

It maintains a bank of queries submitted and uses it to create HoDS. The size of the bank varies depending on the end-user choice. However, the authors show that the optimal size to recognise attack-types is 256.

According to the authors, each attack has its own unique characteristic HoDS. They show empirically that their method generalises across datasets.

**Limitations:**

- The defence has not been evaluated under the $\ell_\infty$ threat model. However, we find that the defence is also vulnerable under the $\ell_2$ threat model.
- It takes a certain number of queries to build HoDS and the mechanism to load HoDS is not addressed in [37]. However, in this work we consider the stronger scenario; one in which HoDS is loaded so it can start detection from the initial stage of the attack.
- GWAD+ is demonstrated to generalise across different datasets, but not comprehensively to unseen attacks.
- The defence has not been evaluated under random restarts of the attack.

#### 2.2.2. Query Provenance Analysis (QPA)

Provenance analysis models relationships among system entities to detect and investigate intrusions [23, 27, 31]. QPA [32] applies this idea to query auditing. Query provenance features are defined by organising historical queries into a provenance graph, where nodes are queries and edges are weighted by similarities between their extracted features

QPA [32] defence uses perceptual image hashing [14] as the default feature extractor to compute similarity between the queries. Upon similarity calculation, QPA constructs graphs where queries with high similarity are grouped into the same graph. QPA then assigns a Provenance Anomaly Score (PAS) to each graph and retains the graphs with the highest score. It calls this phase *statistical analysis*, which is used to filter out suspicious graphs. The graph classifier is then used to flag incoming queries as malicious or benign based on the statistical analysis and graph structure. Malicious queries are rejected.

**Limitations:**

- Although the threat model assumed by the authors allows an adversary to use any query-based black-box attack to cause misclassification, the defence has only been evaluated under $\ell_2$ versions of the query-based black-box attacks. It has not been evaluated under the $\ell_\infty$ versions of most attacks except for Natural Evolutionary Strategies (NES) attack [25].
- The performance of the defence relies on the number of graphs stored in the cache. Detection efficiency can decrease as number of graphs increase. However, authors find number of graphs as 20 to be sufficient.
- The authors argue that in the event of resetting the database, an adversary on average will require 333 days to launch an attack. This depends on the defence's time-to-detect (TTD) which for Square Attack is 17.8 queries on ImageNet dataset [17, 40]. TTD is the number of queries required to get the first detection. However, they assert that the query provenance graph cannot grow infinitely. Therefore, they reset the query provenance graph every 24 hours. An adaptive long-running adversary with an advanced attack will have an opportunity to bypass it.
- The defence has not been evaluated under random restarts of the attack.

#### 2.2.3. AdvQDet

AdvQDet [43] is a detection framework designed to detect adversarial example generation. They proposed ACPT to

fine-tune CLIP image encoder [39]. CLIP image encoder is used for generating embeddings of images. The goal of the proposed fine-tuned CLIP encoder is to generate similar embeddings for two similar images for the purpose of attack detection. Contrastive Learning (CL) is used to train feature extractors [13, 24, 35] in a self-supervised setting. The concept subsequently developed into Adversarial Contrastive Learning (ACL) [26] and Adversarial Prompt Tuning (APT) [44]. The methods combine adversarial training with contrastive learning to design robust encoders. AdvQDet uses Adversarial Contrastive Prompt Tuning (ACPT) to fine-tune the CLIP image encoder based on adversarial contrastive learning (a form of adversarial training) and prompt tokens. The goal of the training is to encourage the encoder to produce similar embeddings for adversarial inputs.

AdvQDet uses a historical embedding bank to store all the embeddings it generates. It also stores their classification predictions in a cache. The embedding of any new query received is then compared with the previous embeddings in the embedding bank. If the cosine similarity score is higher than a threshold, then the query is flagged as an attack. The system then outputs the prediction corresponding to the cached embedding most similar (by cosine similarity) to the query embedding. In terms of storage cost, the authors show that supporting up to 1 million users with each allowed 100 queries will require around 95 GB of storage with float16 precision. AdvQDet is claimed to have strong detection performance with a TTD of 3 against query-based black-box attacks. Moreover, AdvQDet claims zero-shot generalization ability, as it is only trained on ImageNet and demonstrates impressive detection performance across a wide range of datasets and attacks. The authors evaluate their SDM under the $\ell_\infty$ threat model, which is better expected to generalise to other threat models [12].

**Limitations:**

- False detection/collision rate has not been evaluated in the work. It may be possible that the detector flags images belonging to separate classes as an attack, in which case the incorrect prediction is sent back to the user.
- The authors acknowledge storage cost for embedding bank and computational costs for similarity search as a limitation. However, they do not discuss how frequently the embedding bank would have to be pruned in case of memory or storage limitations. They leave it as a future investigation.
- The defence has not been evaluated under random restarts of the attack.

# 3. Evaluating Stateful Defence Models

## 3.1. Threat Model

Given a clean input $x$, the attacker seeks a terminal example $x_{adv}$ such that $\|x_{adv} - x\|_p \leq \epsilon$ and $f(x_{adv}) \neq y$ where $f$ is the black-box classifier, $y$ is the true class, and $p$ can be any metric $\ell_2$, $\ell_0$ or $\ell_\infty$. Only the *terminal* example is constrained; how the attacker searches for $x_{adv}$ (the intermediate queries) is unrestricted. This is the standard adversarial robustness contract under $\ell_p$ [9]. We assume the adversaries can access the model's score outputs (logits) to accomplish this goal.

## 3.2. Evaluation Protocol

Stateful Defence Models (SDMs) that resist the advanced OARS attack still exhibit defence-specific weaknesses. For example, QPA and GWAD report low false-detection rates but require frequent resets of their history banks and were not comprehensively evaluated under an $\ell_\infty$ threat model. Robustness under the $\ell_\infty$ threat model is generally harder to achieve than under $\ell_2$ [15]. By contrast, AdvQDet was evaluated under an $\ell_\infty$ threat model and achieves the smallest TTD, but its false positive rate is unreported.

Furthermore, existing defences have not been systematically evaluated under random restarts (attacks re-initialised with different seeds) of the attack. This setting is critical for assessing the robustness of SDMs, especially given that memory and efficiency constraints often necessitate periodic resets of the historical buffer. Even in systems with a global memory shared across users, a Sybil attacker [19] could launch an attack with one seed, halt upon detection, and then wait for the global memory to expire before initiating a new attack with a different seed. Over time, a persistent adversary could leverage multiple accounts to repeat this process, gradually bypassing the detector.

Similarly, robustness of SDMs to the OARS adaptive attack suggests that a novel attack strategy is required to defeat them. Therefore, we believe that a comprehensive evaluation of the claimed adversarially robust SDMs should be based on the following strategy:

- Evaluate the false detection rate of the defence.
- Evaluate defences under the $\ell_\infty$ threat model to ensure conclusions generalise beyond $\ell_2$.
- Evaluate defences under random restarts to determine if the defences can afford restarting their historical banks.
- Evaluate the defences under different realistic perturbation budgets.
- Evaluate the defences under an adaptive attack designed to bypass detection.

To this end, we have proposed a novel patch attack.

## 3.3. DazzlePatch Attack

**Attack overview.** Our method follows the spirit of query blinding [12]: rather than querying the detector with the target image directly, we submit a *hosted composite* in which the *target core* (a central patch from the target image) is patched into a *uniquely sourced host* image, while the *host borders* (pixels outside the core) come from the host. To

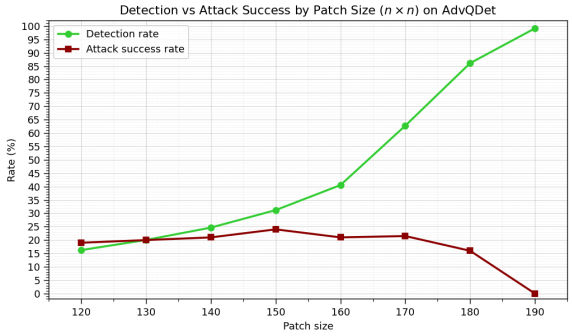


Figure 2. Each patch size is evaluated on a single run of DazzlePatch Attack on AdvQDet.

consistently maintain a low inter-query similarity, we randomly sample a new border without replacement in every query. The composite reduces inter-query similarity and thereby significantly reduces the detection rate of stateful detectors. Any query-based black-box optimiser can then be run *on the target core* while the composite is sent to the system to obtain informative confidence scores from the classifier. The attack is demonstrated in Fig. 1 and the pseudocode is given in supplementary material.

**Constraints.** To reliably produce a valid adversarial example, the attack must satisfy two practical constraints: (i) *Detector blinding:* the composite should substantially reduce the detection rate so that sufficient queries can be issued to optimise the target core. (ii) *De-hosting robustness:* once optimisation completes maximum iterations, we *restore the original target borders* by replacing the host borders with the target's borders (i.e., removing the host content). This is followed to satisfy $\ell_p$ constraint of the threat model.

For the attack to succeed, the adversarial example must stay misclassified after de-hosting, and the detector must fail to identify the target image. De-hosting occurs only at the end to avoid exposing the target's borders to the SDM.

**Confidence-increasing phase.** To meet these requirements, we continue with a *confidence-increasing* phase that explicitly maximises the incorrect-class confidence after the first misclassification on the hosted composite. This creates a robustness margin so that removing the host borders (de-hosting) minimises the chances of reverting the prediction into the correct class.

**Target Patch Size and Position.** The patch size of the target image plays a role in determining both the detection rate and attack success rate as shown in Fig. 2. Using a large patch size will increase the detection rate but may also increase the likelihood that the target remains adversarial after borders are replaced. Finding the right choice depends on the defence. In this work, we use patch sizes of $160 \times 160$ and $170 \times 170$ for the central patch, out of a total image size of $224 \times 224$ on ImageNet.

Patch positioning may also play a role in determining the success of the attack. However, in this work, we always place the target image patch in the center and just change the borders around it. Randomising patch position will make optimisation difficult on the target core because randomisation helps the defender [4, 38].

### 3.4. Experimental Settings

**Metrics.** We use attack success rate (ASR), average hit rate (Avg. Hit Rate), and false positive rate (FP) to measure the performance of the defences. ASR measures the percentage of samples that were misclassified by the defence during an attack run. For attack with restarts, we measure the cumulative ASR, which accumulates successes from previous runs. Samples that were defeated on one run of the attack are not re-evaluated on the next runs of the attack.

Moreover, the average hit rate measures the detection rate of the defence during the entire run over all samples, and the FP rate measures whether an image belonging to one class triggers a false attack alarm, with an image of a different class.

**Datasets and Models.** Dataset and model details are provided in supplementary material.

**Attack configuration.** Stateful defence models were originally proposed to detect an attack during its generation, with an aim of increasing attacker cost to defeat the system. Black-box attacks are usually run with a budget of 10000 queries. If competitive SDMs are installed as a defence, majority of queries will be flagged in the case of an attack. Therefore, an attacker must minimise the queries required to accomplish a successful attack. For this reason, DazzlePatch uses black-box attack optimiser (Square Attack) with a query budget of 700 iterations, which is a smaller budget. On the other hand, since GWAD is significantly vulnerable to vanilla Square Attack, we do not apply DazzlePatch on GWAD. Owing to the low detection rate of GWAD, the Square Attack is allowed to run for 8000 iterations. Furthermore, DazzlePatch attack perturbs the target image using a black-box optimisation procedure to induce misclassification under an $\ell_p$ metric. We find that the success of DazzlePatch-style attacks depends strongly on the optimisation recipe: for the Square Attack [5], inserting multiple squares per iteration substantially strengthens the attack against QPA and AdvQDet. By contrast, results indicate GWAD is already vulnerable to the original Square Attack under the $\ell_\infty$ metric. Interestingly, GWAD detects multi-square Square Attack better than the standard single square Square Attack. Therefore, we apply a multi-square insertion strategy for QPA and AdvQDet, and a single-square strategy when attacking GWAD.

**Perturbation budgets.** Previous works on SDMs have used various perturbation budgets. For example, GWAD

restricts the perturbation budget as $\|x_{adv} - x\|_2/\|x\|_2 \leq 0.1$. They assert that this ensures humans cannot recognise the image distortion. On the other hand, QPA uses multiple $\ell_2$ based attacks and a single $\ell_\infty$ attack and restricts them under a normalised budget of $0.05$. Similarly, AdvQDet evaluates their defence on an $\ell_\infty$ budget of $0.05$.

In this work, we focus on the $\ell_\infty$ metric. Particularly for the DazzlePatch attack, we use $\ell_\infty$ budgets of $0.05$, $0.07$ and $0.09$. The perturbation budget is increased when the attack performance stagnates across random restarts. However, increasing the perturbation budget is not unrealistic as previous works [1, 16, 18, 30, 34] have evaluated classifiers under higher budgets, even on low-resolution datasets such as CIFAR-10 [29]. We show the impact of visual distortion when increasing the perturbation in Fig. 3. The objects in the images are easily recognizable.

Moreover, since GWAD is notably susceptible to the standard Square Attack under the $\ell_\infty$ norm, we include both a small budget $0.01$ and a larger budget $0.05$ in our experiments. We find that the majority of examples generated using Square Attack with an $\ell_\infty$ budget of $0.01$, also satisfy $\|x_{adv} - x\|_2/\|x\|_2 \leq 0.1$.

**Host patch bank.** The *host* image pool was drawn from *Open Images* v7 release [22, 28]. Although the background images can be sampled in various ways, for these experiments we used 993 images annotated as `snowmobile` or `snowplow`; due to environmental similarity their backgrounds are potentially more similar than in random images, allowing the SDMs greater chance of detecting them. However, the results indicate that even a potentially weak attack setting is still successful against the tested defences. In each experiment, we ran a 700-query black-box attack and, at every query iteration, uniquely sourced the host by sampling a distinct image from this pool (without replacement), ensuring no host image was reused within a run. Sampling without replacement decreases the chance of detection.

Moreover, image quality has an impact on detection rate. For example, on a 900-iteration run of DazzlePatch, if the host images are sourced from CIFAR-10 (resized) then the detection rate is $97.32\%$ on AdvQDet. In contrast, the detection rate with source as Open Images is $44.62\%$.

Table 1. Square Attack results on GWAD defence (CIFAR-10) with HoDS size 256 and with a full bank before starting attack optimisation.

| **Metric** (%) | $\ell_\infty = 0.01$ | $\ell_\infty = 0.05$ |
|---|---|---|
| ASR under $\ell_\infty$ | 63.80 | 99.80 |
| ASR under $\ell_2^{normalised} = 0.1$ | 63.70 | 14.90 |
| Avg. hit rate | 8.7 | 0.7 |

# 4. Experimental Results

## 4.1. GWAD

We use Square Attack under the more general $\ell_\infty$ ($\epsilon = 0.01, 0.05$) [5] metric on GWAD, The attack is run for 8000 iterations and on 1000 samples with a $p_{init}$ of $0.8$ and a single square on each iteration.

We show results on GWAD in Tab. 1, which is loaded with initial queries. The optimisation of the attack starts only after 256 iterations (once the bank has loaded). At an $\ell_\infty$ budget of 0.05, the ASR reaches $99.40\%$ with low detectability, while at 0.01 it drops to $63.80\%$. Notably, $63.70\%$ of the 1000 samples also satisfy the original constraint from [37], producing visually imperceptible changes.

## 4.2. Query Provenance Analysis

**Square attack.** QPA [32] does not assume a specific $\ell_p$ norm in the threat model; however, we find that changing the norm makes substantial difference to the attack success rate. With $\ell_2$, Square Attack has only $1.0\%$ ASR against QPA with $\epsilon = 0.5$ and $1.0$; however, with $\ell_\infty$ the picture changes substantially. For a perturbation budget of $0.05$, we find even the default configuration of the Square Attack (1 square) produces a high ASR of $32.5\%$, which jumps up to $41.5\%$ after 4 restarts. However, it does so with a high detection rate of $98.10\%$ and 98.33, respectively. In comparison, increasing the number of squares to 150 – significantly increases the ASR under 4 restarts to $67.50\%$.

**DazzlePatch attack.** Additionally, we launch DazzlePatch attack on QPA to test its robustness, and its results are compared with Square Attack with multiple squares in Tab. 2. DazzlePatch results in a significantly lower hit rate i.e. less than $3.483\%$ as compared to over $17\%$ of Square Attack. While the DazzlePatch has a lower ASR than Square Attack with multiple squares, it still achieves a high ASR ($>50\%$) with a low hit rate. All under a perturbation budget of $0.05$. Raising the perturbation budgets and restarts then increase the ASR to $87.50\%$, which is close to the ASR of Square Attack ($96.00\%$) – asserting the competitive performance of DazzlePatch.

## 4.3. AdvQDet

**False Positive Rate.** AdvQDet has a high false positive detection rate on datasets that have high visual similarity between classes. We show that on German Traffic Sign Recognition Benchmark (GTSRB) [41] it has a high false positive rate on speed limits as shown in Tab. 3. We show the performance on AdvQDet on two different versions. One with cosine similarity threshold set to 0.90 and the other with threshold of 0.95. Further datasets are explored in supplementary material.

**Square attack.** We apply the standard Square Attack to test the base performance of AdvQDet, and find that it

Table 2. Robustness of QPA under adaptive attacks on 200 samples from ImageNet dataset. DazzlePatch achieves competitive ASR.

| | | | **DazzlePatch Attack (Ours)** | | **Square Attack** | |
|---|---|---|---|---|---|---|
| Restart | $\epsilon$ | Num. of Squares | ASR (%) | Avg. Hit Rate (%) | ASR (%) | Avg. Hit Rate (%) |
| 0 | 0.05 | 150 | 34.50 | 2.63 | 59.50 | 17.23 |
| 4 | 0.05 | 150 | 54.50 | 3.48 | 67.50 | 20.13 |
| 9 | 0.07 | 150 | 77.00 | 2.88 | 88.00 | 15.42 |
| 14 | 0.09 | 150 | 87.50 | 2.60 | 96.00 | 10.68 |

only has an ASR of $4.5\%$. All the $4.5\%$ of the samples get misclassified in the first query. SDMs are not designed to defend against such examples, as this is a limitation of the classifier and not the detector. The performance weakens further if we increase the number of squares to 170, with ASR elevating to $14.50\%$.

**DazzlePatch attack.** AdvQDet's minimal TTD (3) necessitates a novel bypass; we therefore propose DazzlePatch, which mounts attacks using Expectation Over Transformation (EOT) [7] with three transformations. The loss is accumulated over every 3 queries (central path perturbation fixed and host border varied). However, within the query budget of 700.

DazzlePatch attack achieves the best performance of $21.00\%$ on the first complete run of the black-box attack, which increases to $39.00\%$ after 9 random restarts under $\ell_\infty$ budget of $0.05$. However, increasing the perturbation to $0.07$ and random restarts to 15 increases the ASR to $54.00\%$. Similarly, increasing the perturbation to $0.09$ and restarts to 21 – shoots the ASR to $75.50\%$. DazzlePatch improves over the Square Attack with multiple squares by $26.50\%$ under 21 restarts and a perturbation budget of $0.09$. The results are given in Tab. 4.

**Random Resized-Cropping.** As proposed by [4], we use a random transformation-based defence against query-based black-box attacks called Random Resized-Cropping (RRC) to increase the defensive abilities of AdvQDet. As the attack relies on replacing the border of the image with a unique image, we find that using RRC crops the border part, thus, increasing the chances of detection. Using RRC results in an increased detection rate, and a lower ASR. The ASR even on the highest perturbation with 22 restarts is $49.0\%$; improving over vanilla AdvQDet by $26.50\%$. However, when the query budget and host images are increased to 5000, EOT is increased to 20, and the central patch size is reduced to 130, the attack performance improves. Under a perturbation budget of $0.05$ and with a single run of the DazzlePatch attack, the ASR increases from $11.00\%$ to $17.00\%$, with an average hit rate of $16.87\%$, suggesting that under RRC, an attacker must expend substantially more queries to achieve a meaningful improvement in ASR.

Table 3. AdvQDet false positive analysis on GTSRB dataset.

| **Label** | **0.90 sim (FP% / Total Samples)** | **0.95 sim (FP% / Total Samples)** |
|---|---|---|
| Label 1 (speed limit 30) | 90.69 / 720 | 31.84 / 200 |
| Label 14 (stop) | 16.41 / 200 | 0.50 / 200 |
| Label 33 (turn right ahead) | 68.09 / 210 | 6.66 / 200 |
| Label 0 (speed limit 20) | 100.00 / 60 | 33.33 / 60 |
| Label 27 (pedestrian) | 40.00 / 60 | 13.33 / 60 |
| Label 22 (bump ahead) | 76.66 / 120 | 40.00 / 120 |
| Label 5 (speed limit 80) | 99.50 / 200 | 72.63 / 200 |

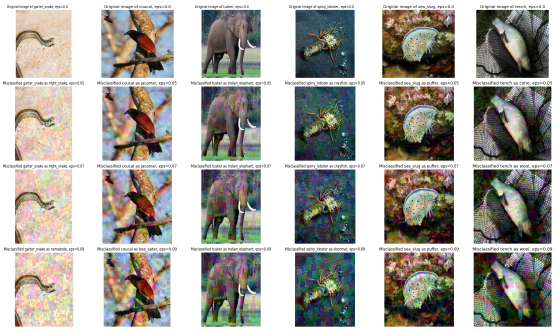

Figure 3. Visualization of images distorted at various levels of perturbations using Square Attack with multiple squares.

### 4.4. Ablation study

**Impact of increasing confidence.** The impact of increasing confidence after a misclassification is induced, contributes significantly to the success of the DazzlePatch attack on AdvQDet. To further demonstrate this, we perform an ablation study in which we do not maximise confidence after a misclassification is achieved during the query phase. We run DazzlePatch on the same seed with and without confidence maximising on the 200 ImageNet samples. We find that without confidence maximisation, the ASR is $13.5\%$, as compared to a ASR of $21.0\%$ with confidence maximisation. This suggests that confidence maximisation increases the chances of a sample staying adversarial after the original borders are restored to satisfy the $\ell_\infty$ constraint.

**Images vs. random patterns in border.** Our experiments on ImageNet had shown that the CLIP encoder in AdvQDet is largely invariant to non-semantic patterns such as checkerboard or noise, and instead focuses on the central patch when computing similarity, helping detection. In con-

Table 4. Robustness of AdvQDet under adaptive attacks on 200 samples from ImageNet dataset. DazzlePatch significantly outperforms Square Attack.

| Attack | | | **Square** | | **DazzlePatch (Ours)** | | | |
|---|---|---|---|---|---|---|---|---|
| Defence | | | **AdvQDet** | | **AdvQDet** | | **AdvQDet + RRC** | |
| Restart | $\epsilon$ | No. of Squares | ASR (%) | Avg. Hit Rate (%) | ASR (%) | Avg. Hit Rate (%) | ASR (%) | Avg. Hit Rate |
| 0 | 0.05 | 170 | 14.00 | 99.71 | 21.00 | 40.55 | 11.00 | 55.90 |
| 4 | 0.05 | 170 | 23.00 | 99.71 | 34.50 | 42.68 | 20.0 | 58.33 |
| 9 | 0.05 | 170 | 24.50 | 99.71 | 39.00 | 54.67 | 21.50 | 69.00 |
| 15 | 0.07 | 170 | 31.00 | 99.71 | 54.00 | 53.79 | 35.00 | 68.03 |
| 21 | 0.09 | 170 | 49.00 | 99.71 | 75.50 | 52.98 | 49.00 | 67.98 |

trast, introducing semantically explainable elements in the borders present greater diversity; this enables the composite image to evade similarity detection. Particularly, the ASR based on a central patch size of 160 with borders as random noise is $5.0\%$ with an avg. hit rate of $99.57\%$.

## 5. Discussion & Limitations

**The failure of OARS attack** AdvQDet, QPA, and GWAD all claim robustness against OARS [21], suggesting that OARS is insufficient for bypassing SDMs. For instance, AdvQDet detects OARS within the first 3 queries and continues to consistently flag the attack for the remainder of the query budget. Such early and persistent detection renders the adapt-and-resample strategy of OARS ineffective, as the attack is unable to adapt in the presence of immediate and consistent detection. In particular, OARS-square achieves an ASR of $0\%$ with an average hit rate of $99.57\%$ against AdvQDet. Out of 700 queries, only the first 3 are undetected, and even these are expended during OARS's internal optimisation process for determining the number of squares.

OARS operates on the assumption that even if some queries are detected, it can adapt its strategy to evade SDMs. However, AdvQDet demonstrates that this assumption does not hold when robust detectors are employed. AdvQDet flags minor perturbation differences between similar images reliably. The robustness of AdvQDet stems from its use of an adversarially trained CLIP encoder. Consequently, bypassing such a robust detector requires a fundamentally different approach. Notably, where the adapt-and-resample strategy of OARS fails, DazzlePatch succeeds by using border replacement and loss maximisation.

**High perturbation budget.** Our proposed attack can defeat AdvQDet under the perturbation assumed in the original work with limited success ($ASR = 39\%$). Achieving higher ASR requires a slight increment in the perturbation budget, which may not be possible on all datasets. However, our work proposes DazzlePatch as a framework to bypass and defeat SDMs. The target patch in our attack could be optimised using various black-box techniques. Works that focus on black-box attacks assume higher perturbation budgets than white-box attacks as black-box attacks are more difficult. We believe that this work will motivate the development of attacks that produce valid adversarial examples with imperceptible distortions.

**Hard-label attacks.** We have performed the attacks in a setting where the defender sends back output scores instead of labels. In the case where the defender only sends hard-labels, our proposed attack could still be utilised by using a hard-label black-box attack on the target image. We leave this extension for future investigation.

## 6. Conclusion

According to our evaluation, QPA and GWAD can be compromised with only the Square Attack under $\ell_\infty$ metric. Similarly, AdvQDet and QPA are also significantly vulnerable to our proposed DazzlePatch attack. Our proposed attack achieves the best performance on AdvQDet, which is the most robust SDM. SDMs face inherent constraints in size and efficiency, which a persistent attacker can exploit. Hence, our proposed attack uses random restarts to significantly degrade SDMs performance. QPA is compromised with only 15 runs (different seeds) of the attack. Similarly, AdvQDet can be significantly compromised in under 22 resets of the historical bank with our attack. Our attack is the only attack to bypass AdvQDet, yet. It is also the first time that defences resistant to OARS attack have been bypassed.

## 7. Acknowledgements

This work was supported by a doctoral grant from Covatic.

## References

[1] Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R Venkatesh Babu. Scaling adversarial training to large perturbation bounds. In *European Conference on Computer Vision*, pages 301–316. Springer, 2022. 6

[2] Manjushree B Aithal and Xiaohua Li. Mitigating black-box

adversarial attacks via output noise perturbation. *IEEE Access*, 10:12395–12411, 2022. 2

[3] Ziad Tariq Muhammad Ali, Ameer Mohammed, and Imtiaz Ahmad. Evaluating adversarial robustness of secret key-based defenses. *IEEE Access*, 10:34872–34882, 2022. 1

[4] Ziad Tariq Muhammad Ali, R Muhammad Atif Azad, Muhammad Ajmal Azad, James Holyhead, Iain Rice, and Ali Shariq Imran. Random transformations to improve mitigation of query-based black-box attacks. *Expert Systems with Applications*, 264:125840, 2025. 2, 5, 7

[5] 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. 2, 3, 5, 6

[6] Maungmaung Aprilpyone and Hitoshi Kiya. Block-wise image transformation with secret key for adversarially robust defense. *IEEE Transactions on Information Forensics and Security*, 16:2709–2723, 2021. 1

[7] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 274–283. PMLR, 2018. 7

[8] Brian R Bartoldson, James Diffenderfer, Konstantinos Parasyris, and Bhavya Kailkhura. Adversarial robustness limits via scaling-law and human-alignment studies. In *International Conference on Machine Learning*, pages 3046–3072. PMLR, 2024. 2

[9] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 1, 4

[10] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1739–1747, 2020. 2

[11] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020. 2

[12] Steven Chen, Nicholas Carlini, and David Wagner. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pages 30–39, 2020. 1, 2, 3, 4

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 4

[14] Seok-Hwan Choi, Jinmyeong Shin, and Yoon-Ho Choi. Piha: Detection method using perceptual image hashing against query-based adversarial attacks. *Future Generation Computer Systems*, 145:563–577, 2023. 2, 3

[15] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. 2, 3, 4

[16] Palma Alessandro De, Durand Serge, Chihani Zakaria, Terrier François, and Caterina Urban. On using certified training towards empirical robustness. *Transactions on Machine Learning Research Journal*, 2025. 6

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[18] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*, 2018. 6

[19] John R Douceur. The sybil attack. In *International workshop on peer-to-peer systems*, pages 251–260. Springer, 2002. 4

[20] Bardia Esmaeili, Amin Azmoodeh, Ali Dehghantanha, Hadis Karimipour, Behrouz Zolfaghari, and Mohammad Hammoudeh. Iiot deep malware threat hunting: from adversarial example detection to adversarial scenario detection. *IEEE Transactions on Industrial Informatics*, 18(12):8477–8486, 2022. 2, 3

[21] Ryan Feng, Ashish Hooda, Neal Mangaokar, Kassem Fawaz, Somesh Jha, and Atul Prakash. Stateful defenses for machine learning models are not yet secure against black-box attacks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 786–800, 2023. 2, 3, 8

[22] Google Research. Open images v7. https://storage.googleapis.com/openimages/web/index.html, 2021. Dataset with 9 million images annotated with image-level labels, object bounding boxes, segmentation masks, and visual relationships. 6

[23] Xueyuan Han, Thomas Pasquier, Adam Bates, James Mickens, and Margo Seltzer. Unicorn: Runtime provenance-based detector for advanced persistent threats. *arXiv preprint arXiv:2001.01525*, 2020. 3

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[25] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018. 2, 3

[26] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in neural information processing systems*, 33: 16199–16210, 2020. 4

[27] Samuel T King and Peter M Chen. Backtracking intrusions. In *Proceedings of the nineteenth ACM symposium on Operating systems principles*, pages 223–236, 2003. 3

[28] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper

Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. In *Dataset release, https://storage.googleapis.com/openimages/web/index.html*, 2017. 6

[29] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Citeseer*, 2009. 6

[30] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Blacklight: Scalable defense for neural networks against {Query-Based}{Black-Box} attacks. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2117–2134, 2022. 2, 3, 6

[31] Shaofei Li, Feng Dong, Xusheng Xiao, Haoyu Wang, Fei Shao, Jiedong Chen, Yao Guo, Xiangqun Chen, and Ding Li. Nodlink: An online system for fine-grained apt attack detection and investigation. *arXiv preprint arXiv:2311.02331*, 2023. 3

[32] 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. 2, 3, 6

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 2

[34] Maung Maung, April Pyone, and Hitoshi Kiya. Encryption inspired adversarial defense for visual classification. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1681–1685. IEEE, 2020. 6

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[36] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 1

[37] 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. 2, 3, 6

[38] Zeyu Qin, Yanbo Fan, Hongyuan Zha, and Baoyuan Wu. Random noise defense against query-based black-box attacks. In *Advances in Neural Information Processing Systems*, 2021. 1, 2, 5

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 4

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[41] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 6

[42] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 2

[43] 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. 2, 3

[44] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *European conference on computer vision*, pages 56–72. Springer, 2024. 4

[45] Meixi Zheng, Xuanchen Yan, Zihao Zhu, Hongrui Chen, and Baoyuan Wu. Blackboxbench: A comprehensive benchmark of black-box adversarial attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2