

BIRMINGHAM CITY
University

On the feasibility of mitigating query-based black-box attacks

By

Ziad Tariq Muhammad Ali

School of Architecture, Built Environment, Computing and Engineering
Birmingham City University

A thesis submitted for the degree of
Doctor of Philosophy

Oct 2026

Supervisor(s): Prof. Raja Muhammad Atif Azad, Dr. Muhammad Ajmal Azad and Prof. Iain Rice

ABSTRACT

Query-based black-box attacks pose a practical threat to deployed machine learning systems because they require only repeated access to model outputs rather than internal knowledge of the target model. The thesis examines the landscape of literature on defences against query-based black-box attacks and focuses on two of the most popular streams of defences. These methods are input randomisation and stateful defence models (SDMs).

Building on this foundation, the thesis re-examines the efficacy of the state-of-the-art methods in randomisation based defences, namely, the random-noise defence and shows that methods based on popular transformations commonly used to augment data, particularly random rotation and random crop-resize, are significantly more effective than the noise-based randomisation for mitigating query-based attacks.

The thesis next examines startling claims of foolproof security made by modern stateful defence models against query-based black-box attacks. The presented work challenges these claims via an effective experimental setup that disproves these claims for the first time. These findings are significant because they demonstrate that overstated robustness claims can create a false sense of security. The thesis also presents a novel, bespoke attack, DazzlePatch, which effectively breaks the most resilient SDM called AdvQDet. DazzlePatch increases the attack success rates (ASR) while decreasing the detection rates. The work therefore contributes towards advancement of adversarial defences by presenting such an effective additional testbed.

In a departure from common practice, the thesis then investigates attack capabilities and defence robustness assuming a threat model where the attacker is highly-resourced. Although this model is computationally demanding for the average experimenter, it remains well within reach of a motivated and resourceful attacker. The results show that the performance of even the best-known defences decreases under such attack scenarios. This prompts the community to rethink the defences of systems exposed to large user queries and puts up a case that defence performances should not merely be evaluated in typical settings, which assume a moderate to low attacker capability.

DEDICATION

To my parents, Tariq Ali and Tehmina Saeed Butt, who dedicated their lives to giving their children the greatest gift they could offer: education.

You invested not only your resources, but your hopes, dreams, and sacrifices into our future. You placed our education above your own comforts and devoted everything you could to opening doors of opportunity for us. To my mother, whose extraordinary focus on our education and relentless encouragement carried us forward; and to my father, whose hard work and support made this journey possible.

To my late grandparents, Muhammad Ali and Muhammad Saeed Butt, whose hard work and courage laid the foundations for pursuit of education in my family.

This achievement is a reflection of your sacrifices and your dreams.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my principal supervisor, Professor Atif Azad, whose exceptional guidance have been central to the completion of this PhD. His expertise, patience, and thoughtful advice helped me navigate every stage of this research. I am particularly grateful for the time and confidence he invested in me. His mentorship has shaped not only this thesis but also my development as a researcher, and I feel truly privileged to have worked under his supervision.

I would also like to sincerely thank my supervisors, Dr. Ajmal Azad and Professor Iain Rice, for their valuable guidance, constructive feedback, and continued support throughout my PhD. Their knowledge and perspectives have made an important contribution to this research.

My sincere appreciation goes to Covatic and Nick Pinks for funding this PhD and making this research possible. Their generous financial support and collaboration provided me with an invaluable opportunity to undertake this doctoral work.

I am also especially grateful to Dr Umar Daraz for establishing the connection between Covatic and Birmingham City University. His efforts played an important role in creating the opportunity from which this PhD emerged, and I greatly appreciate his support.

Finally, I would like to thank Professor Ali Ansari for introducing me to Professor Atif Azad. I appreciate his support in bringing us into contact and for helping establish this valuable academic connection.

Contents

Page

Abbreviations **xiv**

1 Introduction **1**

1.1 Overview of the Thesis 1

1.1.1 Aim 2

1.1.2 Motivation 2

1.1.3 Gap in Knowledge 3

1.2 Research Questions 4

1.3 Key findings 5

1.4 Contribution of the Thesis 6

1.5 Publications 7

1.6 Outline of the Thesis 8

2 A review of defences against query-based attacks **10**

2.1 Background 10

2.1.1 Threat Model 10

2.1.2 Metrics 11

2.1.3 Black-box Attacks 12

2.1.4 Adaptive Attacks 13

2.2 Review Queries 13

2.3 Literature Review 16
2.3.1 Adversarial training 17
2.3.2 Processing-Based Defences 19
2.3.3 Detection-Based Defences 27
2.3.4 Architecture 35
2.4 Conclusion 39

3 Random Transformation is more effective than random noise to mitigate query-based attacks 41
3.1 Introduction 41
3.1.1 Contributions 43
3.2 Black-box attacks 44
3.2.1 Weakness of black-box attacks 44
3.2.2 Adaptive black-box attacks 45
3.3 Preliminaries 46
3.3.1 Threat Model 46
3.3.2 Random Noise Defence 47
3.4 Proposed Methodology 49
3.4.1 Randomised transformations 49
3.4.2 Evaluation 52
3.5 Experimental Settings 54
3.5.1 Datasets and Models 55
3.5.2 Data Augmentations 56
3.5.3 Defences Evaluated 57
3.5.4 Model Prediction 57
3.5.5 Metrics 59
3.6 Experimental Results 59

3.6.1 CIFAR-10 . . . 59
3.6.2 ImageNet . . . 60
3.6.3 Tiny ImageNet . . . 62
3.6.4 Indicative Comparison with RFD and Random Padding-Cropping . 62
3.6.5 Evaluation on hard-label attack . . . 63
3.6.6 Ablation Study . . . 64
3.6.7 Increasing Attacker Noise . . . 65
3.6.8 Statistical Significance Test . . . 70
3.6.9 Conclusion . . . 70

4 Re-evaluating the robustness of Stateful Defence Models 73

4.1 Introduction . . . 73
4.2 Related Work . . . 75
4.3 Evaluating Stateful Defence Models . . . 80
4.3.1 Threat Model . . . 80
4.3.2 Evaluation Protocol . . . 81
4.3.3 DazzlePatch Attack . . . 82
4.3.4 Experimental Settings . . . 86
4.4 Experimental Results . . . 90
4.4.1 GWAD . . . 91
4.4.2 Query Provenance Analysis . . . 91
4.4.3 AdvQDet . . . 93
4.4.4 Ablation study . . . 98
4.5 Conclusion . . . 98

5 Evaluating defences against black-box attacks under a resourceful attacker 100

5.1 Introduction . . . 100
5.2 Input Randomisation under Resourceful Attacker . . . 101

5.2.1 Experimental settings 102
5.2.2 Random Noise Defence under high EOTs 103
5.2.3 Random Rotation Defence under high EOTs 107
5.2.4 Highly confident misclassified samples under RND 110
5.2.5 Practical Feasibility of EOT-2000 Attacks 113
5.3 SDM+Randomisation under Resourceful Attacker 114
5.4 Conclusion 115

6 Conclusions **117**
6.1 Answers to the Research Questions 117
6.2 Contributions to Knowledge 120
6.3 Limitations 121
6.4 Future Work 122
6.5 Implications for Knowledge and Practice 124
6.6 Final Remarks 125

References **126**

List of Figures

2.1 PRISMA flow of the database search. 15

2.2 Taxonomy of defences against black-box attacks 16

2.3 Conceptual framework of black-box attacks and defences 17

3.1 Original image of a Goldfish from the ImageNet dataset. 51

3.2 Random Resize-Cropping (RRC) variants of an image of a Goldfish from the ImageNet dataset. Our preliminary experiments revealed superior performance of 80% cropping value over smaller cropping values. 51

3.3 Random rotation variants of an image of a Goldfish from the ImageNet dataset. 51

3.4 Illustration of how the model prediction is used for random transformation defences in this work. 58

3.5 The attack success rate graph of the randomised defences on square attack with EOT-50 on ResNet-18 model architecture and CIFAR-10 dataset. The attack success rate is measured across the iterations of the attack. 60

3.6 The improvement in robust accuracy of transformation based defences over RND as attacker noise is increased. 66

3.7 Representative SSIM example 1 under increasing attacker noise. 67

3.8 Representative SSIM example 2 under increasing attacker noise. 67

3.9 Representative SSIM example 3 under increasing attacker noise. 68

3.10 Representative SSIM example 4 under increasing attacker noise. 68

3.11 Representative SSIM example 5 under increasing attacker noise. 69

4.1 Proposed DazzlePatch replaces the border of a target image to avoid detection from SDMs while allowing Square Attack to make adversarial changes. In the final iteration, the target′ image is fully restored with its original borders to satisfy ℓ_p constraint. 83

4.2 The relationship between detection rate and attack success rate on each patch size. Each patch size is evaluated on a single run of DazzlePatch Attack on AdvQDet. 85

4.3 Visualisation of images distorted at various levels of perturbations using Square Attack with multiple squares. 97

5.1 Per-class distribution of the randomly sampled 1000 images from the CIFAR-10 dataset. 102

5.2 Per-class samples (a) and per-class percentage of samples (b) made robust by defences such as adversarial training and randomisation. The defensive impact of randomisation reduces as the EOT increases. 104

5.3 Per-class samples (a) and per-class % of samples (b) that remain resilient under various levels of EOTs (RND only). It is evident that randomisation disproportionately effects the classes. Frog and Truck samples are the most effected by randomisation. 106

5.4 Per-class samples (a) and per-class % of samples (b) made resilient by defences such as adversarial training and randomisation. The defensive impact of randomisation reduces as the EOT increases. 108

5.5 Per-class samples (a) and per-class % of samples (b) that remain resilient under various levels of EOTs (only RR improved samples). It is evident that randomisation disproportionately effects the classes. Frog and Truck samples are the most effected by randomisation. 109

5.6 Per-class % of samples that remain resilient under various levels of RND magnitudes. It is evident that randomisation disproportionately effects the classes. 111
5.7 Per-class resilient number of samples under various randomisation techniques. 113

List of Tables

2.1 Summary of existing processing and model architecture defence methods against black-box attacks. 20

2.2 Comparison of processing-based defence techniques against query-based black-box attacks. $\sim$ means partially resistant. 26

2.3 Summary of existing detection based defences against black-box attacks. . 27

3.1 Performance of randomised defences under Square Attack 61

3.2 Performance of randomised defences on CIFAR-10 and ImageNet datasets under confident square attack 62

3.3 Performance of additional randomised defences on first 200 samples 63

3.4 Performance of randomised defences on CIFAR-10 under the hard-label RayS attack 63

3.5 Ablation: multipass vs single pass 64

3.6 SSIM comparison at the highest attacker-noise setting across five representative examples. Higher SSIM means less degradation. 69

4.1 Experimental settings used to evaluate the three stateful defence models. Each defence is evaluated under its original experimental setting to ensure fidelity. 87

4.2 Square Attack results on GWAD defence (CIFAR-10) with HoDS size 256 and with a full bank before starting attack optimisation. 91

4.3 Robustness of QPA under adaptive attacks on 200 samples from ImageNet dataset. DazzlePatch achieves competitive ASR. 92

4.4 AdvQDet false positive analysis on GTSRB dataset. 93
4.5 AdvQDet false positive analysis on ImageNet. 94
4.6 False Positive Analysis on Chinese Traffic Dataset 95
4.7 Robustness of AdvQDet under adaptive attacks on 200 samples from the ImageNet dataset. DazzlePatch significantly outperforms Square Attack. . . . 96

5.1 Comparison of low and high query budget settings for evaluating SDM+Randomisation (RRC) under DazzlePatch attack. 115

Abbreviations

Abbreviation	**Meaning**
AI	Artificial Intelligence
AT	Adversarially Trained/Training
ASR	Attack Success Rate
EOT	Expectation over Transformation
ML	Machine Learning
OARS	Oracle-Guided Adaptive Rejection Sampling
RND	Random Noise Defence
RR	Random Rotation
RRC	Random Resized-Cropping
RT	Random Transformations
SDM	Stateful Defence Model
TTD	Time-to-Detect

Chapter 1

Introduction

1.1 Overview of the Thesis

As capable as modern artificial intelligence systems have become, they remain vulnerable to unexpected failure modes in which carefully crafted inputs can cause them to make incorrect predictions. These inputs, known as *adversarial examples*, are generated through adversarial attacks that exploit weaknesses in machine learning models (Szegedy et al., 2014). In some cases, perturbations that are nearly imperceptible to humans can cause a model to misclassify images of everyday objects, raising concerns about the reliability and security of AI systems deployed in real-world applications such as autonomous vehicles, healthcare, and surveillance.

Among the many forms of adversarial attacks, *query-based black-box attacks* represent a particularly realistic and practical threat model (Carlini et al., 2019). In contrast to the white-box attacks, which assume access to internal information such as model parameters, architecture, or training data, query-based black-box attacks require only access to the model's outputs. By repeatedly querying the target model and observing its responses, an adversary can gradually optimise an adversarial perturbation capable of inducing misclassification. This threat model is especially relevant in modern machine learning deployments, where models are frequently exposed through public APIs or cloud-based services while their internal details remain hidden.

1.1.1 Aim

In this thesis, we investigate the feasibility of mitigating query-based black-box attacks through two complementary lines of defence: randomisation-based defences and stateful defence models (SDMs). These strategies are the current state-of-the-art to defend against query-based black-box attacks (Qin et al., 2021; Chen et al., 2020c). Randomisation-based defences introduce stochastic transformations to the input or inference process, making it more difficult for an attacker to obtain reliable optimisation signals across successive queries. By disrupting the consistency of the model's responses, these methods aim to reduce the effectiveness of iterative attack strategies. In contrast, stateful defence models seek to detect malicious query behaviour by maintaining historical information about previously submitted inputs. Since query-based attacks typically rely on issuing many closely related queries, stateful approaches attempt to identify and block such suspicious patterns before a successful attack can be completed (Chen et al., 2020c).

The selection of these two defence paradigms is motivated by their fundamentally different approaches to the problem. Randomisation seeks to weaken the attacker's optimisation process, whereas stateful defences aim to identify the attack itself. By studying both approaches, this thesis provides a broader understanding of the strengths and limitations of current defensive strategies against query-based black-box attacks and contributes to the ongoing effort to improve the robustness and security of machine learning systems.

1.1.2 Motivation

The growing adoption of AI systems in online platforms and social media amplifies the risks posed by adversarial examples. Modern AI models are increasingly used to analyse images and provide information about products. For instance, a user may upload an image

of a common over-the-counter medication and ask an AI assistant whether the product is safe when used as directed. Under normal circumstances, the model should correctly identify the medication and provide an appropriate response. However, an adversarially perturbed version of the same image may cause the model to misclassify the medication as a different substance with significantly different properties. Such misclassification could lead to incorrect or misleading information being communicated to users (Zhang et al., 2026). When AI-generated responses are widely shared across social media platforms, the effects of these errors may be amplified, contributing to the spread of misinformation and reducing trust in AI-assisted decision making.

As machine learning systems continue to be integrated into safety-critical and information-sensitive applications, protecting them against adversarial manipulation becomes increasingly important. Consequently, this thesis investigates defensive strategies against query-based black-box attacks, with a particular focus on randomisation-based defences and stateful defence models. By evaluating the effectiveness and limitations of these approaches, this work aims to contribute towards the development of more robust and trustworthy machine learning systems.

1.1.3 Gap in Knowledge

We follow two main lines of investigation in this thesis: the first focuses on improving existing defence techniques, while the second explores stronger adaptive attacks to provide a more rigorous evaluation of these defences.

Despite the growing body of work on defending machine learning systems against query-based black-box attacks, important gaps remain in the literature. First, although randomisation-based defences such as Random Noise Defence (RND) have been established

as the state-of-the-art (Qin et al., 2021), additive noise can degrade image quality, yet no prior work has rigorously compared it against non-noise, transformation-based alternatives (e.g., random rotation and random resize-cropping) that are already widely used for data augmentation. It therefore remains unclear whether such transformations can mitigate query-based attacks as effectively as noise-based approaches while preserving classification performance.

Second, recently proposed stateful defence models (SDMs) that are resistant to the OARS adaptive attack have reported extraordinary robustness claims, including up to 100% detection in some settings. However, the adversarial machine learning literature has repeatedly shown that apparently robust defences can fail once evaluated under stronger, defence-aware adaptive attacks (Athalye et al., 2018). At the time this thesis began, these specific claims of complete robustness had not yet been subjected to such rigorous, adaptive scrutiny.

Third, while input randomisation and SDMs represent two fundamentally different mitigation strategies, no prior work had evaluated whether combining them yields additional robustness, nor had either paradigm, alone or combined, been tested under a threat model in which the attacker is permitted a substantially larger query budget and computational resources than is typical in the literature. Consequently, the robustness of these defences against realistic, highly-resourced adversaries remained largely unknown.

1.2 Research Questions

To address the challenges associated with defending against query-based black-box attacks, this work investigates the following research questions:

- What are the key approaches proposed for countering query-based black-box attacks?

- Given that randomisation is intended to mislead attackers and reduce the effectiveness of their queries, can we use non-noise-based inference-time randomisation to improve mitigation of query-based attacks?
- Can detection-based defences against query-based black-box attacks provide complete robustness, as claimed by prior work?
- How do the defences against black-box attacks scale down as the attacker becomes significantly more resourceful?

1.3 Key findings

We evaluate state-of-the-art inference-time randomisation defences that rely on additive noise to mitigate adversarial attacks. Our findings indicate that their defensive performance can be surpassed by employing randomised image transformations commonly used for data augmentation during model training. A key limitation of noise-based perturbations is that they can degrade image quality, which can make their applications in certain domains difficult. Motivated by this limitation, we investigate whether non-noise-based transformations can provide more robust protection against query-based attacks while preserving classification performance. We also find that noise-based defences are significantly more vulnerable to adaptive attackers than transformation-based defences.

Furthermore, we question recent claims of 100% detection against query-based black-box attacks by proposing a novel adaptive attack strategy that can evade stateful defence models (SDMs). Existing SDMs operate under the assumption that query-based attacks rely on submitting inputs that are highly similar in the representation space. We demonstrate that this assumption does not necessarily hold, as an adversary can generate successful adversarial examples without issuing highly similar queries as per the popular measures of

similarity adopted by the most successful SDMs.

In addition, our findings reveal that the two major defence paradigms—input randomisation and SDMs—are not orthogonal; rather, they can be combined to provide improved robustness against adversaries that adapt their attacks based on the capabilities of the defender also known as adaptive attackers. However, the field of adversarial defences has not yet discovered any method that is foolproof and our proposals are no exceptions. Nevertheless, they are rigorously tested against the state-of-the-art benchmarks and they improve upon the previous results in various settings described in the thesis. Through testing in rigorous settings, we find that adaptive attackers with abundant resources may still be able to compromise these defences.

1.4 Contribution of the Thesis

We address the research questions by making the following contributions:

- We review the literature systematically and taxonomy of the principal defences proposed against query-based black-box attacks, showing that the field is organised mainly around randomisation-based mitigation and stateful detection.
- We demonstrate empirically that transformation-based randomisation, particularly random rotation and random resize-cropping, is more effective than noise-based randomisation for mitigating query-based black-box attacks, staying considerably more resistant to adaptive attackers than noise-based randomisation.
- We rigorously evaluate the 100% robustness claims concerning some recent SDMs and find that those claims are not true. In one instance where the defence method indeed proved robust to common attacks, we challenged it by proposing a novel attack

called DazzlePatch that effectively defeated the defence. We therefore contribute an additional testbed for the evolving body of SDMs.

- We identify limitations of existing defences against query-based black-box attacks by showing that the benefits of randomisation diminish as the attacker resources increase. We further show that although combining SDMs with the proposed randomisation-based defences improves robustness, much like any defence method known today, this combination is not foolproof. In general, increasing the capability of the attacker to very high levels challenges these defences as well. We thus posit that the community must consider these high intensity attack scenarios as well while presenting the claims of robustness.

1.5 Publications

The following research papers have been published as a result of work towards the thesis:

- Ali, Ziad Tariq Muhammad, Raja Muhammad Atif Azad, Muhammad Ajmal Azad, Iain Rice, Umar Daraz, Ali Shariq Imran, and James Holyhead. "On Evaluating Stateful Defence Models against Query-Based Black-Box Attacks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 809–818. 2026.[1]

- Ali, Ziad Tariq Muhammad, 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 (2025): 125840.[2]

[1] Code: https://github.com/ziyadt1995/evaluating_stateful_defences
[2] Code: https://github.com/ziyadt1995/RT_to_mitigate_QBBAs

1.6 Outline of the Thesis

An outline of each chapter now follows.

Chapter 1 - Introduction. This chapter introduces the thesis and outlines the chapters along with their descriptions.

Chapter 2 - A review of defences against query-based attacks. This chapter provides a systematic literature review on defences proposed to counter query-based black-box attacks. It categorises the defences that have been proposed to counter query-based attacks. The key categories outlined are pre-processing techniques and detection techniques.

Chapter 3 - Random Transformation is more Effective than Random Noise to Mitigate Query-Based Attacks. This chapter evaluates randomisation-based methods in rigorous settings and provides experimental evidence that random transformations outperform random noise-based defences–the prior state-of-the-art for mitigating query-based attacks. The work in this chapter is published in Ali et al. (2025).

Chapter 4 - Re-evaluating the robustness of Stateful Defence Models As identified in Chapter 2, as the other main class of defences against query-based black-box attacks, in this chapter we re-examine the extraordinary claims of 100% robustness of some of the recently proposed SDMs. We find that these claims do not hold under comprehensive evaluation by modifying existing attacks. For the defence that remained resilient against existing attacks, we introduce a novel adaptive attack, DazzlePatch, which successfully bypasses its protection. Consequently, this work contributes a new benchmark for the rigorous evaluation of current and future SDMs. The work in this chapter is published in Ali et al. (2026).

Chapter 5 - Evaluating defences against black-box attacks under a re-

sourceful attacker.

This chapter investigates the robustness of randomisation-based defence strategies against query-based black-box attacks in the presence of increasingly resourceful adaptive adversaries. Using *Expectation over Transformation (EOT)*, which estimates model outputs by averaging predictions over a large number of randomised transformations (Athalye et al., 2018), we show that randomisation either as input alone or when coupled with detectors can substantially increase the query complexity and computational effort required to mount a successful attack. Nevertheless, our evaluation further demonstrates that, given sufficiently large computational resources and query budgets, adaptive adversaries can progressively recover attack effectiveness. These findings highlight that randomisation provides a meaningful increase in adversarial cost, while emphasising the importance of evaluating such defences under high-resource threat models to establish realistic robustness guarantees.

Chapter 6 - Conclusion and Future Work. This chapter concludes the thesis by summarising the findings and giving directions for future research. While the thesis advances the state-of-the-art in the various defence methods, adversarial defence remains an area of continuous improvement as the arms race between the defence and attacks continues.

Chapter 2

A review of defences against query-based attacks

In this chapter, we conduct a systematic literature review on the works proposed to counter query-based black-box attacks and lay out a taxonomy of the approaches used in this line of defence. Various studies were collected and categorised based on their working methodology. We find two key strategies for defending against query-based attacks, namely processing and detection-based defences. While the focus of this thesis is on the defences, we briefly review the black-box attacks that have prompted the need for the defences.

This chapter addresses **RQ1**, "What are the key approaches proposed for countering query-based attacks?" The chapter is organised as follows, Section 2.1 gives information about the preliminaries, Section 2.2 shows how literature review was performed and Section 2.3 describes the defences proposed against query-based black-box attacks.

2.1 Background

2.1.1 Threat Model

Consider a machine learning classifier f that classifies inputs $\mathbf{x} \in \mathbb{R}^d$ into labels $y_{\text{pred}} \in \mathbb{Z}_k$, where $f : \mathbb{R}^d \to \{1, \ldots, k\}$, d is the dimension of the input, and k is the number of classes.

Let $y \in \mathbb{Z}_k$ denote the ground-truth label of a given input. The objective of the classifier is to correctly classify all input instances such that $f(\mathbf{x}) = y$. In contrast, the objective of the adversary is to construct an adversarial example $\mathbf{x}_{\text{adv}} = \mathbf{x} + \delta$ such that $f(\mathbf{x}_{\text{adv}}) \neq y$ (Carlini et al., 2019). The adversarial input $\mathbf{x}_{\text{adv}}$ is perturbed in such a way that it remains close to the original input $\mathbf{x}$ according to an ℓ_p distance metric,

$$\|\mathbf{x} - \mathbf{x}_{\text{adv}}\|_p \leq \epsilon,$$

where p may correspond to ℓ_2, ℓ_0, or ℓ_∞. The perturbation budget ϵ constrains the maximum allowable change to the input while ensuring that the adversarial example remains sufficiently similar to the original sample yet causes the classifier to misclassify it.

2.1.2 Metrics

In general, the metrics used in this thesis are the following:

$$\text{Clean Accuracy (CA)} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(f(x_i) = y_i) \times 100, \tag{2.1}$$

$$\text{Robust Accuracy (RA)} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big(f(x_i^{\text{adv}}) = y_i\big) \times 100, \tag{2.2}$$

$$\text{Attack Success Rate (ASR)} = \frac{\sum_{i=1}^{N} \mathbb{I}(f(x_i) = y_i) \cdot \mathbb{I}\big(f(x_i^{\text{adv}}) \neq y_i\big)}{\sum_{i=1}^{N} \mathbb{I}(f(x_i) = y_i)} \times 100. \tag{2.3}$$

where N denotes the total number of test samples, x_i is the i-th clean input, x_i^{adv} is its corresponding adversarial example, y_i is the ground-truth label, $f(\cdot)$ denotes the classifier, and $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the condition is true and 0 otherwise.

Clean Accuracy (CA) measures the percentage of clean, unperturbed test samples that are classified correctly by the model. It reflects the model's predictive performance under normal operating conditions.

Robust Accuracy (RA) measures the percentage of adversarial examples that are still classified correctly. It is therefore a direct measure of the model's robustness to adversarial perturbations, with higher values indicating greater resilience.

Attack Success Rate (ASR) measures the proportion of adversarial attacks that successfully cause misclassification among samples that were originally classified correctly. A higher ASR indicates a more effective attack, whereas a lower ASR indicates a more effective defence.

2.1.3 Black-box Attacks

Black-box attacks can succeed either by repeatedly querying the target model or by exploiting the transferability of adversarial examples. Query-based attacks can be broadly categorised into score-based and hard-label/decision-based attacks. Score-based attacks, such as the state-of-the-art Square Attack (Andriushchenko et al., 2020), leverage the model's output scores (e.g., logits or class probabilities) to estimate gradients that lead in the adversarial direction and thereby optimise perturbations that induce misclassification. In contrast, hard-label attacks, such as the *RayS Attack* (Chen et al., 2020b), can only access the final predicted class label and must optimise adversarial perturbations without confidence information. Since many real-world machine learning services expose only the predicted class rather than confidence scores, hard-label attacks are generally considered the more realistic threat model. Unlike their white-box counterparts that utilise the model access to compute gradients, query based attacks use zeroth-order optimisation techniques that are derivative-free approaches.

In contrast, another class of black-box attacks called *transferability attacks* generate adversarial examples using surrogate models and then transfer them to the target model.

A *surrogate model* is a substitute model that is accessible to the attacker and is used to craft adversarial examples that can subsequently transfer to the target model due to the similarity of their learned decision boundaries (Zheng et al., 2023). However, the attack success depends significantly on the similarity between target and surrogate models (Zheng et al., 2023), making success challenging. Therefore, investigating query-based attacks which have been proven to be more successful and difficult to defend against (Zheng et al., 2023) is critical.

2.1.4 Adaptive Attacks

Another class of attacks that is increasingly used in the modern literature is titled *adaptive attacks*. An *adaptive attack* is one in which the adversary explicitly tailors the attack strategy to circumvent a specific defence mechanism. For example, many inference-time randomisation defences can be weakened using *Expectation Over Transformation (EOT)* (Athalye et al., 2018), which averages the model outputs over multiple randomised transformations to reduce output variance and obtain more reliable gradient estimates. Likewise, in the context of detection-based defences, the *Oracle-guided Adaptive Rejection Sampling (OARS)* (Feng et al., 2023b) attack adapts existing query-based black-box attacks by modifying their perturbation proposal distributions, enabling them to evade stateful detection mechanisms while maintaining attack effectiveness. To add to this line of adaptive attack, we also propose a novel attack on SDMs called the DazzlePatch attack, which will be introduced in Chapter 4.

2.2 Review Queries

The study followed a rigorous Systematic Literature Review (SLR) approach, adhering to the guidelines proposed by the PRISMA statement (Preferred Reporting Items for Systematic

Reviews and Meta-Analyses) (Moher et al., 2009; Khan et al., 2026). The PRISMA flow as shown in Figure 2.1 illustrates the structured selection of studies.

We performed the following query on IEEE Xplore and Scopus engines, 'black-box' AND 'query' AND 'defense' AND 'adversarial attack', and retrieved 134 and 160 documents, respectively. This search was performed on 9 June 2026. We focus on works that relate to mitigating query-based black-box attacks. Furthermore, we used the following prompt "bring me all research papers that propose defense against query-based black-box attacks" on *Google Scholar Lab* that did not return any more relevant works other than the ones already found. We focus specifically on works that were proposed to counter query-based black-box attacks in computer vision. We found 25 relevant papers that deal with proposing defences against query-based adversarial attacks in vision using the database search. We also used snowballing (Wohlin, 2014) on seminal works by checking papers that cite them (Chen et al., 2020c), (Pang et al., 2020) and (Qin et al., 2021), and found 6 additional relevant papers.

Based on the search, 294 documents were identified and after duplication removal, 172 of them were excluded because they were unrelated to query-based attacks. 50 full-text articles were read and assessed for eligibility to our topic, and 25 of them were excluded because they were not proposed to defend against query-based attacks. In contrast, 25 were selected because they were directly related to defending against query-based attacks.

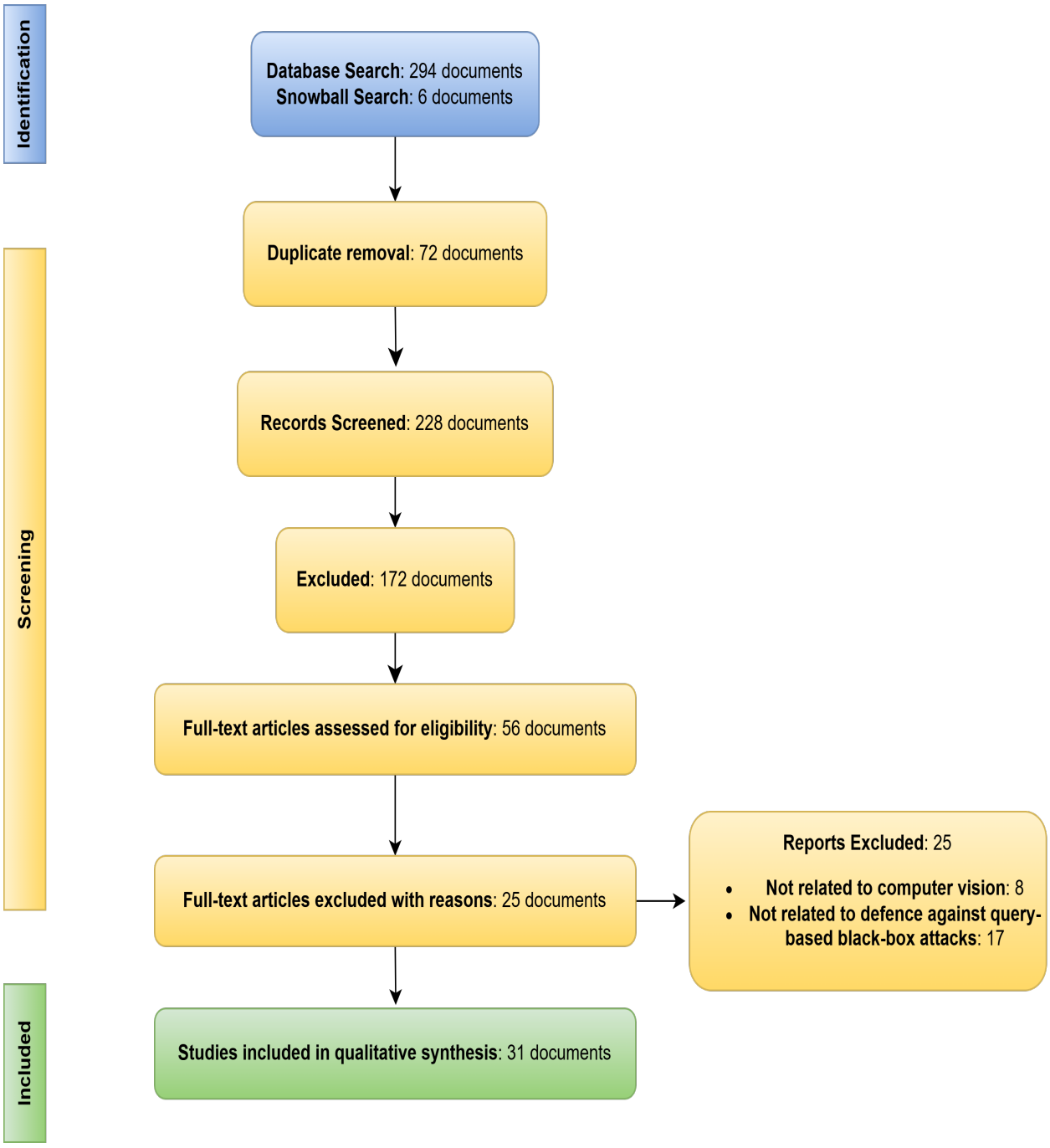


Figure 2.1: PRISMA flow of the database search.

2.3 Literature Review

We provide the taxonomy of the methods proposed that work effectively against query-based attacks as shown in Figure 2.2. The works can be categorised into adversarial training, processing, detection-based, and model-architecture.

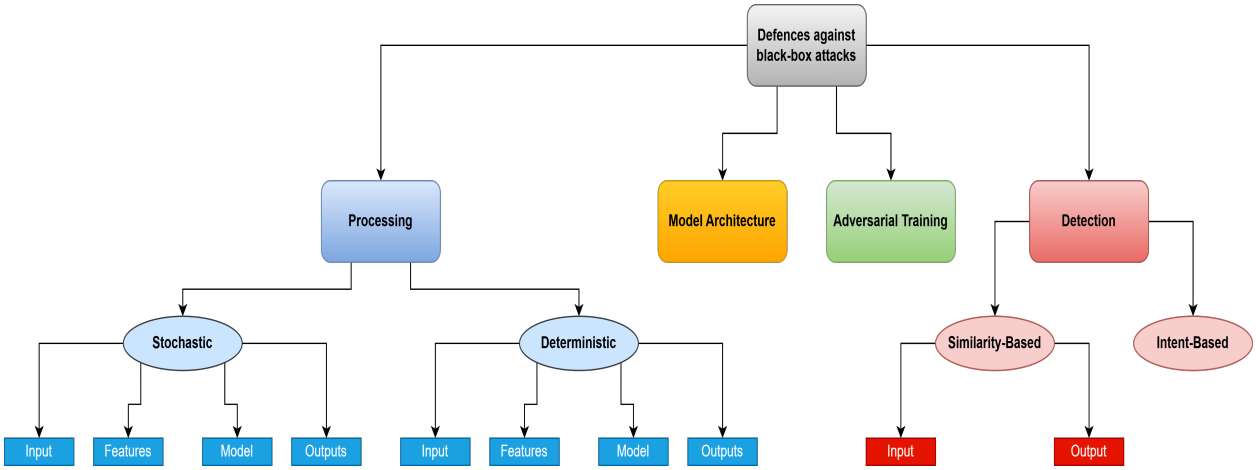


Figure 2.2: Taxonomy of defences against black-box attacks

Adversarial training counters white-box attacks, which operate under the assumption that model parameters are available to the attacker. Robustness against white-box attacks naturally extends to black-box attacks, as the white-box setting represents the stronger adversarial threat model. However, samples that remain susceptible to white-box attacks generally continue to be vulnerable in the black-box setting as well (Andriushchenko et al., 2020). In contrast, processing techniques, model architecture enhancements, and detection-based methods have been specifically developed to mitigate black-box attacks. Accordingly, these defence strategies form the primary focus of this literature review. We provide the conceptual framework of the black-box attacks and defences in Figure 2.3.



Figure 2.3: Conceptual framework of black-box attacks and defences

2.3.1 Adversarial training

Adversarial training is the procedure of enhancing the robustness of deep neural networks (DNNs) by training them on images that are perturbed using the standard white-box attacks (Madry et al., 2018). White-box attacks use gradient-based methods to generate adversarial

examples. The idea is that the once the model is trained on such adversarial examples it will become resilient to unseen adversarial examples - enhancing the underlying robustness of the model. Since the robustness gains enhance the underlying "true robustness" of the model, the model also becomes more resilient to black-box attacks (Madry et al., 2018).

Adversarial training was initially introduced by Goodfellow et al. (2015), who generated adversarial examples using the Fast Gradient Sign Method (FGSM). Although this approach improved robustness against simple attacks, it remained vulnerable to stronger iterative attacks such as the Basic Iterative Method (BIM). To address this limitation, Madry et al. (2018) proposed adversarial training based on Projected Gradient Descent (PGD), demonstrating substantial robustness improvements on the MNIST and CIFAR-10 datasets. This PGD-based adversarial training remains one of the most effective empirical defence strategies. Nevertheless, despite its effectiveness, adversarial training continues to face several challenges, including robust overfitting (Rice et al., 2020), reduced performance on clean data and vulnerability to black-box attacks.

Although adversarial training improves robustness, robust overfitting remains a major challenge. On CIFAR-10, Madry et al. (2018) achieved approximately 48% robust accuracy compared to 87% clean accuracy. Schmidt et al., 2018 suggested that overcoming this limitation requires significantly more training data. Using 500,000 additional pseudo-labelled images, Carmon et al., 2019 increased robustness to around 63%. Further improvements were obtained through data augmentation by Rebuffi et al., 2021 and advanced training techniques by Wang et al., 2023, reaching 70.69% robust accuracy. Nevertheless, a substantial gap between robust and natural accuracy persists on CIFAR-10.

Despite the improvements brought by adversarial training, the technique is intended to improve white-box robustness. Improvements in black-box robustness is a bonus. However, adversarial training has limitations. Firstly, the method is still far from achieving full

robustness; gains even on toy datasets such as CIFAR-10 plateau around 50-70% (Croce et al., 2021). Secondly, it incurs a significant computational overhead. Thirdly, the performance on clean examples is compromised and lastly it remains vulnerable to black-box attacks. This motivates the need for techniques specialised to enhance robustness against black-box attacks. Such techniques may provide resilience against practical attacks without the limitations of adversarial training. Therefore, the focus of our literature review is on defences proposed specifically for resilience against black-box attacks.

2.3.2 Processing-Based Defences

Processing based defences aim to disrupt attack optimisation and counter attacker perturbations by modifying the model's response or internal computation. These techniques increase variance in gradient estimation (Qin et al., 2021). The attacker relies on gradient estimation to find adversarial directions. However, variance in gradient estimation disrupts the attack from reaching the boundary of the sample.

The processing-based defences can be broadly categorised into *input-processing*, *feature-processing*, and *output-processing* approaches. Input-processing defences transform the input before it is passed to the model, whereas feature-processing defences modify the model's intermediate latent representations to reduce the effectiveness of adversarial perturbations. In contrast, output-processing defences operate on the model's outputs, such as logits or softmax probabilities, to mitigate the information available to an attacker. The summary of the works along this line of defences is given in Table 2.1.

Table 2.1: Summary of existing processing and model architecture defence methods against black-box attacks.

Work Title	**Defence Type**	**Limitations**
How to robustify black-box ml models? a zeroth-order optimization perspective (Zhang et al., 2022)	Denoising-Input	EOTs
Adversarial attack on attackers: Post-process to mitigate black-box score-based query attacks (Chen et al., 2022)	Randomisation-Output	Hard-label attacks
Counter-Samples: A Stateless Strategy to Neutralize Black-Box Adversarial Attacks (Bokobza et al., 2025)	Randomisation-Input and Deterministic Output Processing	Upper-bounded by success of Noise Defences
Mitigating black-box adversarial attacks via output noise perturbation (Aithal et al., 2022b)	Randomisation-Output	Hard-label attacks & EOTs
It is double pleasure to deceive the deceiver: disturbing classifiers against adversarial attacks (Zago et al., 2020)	Randomisation-Output	Hard-label attacks & EOTs
Boundary defense against black-box adversarial attacks (Aithal et al., 2022a)	Randomisation-Output	Hard-label attacks & EOTs
Random transformations to improve mitigation of query-based black-box attacks (Ali et al., 2025) (Ours)	Randomisation-Input	EOTs
Random noise defense against query-based black-box attacks (Qin et al., 2021)	Randomisation-Input	EOTs
Closing the gap: achieving better accuracy-robustness tradeoffs against query-based attacks (Zimmer et al., 2024)	Randomisation-Input	EOTs
Defense against query-based black-box attack with small Gaussian-noise (Zhu et al., 2022)	Randomisation-Input	EOTs
Understanding the robustness of randomized feature defense against query-based adversarial attacks (Nguyen et al., 2024)	Randomisation-Features	EOTs
Exploring non-additive randomness on ViT against query-based black-box attacks (Gu et al., 2023)	Randomisation-Model	Architecture Specific and EOTs
Can Model Randomization Offer Robustness Against Query-Based Black-Box Attacks? (Vo et al., 2025)	Randomisation-Model	EOTs
Divergent Twins Fencing: Protecting Deep Neural Networks Against Query-based Black-box Adversarial Attacks (Farshadfar et al., 2025)	Model Architecture	EOTs
Towards Robust Synthetic Aperture Radar Classification: Counteracting Black-Box Adversarial Attacks (Wang et al., 2025a)	Model Architecture	Hard-label attacks
Combating adversaries with anti-adversaries (Alfarra et al., 2022)	Model Architecture	Hard-label attacks
A Novel Method to Mitigate Adversarial Attacks Against AI-as-a-Service Functionality (Tuna et al., 2024)	Model Architecture	EOTs
AlignFix: Fixing Adversarial Perturbations by Agreement Checking for Adversarial Robustness against Black-box Attacks (Nirala et al., 2025)	Model Architecture	EOTs
Dashed Line Defense: Plug-And-Play Defense Against Adaptive Score-Based Query Attacks (Fu et al., 2026)	Deterministic Output Processing	Hard-label attacks

2.3.2.1 Input-processing

Input-processing techniques aim to disrupt the attacker by applying mitigation techniques on the model inputs.

Denoising Defence: Zhang et al. (2022) propose a denoising-based defence for protecting machine learning models against black-box adversarial attacks in privacy-preserving settings. The central idea is to remove adversarial perturbations from the input before it is processed by the target model, thereby reducing the effectiveness of the attack. Unlike conventional denoising approaches, the defence assumes that the target model is completely inaccessible, even to the defender. Consequently, the denoiser cannot be trained using gradients from the target model.

To address this challenge, the authors introduce **ZO-AE-DS** (**Z**eroth-**O**rder **A**uto **E**ncoder-based **D**enoised **S**moothing), which combines certified denoised smoothing with zeroth-order optimisation. Since the target model's parameters are inaccessible, the denoiser is trained using gradient estimates obtained solely through model queries. However, zeroth-order gradient estimation becomes increasingly noisy and computationally expensive in high-dimensional input spaces. To overcome this limitation, the method prepends an autoencoder to the target model, projecting inputs into a lower-dimensional latent space where gradient estimation is considerably more accurate and efficient. The denoiser is then trained in this latent space before the reconstructed input is passed to the target classifier.

Experimental results on image classification benchmarks (CIFAR-10, STL-10, and ImageNet) and image reconstruction (MNIST) demonstrate that ZO-AE-DS significantly outperforms conventional zeroth-order training methods. On CIFAR-10, for example, the proposed approach increases certified robust accuracy from approximately 5% to 35% at an ℓ_2 perturbation radius of 0.5, while achieving robustness that is competitive with several

white-box first-order defence methods. Even though there is improvement in robust accuracy, there is still a large room for improvement.

Input-noise defences: The core idea of input noise defences is to inject Gaussian noise into model inputs at inference time to disrupt the attacker's gradient estimation. The methods include *Random Noise Defence (RND)* (Qin et al., 2021), *Small Noise Defence (SND)* (Byun et al., 2022), *Adaptive Noise Defence* (Jia et al., 2023) and the *Gaussian Noise Defence (GND)* (Zhu et al., 2022). While all four methods follow the same principle, *SND* advocates injecting lower-magnitude noise to better preserve clean accuracy, whereas *Adaptive Noise Defence* adjusts noise dynamically based on attack likelihood. It increases the noise magnitude when suspicious behaviour is detected.

In addition, *RND* demonstrates that the ratio between the defender's noise magnitude and the attacker's perturbation budget is a critical factor in determining defensive effectiveness. The authors further show that *RND* remains effective against several query-based black-box attacks, including adaptive adversaries employing Expectation over Transformation (EOT) with 10 samples. In fact, RND has become the state-of-the-art defence in defending against black-box attacks because when combined with adversarially trained models, it leads to substantial improvements in robust accuracy (Qin et al., 2021; Zimmer et al., 2024) .

Low-confidence input randomisation: Zimmer et al. (2024) propose a defence technique that applies input randomisation techniques such as random resize-cropping, *RND*, or JPEG compression only on samples where the output confidence is lower than a certain threshold. They show that this technique which applies randomisation only on low-confidence samples is enough to give robustness gains over a vanilla model. They show that RND outperforms other techniques. However, this defence is designed to complement input randomisation defences. Moreover, the method is not tested under EOT and therefore its efficacy

is not tested as robustly as RND, which remains the widely used benchmark in such line of defences.

2.3.2.2 Feature-processing Defences

Instead of manipulating the raw pixel values of the input image, feature-based processing techniques modify the feature representations of the model to disrupt the attacker.

Randomised Feature Defence: *Randomised Feature Defence (RFD)* (Nguyen et al., 2024) injects stochasticity into intermediate feature representations rather than the input image. They select the magnitude of the noise depending on its impact on the performance of the model on clean (natural) images. However, the results reported in this thesis show that it is only partially effective in improving mitigation of query-based attacks. Moreover, the results also show that EOT can compromise the performance of the defence.

Non-Additive Randomness: *Non-Additive Randomness* (Gu et al., 2023) is an architecture-based random processing defence designed to mitigate query-based black-box attacks. Rather than injecting noise into the input, the method introduces stochasticity within the model by randomly permuting or dropping the positional encodings associated with the patch embeddings in a Vision Transformer (ViT). This internal randomisation causes the model's predictions to vary across repeated queries. Results show superior performance to other methods such as SND (Byun et al., 2022). However, the defence is not tested under EOTs as the authors claim that EOTs require a large number of queries.

2.3.2.3 Output-processing

Unlike the input processing methods, the output-processing techniques aim to disrupt the attacker by applying mitigation techniques on the model outputs. However, output-processing

struggle to defend against hard-label attacks (Chen et al., 2022), which is not the case with input processing methods.

Adversarial attack on attackers (AAA): Chen et al. (2022) introduce *Adversarial attack on attackers (AAA)* that deliberately manipulates the returned output logits so that the attacker sees a misleading loss trend. The methodology post-processes the logits at inference time by solving a small optimisation problem that moves the output toward a target loss value. The paper designs periodic "attractor" intervals in loss space and then reshapes the apparent loss curve-using either a linear or sine-based design-so that greedy score-based attackers are guided toward the wrong update direction. The results show that this method can defend against score-based attacks. However, the main limitation is that *AAA* is not able to defend against hard-label attacks, which is acknowledged by the authors in the paper.

Output noise perturbation: Aithal et al. (2022b) and Zago et al. (2020) propose a defence that adds noise directly to the model's output for every query. The goal is to corrupt the gradient signal used by black-box attackers. The paper's main contribution is its theory: it analyses the signal-to-noise ratio of the attacker's estimated gradient and derives how the required query count grows as a function of the standard deviation of the defender's noise. The core argument is that even very small output noise can be strongly amplified when attackers rely on tiny output differences to estimate gradients. Empirically, under constrained perturbation budgets, the method reduces the success of query-based attacks while causing little clean-accuracy degradation when the noise is kept small. However, the method is not tested under a strong adaptive attacker with EOT.

Boundary Defence: Aithal et al. (2022a) refine the output noise idea by applying Gaussian noise only to likely boundary samples rather than to all queries. The methodology is based on the observation that query-based attacks, especially decision-based ones,

usually need to probe low-confidence points near the decision boundary. The defence therefore uses the model's confidence to detect such boundary-region queries and injects white Gaussian noise into their logits only when the confidence is below a threshold. This makes the defence more targeted than plain output-noise perturbation and gives it a better clean-accuracy/robustness trade-off, because high-confidence benign samples are left untouched.

The results reveal that RND+AT can outperform the defence under square attack. Furthermore, the defence is neither tested under an EOT adaptive attacker nor under the ℓ_∞ distance metric.

Counter Samples: Bokobza et al. (2025) propose a defence that is conceptually different from the others and is not really a noise-on-output defence. Its methodology is to respond to each query with a counter-sample: a defender-generated, slightly modified version of the query that is optimised towards the current predicted class. The paper argues that this exploits an attacker-defender asymmetry: the attacker only has black-box query access, while the defender has white-box access and can perform a targeted optimisation per query.

In practice, the paper says even one PGD-style step per query is enough to distort the attacker's local landscape, reduce or misdirect gradients, and significantly raise attack failure rates, while only minimally affecting clean inputs. Its biggest limitation is practicality: unlike simple noise injection, it requires the defender to run an optimisation step for every incoming query, which may be expensive for large-scale or latency-sensitive deployments.

Furthermore, it is vulnerable to hard-label based attacks. To address this, the authors introduce a small gaussian noise in the input to make it resilient to hard-label attacks which simply reduces the defence to RND.

Dashed Line Defence: Fu et al. (2026) propose *dashed line defence* which is a runtime, output-manipulation defence designed specifically for score-based black-box query

attacks. Although the work is non-peer-reviewed, we add it for the sake of completion. It belongs to the class of plug-and-play post-processing defences, meaning it does not modify the underlying model or require retraining; instead, it alters the information returned to the attacker at inference time.

Despite its advantages, the defence is like many runtime defences that manipulate outputs, it primarily targets score-based query attacks, since these attacks depend heavily on the loss or probability values returned by the model. In fact, the paper does not evaluate performance against hard-label attacks.

Table 2.2: Comparison of processing-based defence techniques against query-based black-box attacks. $\sim$ means partially resistant.

Defence	**Resistant to Hard-label Attacks?**	**Resistant to Score-based Attacks?**	**Resistant to EOT?**
Input Randomisation	✓	✓	$\sim$
Feature Randomisation	✓	✓	$\sim$
Output Randomisation	×	✓	$\sim$

Processing-based defences primarily use randomisation to mitigate query-based attacks. The major types of processing defences can be classified into input-level, feature-level and output-level. We compare these techniques based on their vulnerabilities in Table 2.2. All processing-based defences provide some level of protection against black-box attacks. However, the differences in performance among these defences are generally marginal, as will be demonstrated in Chapter 3.

Table 2.3: Summary of existing detection based defences against black-box attacks.

Work Title	**Defence Type**	**Limitations**
Stateful detection of black-box adversarial attacks (Chen et al., 2020c)	Detection - Input Similarity	OARS
Adaptive Clustering Buffer-Based Detection (Alshanqiti et al., 2022)	Detection - Input Similarity	Adaptive Attacks
IIoT deep malware threat hunting: from adversarial example detection to adversarial scenario detection (Esmaeili et al., 2022)	Detection - Input Similarity	OARS
PIHA: Detection method using perceptual image hashing against query-based adversarial attacks (Choi et al., 2023)	Detection - Input Similarity	OARS
Malprotect: Stateful defense against adversarial query attacks in ml-based malware detection (Rashid et al., 2023)	Detection - Input Similarity	OARS
Blacklight: Scalable Defense for Neural Networks against Query-Based Black-Box Attacks (Li et al., 2022)	Detection - Input Similarity	OARS
AdvQDet: Detecting query-based adversarial attacks with adversarial contrastive prompt tuning (Wang et al., 2024)	Detection - Input Similarity	DazzlePatch
Mind the Gap: Detecting Black-box Adversarial Attacks in the Making through Query Update Analysis (Park et al., 2025)	Detection - Input Similarity	Square Attack
Query Provenance Analysis: Efficient and Robust Defense against Query-based Black-box Attacks (Li et al., 2025)	Detection - Input Similarity	DazzlePatch and Square Attack
Advmind: Inferring adversary intent of black-box attacks (Pang et al., 2020)	Intent-based	Sybil Attack
Detecting Black-box Model Probing Attacks Through Probability Scores (Wang et al., 2025b)	Detection - Output Similarity	Benign Query Injection
MalShield: Enhancing Android Malware Detection with Stateful Defense against Query Attacks (Zhang et al., 2025)	Detection - Input Similarity	Benign Query Injection?

2.3.3 Detection-Based Defences

Detection-based defences attempt to identify adversarial attacks during the query process by monitoring sequences of queries. Detection-based defences can either be similarity-based or intent-based. Similarity-based defences check whether the queries in history are similar to each other to detect an attack. Intent-based defences go a step further and infer the target class that the adversary is trying to move the sample towards. They both share the same goal of detecting attacks. The summary of the works is given in Table 2.3.

2.3.3.1 Intent-Based detection defences

Pang et al. (2020) introduce *AdvMind*, designed to infer the adversary's intent by predicting the target class toward which an input sample is being manipulated. The defender is considered successful if the intended target class can be identified before the adversary completes the generation of the adversarial example. While *AdvMind* demonstrates promising capabilities in anticipating adversarial objectives, it is developed under the assumption of a single adversary operating in isolation. Consequently, it can be susceptible to a *Sybil attack* (Douceur, 2002), in which a malicious entity creates and coordinates multiple fake identities or adversarial agents to influence the system's observations and decision-making process. Such coordinated behaviour can obscure the true adversarial intent, potentially reducing the effectiveness of the defence mechanism.

2.3.3.2 Similarity-based detection defences

Similarity-based detection defences are founded on the premise that query-based attacks typically involve the submission of a sequence of closely related inputs or outputs. As these queries tend to occupy neighbouring regions in the representation space, their similarity can be leveraged to identify and flag potentially malicious activity. These methods maintain history banks of queries to perform detection.

Encoder-based detectors using average similarity with past queries

These similarity-based detection defences generate latent representations of incoming inputs and compare them against previously observed queries by measuring their average similarity.

Stateful Detection of Black-Box Adversarial Attacks (OSD): Chen et al.

(2020c) introduced stateful detection of black-box adversarial attacks also referred to as the original stateful detector *(OSD)*. *OSD* detects query-based adversarial attacks by monitoring and analysing the sequence of queries submitted to the machine learning service. Because it is a similarity-based detector, *OSD* maintains a historical bank of queries, and applies similarity encoding to queries for assessing whether an attack is being generated. If the attack generation process is detected then the authors propose to block or cancel the account.

They also propose an adaptive attack called the 'query-blinding' attack, which attempts to apply randomised processing techniques such as random-cropping, random rotation and random gaussian noise on the images to help evade detection. Despite this adaptive strategy, the authors report that their defence remains robust against query-blinding attacks.

OSD uses a pretrained similarity encoder to convert each image into a latent space vector. *OSD* flags the query as adversarial if the average latent similarity (computed for the submitted query), exceeds a certain threshold for the user. However, because *OSD* maintains query histories on a per-user basis, it is inherently vulnerable to Sybil attacks, where an adversary distributes queries across multiple identities to avoid accumulating sufficient similarity for detection (Douceur, 2002).

Adaptive Clustering Buffer-Based Detection: Alshanqiti et al. (2022) propose a buffer-based adversarial attack detection framework that operates as a firewall in front of an image classifier. Each incoming image is stored in a fixed-size historical buffer that maintains recent queries. To reduce the computational cost of comparing a new image against all buffered samples, the authors employ a Self-Organising Map (SOM) to cluster buffered images into groups of similar instances. When a new image arrives, it is assigned to its best-matching SOM cluster, and similarity comparisons are performed only against images within that cluster using distance measures such as Euclidean, Manhattan, cosine, and ℓ_∞ norms. A distance score is then computed and compared with a predefined threshold T;

images whose similarity patterns indicate suspiciously small variations from previous queries are flagged as adversarial, while the remaining images are forwarded to the classifier.

The SOM is continuously updated as new samples enter the buffer, allowing the system to adapt dynamically to incoming data. Experimental evaluation on MLP, VGG19, and ResNet50 models against multiple white-box and black-box attacks demonstrated that the adaptive clustering buffer can provide an additional layer of protection with limited performance overhead, although the achieved detection accuracy was approximately 50% on average, suggesting that the approach is more suitable as a complementary defence mechanism rather than a standalone solution. The defence has not been tested against adaptive attackers.

Hashing-based detectors

Instead of using encoders, hashing-based similarity detection defences aim to generate hash representations of the inputs for comparison and detection of malicious activity. The idea is that if two images generate a hashes that sufficiently overlap, then it must be a part of an attack.

Blacklight: Li et al. (2022) propose *Blacklight*, a defence that detects query-based black-box attacks using probabilistic content fingerprints. The authors argue that existing similarity measures are either impractical or inaccurate for this task: direct comparison of queries using pixel-wise or ℓ_p distances incurs prohibitive computational overhead, while perceptual hashing methods produce excessive false positives by matching naturally similar benign images. Blacklight addresses these limitations by constructing compact probabilistic fingerprints that efficiently preserve the similarity between adversarial queries while remaining scalable The proposed defence stores probabilistic fingerprints in a database to detect attacks. The experimental results show that major black-box attack algorithms such as (Chen et al., 2020a) and (Maho et al., 2021) are detected efficiently - reducing the attack

success rate to 0%. The complete detection rates were later challenged by Feng et al. (2023b).

PIHA (Perceptual Image Hashing): Choi et al. (2023) further improve detection through hashing by using perceptual hashing instead of general hashing. The idea is that perceptual hashing produces similar hashes for images that are slightly perturbed, whereas general hashing produces different hashes for slightly perturbed images that does not help to detect attacks. Perceptual hashes are robust to small brightness or noise perturbations. Experimental evidence shows superior performance to *blacklight* and *OSD*.

SDMs resistant to OARS

Here we present the similarity-based detection defences that are resistant to an adaptive attack called *Oracle-Guided Rejection Sampling*. *OARS* was proposed to expose the vulnerability in similarity-based defences such as *blacklight*, *OSD* and *PIHA*. However, following this work, new defences that are resistant to OARS claim perfect detection rates on query-based attacks.

Gradient's Watch Adversarial Detection (GWAD): GWAD (Park et al., 2025) detects attacks by training a classifier on the *Histogram of delta similarities (HoDS)*. The classifier detects attack by classifying a histogram of cosine similarity, which is generated based on the perturbation differences (δ) between consecutive queries, capturing the signature of adversarial behaviour. Essentially, it analyses the query update patterns, that is HoDS, rather than the direct 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.

Furthermore, the authors show that GWAD is vulnerable to benign-query injection

attacks, in which an adversary deliberately submits benign queries to displace adversarial queries from the query banks, thereby reducing the likelihood of detection. To mitigate this weakness, they propose a screening mechanism that identifies and filters benign queries before they are stored in the query banks.

Query Provenance Analysis (QPA): Provenance analysis models relationships among system entities to detect and investigate intrusions (Han et al., 2020; King et al., 2003; Li et al., 2023). QPA (Li et al., 2025) 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 extracts a feature representation using a pretrained feature extractor and subsequently generates a perceptual hash using PIHA (Choi et al., 2023).

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.

AdvQDet: AdvQDet (Wang et al., 2024) was proposed to address the limitations of earlier stateful detection methods, which rely on handcrafted similarity measures that often struggle to distinguish benign queries from adversarial optimisation trajectories. Instead, AdvQDet learns a robust similarity representation by fine-tuning the CLIP image encoder (Radford et al., 2021) using Adversarial Contrastive Prompt Tuning (ACPT). The fine-tuning serves as adversarial training and its goal is to encourage the encoder to produce similar embeddings for adversarial inputs.

To detect attacks, AdvQDet maintains a historical embedding bank containing the

embeddings of previously submitted queries together with their predicted labels. Each incoming query is encoded using the fine-tuned CLIP encoder and compared against the embedding bank using cosine similarity. If the similarity exceeds a predefined threshold, the query is identified as part of an ongoing adversarial attack. Rather than forwarding the query to the target classifier, AdvQDet returns the cached prediction associated with the most similar historical embedding, thereby disrupting the attack optimisation process.

The authors report a Time-to-Detect (TTD) of three queries against query-based black-box attacks while demonstrating strong zero-shot generalisation across multiple datasets and attack algorithms despite training only on ImageNet. Furthermore, they show that the memory requirements remain practical, estimating that supporting one million users with a history of 100 queries each requires approximately 95 GB of storage using float16 precision. The framework is evaluated under the ℓ_∞ threat model, which is generally expected to transfer more effectively to other threat models than detectors trained under alternative perturbation constraints (Chen et al., 2020c).

Output-Based Detectors

Wang et al. (2025b) propose to detect query-based black-box attacks by using output probabilities. Unlike the methods that detect similarity across input query features, these defences compare the top-2 output probabilities (for a multi-class classification) produced by the queries. They show that query-based attacks can be detected through top-2 probability scores as they form distinct patterns unlike benign queries.

For detection, the authors show that either Transformer or Bidirectional Long Short-Term Memory (BiLSTM) can be used. They show that the defence can also detect OARS using this strategy. However, their method is vulnerable to benign-query injection - an attack which compromised GWAD as well. In this attack, the attacker randomly injects benign queries between submission of attack queries. The results show that BiLSTM is

significantly compromised using this method whereas the accuracy of the transformer is also reduced.

Malware Detection

Techniques used to detect adversarial samples in image domain have also been applied to detect query-based attacks on malware classification. These methods employ the same techniques but for a different application domain. These techniques generally convert raw bytecodes from extensions to image forms with pixel values. The images can then be used for classification or manipulated to launch adversarial attacks.

IIOT-SDA: Esmaeili et al. (2022) apply the stateful query analysis in the context of *Industrial Internet of Things (IIoT)*. They used a dataset made for robust malware detection. The encoder in this work uses Mahalanobis distance instead of Euclidean distance used by Chen et al. (2020c). The work claims superior attack detection abilities than *OSD*. This technique was also shown to be vulnerable to *OARS* (Feng et al., 2023b).

Malprotect: Rashid et al. (2023) propose *malprotect* to detect query-based attacks. *malprotect* generates a score vector composed of threat indicators. Threat indicators rely on a range of criteria, primarily grounded in anomaly detection techniques. These include evaluating the autoencoder reconstruction loss of incoming queries, measuring distances between queries, analysing shared feature patterns, and assessing whether query characteristics align with expected distributions. Based on the accumulated query history, each indicator produces a score that reflects the likelihood of anomalous or adversarial behaviour. It shows superior performance compared to other detection techniques such as *blacklight* and *OSD*.

Malshield: Zhang et al. (2025) propose *malshield* which works similar to malprotect for malware detection but applies a dynamic sliding window rather than a fixed-length sliding window. Experimental results show near-perfect accuracy of the defence as compared to

other stateful techniques such as blacklight and QPA (Li et al., 2025).

Stateful defence models encompass multiple strategies, including encoder-based detectors that measure the average similarity between current and previous queries, hash-based detectors that generate query fingerprints, and intent-based detection methods. All of these stateful defence models are vulnerable to adaptive attacks such as Sybil and OARS, while output-based detectors are susceptible to the Benign Query Attack. However, as we have seen, some SDMs demonstrate resistance to adaptive attacks such as Sybil and OARS. This motivates the investigation of these stateful defence models under novel adaptive attack settings, which is explored in Chapter 4.

2.3.4 Architecture

Architecture-based defences aim to make enhancements to the model architecture to improve mitigation of query-based attacks. These techniques may use model randomisation or tweak alignment of the models to enhance defensive capabilities. However, their main limitation is the computational overhead they introduce, which may make their deployment difficult. Nevertheless, the main focus of this thesis is on processing-based and detection-based defences because they are the most established branches of defences in this field. However, for the sake of completion we also include architecture based defences in the review as it is an emerging field. The works along this line are also summarised in Table 2.1.

Combating adversaries with anti-adversaries: Alfarra et al. (2022) propose a defence mechanism against adversarial attacks by introducing an anti-adversary layer that can be prepended to an existing neural network classifier. The key idea is that adversarial attacks typically push an input toward a decision boundary to reduce the model's confidence in the correct label. The proposed method counteracts this by generating a small pertur-

bation in the opposite direction that maximises the classifier's confidence in its predicted label.

For an input x, the model first predicts a pseudo-label using the base classifier, then computes a perturbation by minimising the classification loss with respect to this predicted label through a few iterations of signed gradient descent. The resulting perturbed input is then passed to the original classifier for prediction.

The approach is training-free, meaning it can be applied to already trained models without retraining, and experiments on CIFAR-10, CIFAR-100, and ImageNet show improved robustness against black-box and white-box attacks while maintaining similar clean accuracy. However, the method has not been evaluated on hard-label attacks.

Model Randomisation: Vo et al. (2025) investigate whether model randomisation can improve robustness against query-based black-box adversarial attacks, where attackers iteratively query a model and analyse its responses to construct adversarial examples. The work is unpublished. However, we add it for the sake of completion. The authors observe that such attacks rely on the assumption that the model parameters remain static across queries, allowing attackers to estimate gradients or search directions from successive responses. To break this assumption, the paper proposes a defence mechanism that randomly samples a model from a diverse ensemble for each query, thereby injecting uncertainty into the responses.

Instead of adding noise to inputs or features, the method randomises in the function/parameter space by training a set of diverse yet well-performing models and randomly selecting a subset to generate predictions. The authors also introduce a training objective that encourages both model diversity and individual model performance, ensuring the ensemble remains accurate while producing varied outputs.

However, the approach has several limitations. The defence requires training and maintaining multiple models, which increases computational and memory overhead compared to single-model defences. Because predictions depend on random model selection, the same input can produce different outputs across repeated queries, which can reduce reliability and require repeated evaluations to ensure stable predictions. The defence has also not been tested under adaptive attackers.

Model Retraining: Tuna et al. (2024) propose a defence mechanism designed to protect machine learning models deployed in AI-as-a-Service (AIaaS) environments. The core idea is to leverage uncertainty estimation during inference to detect suspicious inputs that are likely to be adversarial. When an input is submitted, the model first produces a prediction and computes the associated aleatoric uncertainty using Monte Carlo dropout–based Bayesian approximation. If the uncertainty value exceeds a predefined threshold, the system performs a temporary one-time update of the model weights in a direction that minimises the uncertainty for that specific input. The updated model then generates a new prediction, after which the system reverts back to the original model parameters. Additionally, if the updated prediction differs from the original prediction, the input can be flagged as a potential adversarial sample and stored for later adversarial training.

The results show that the method significantly reduces attack success rates while maintaining almost unchanged accuracy on clean data. However, the method is tested under low perturbation budgets and is not evaluated under adaptive attacks such as EOT. Furthermore, the defence introduces additional computational overhead and latency, since uncertainty estimation and temporary model updates must be performed during inference.

Models alignment: *AlignFix* (Nirala et al., 2025) pairs a naturally trained model (weakM, texture-biased) with an adversarially trained model (strongM, shape-biased), exploiting the fact that these models rely on such different features that attacks rarely fool

both simultaneously. When an input arrives, it nudges the input one step toward weakM's predicted class using a combined loss from both models, then accepts the corrected prediction if strongM now agrees with weakM, otherwise falling back to strongM's original output, effectively expanding strongM's decision boundary without needing access to attack internals.

The defence outperforms prior defences including AAA across score-based query attacks, decision-based attacks, and transfer attacks - the latter being a category most defences neglect entirely. The main limitations are a roughly $5\times$ inference slowdown due to the extra forward and backward passes.

Synthetic Aperture Radar (SAR): Wang et al. (2025a) prepend a lightweight layer to the classifier that is optimised at inference time to distort gradient information available through model outputs. By recalibrating this layer so that the gradient inferred from output scores is orthogonal to the true loss gradient, the defence aims to disrupt score-based black-box attacks that rely on gradient estimation from queries. Importantly, the modulation is constrained to remain close to the identity (1), ensuring logit consistency and high clean accuracy while misleading attackers into following ineffective update directions.

The defence relies on local gradient obfuscation that is effective only under limited query budgets. Furthermore, the proposed work does not evaluate their defence under any hard-label attack. Since a hard-label based attack does not use score outputs and only relies on the classification label, it is unlikely to be impacted by this defence as this defence explicitly aims to only change the output score slightly rather than changing the classification.

The authors in the paper criticise hard-label-based attacks for consuming a large number of queries. The justification provided is weak as many black-box attacks rely on a high number of queries to become successful. Consequently, the method should be viewed as a practical query-throttling defence rather than a principled robustness mechanism.

Divergent Twins: Farshadfar et al. (2025) introduce a defence that trains two models, which are used by randomly alternating between them. This defence aims to disrupt the black-box attacker with the alternation and also by introducing noise to the output if increasing the number of models is not possible. The work trains the two models by encouraging their outputs to be different for the same inputs, which is accomplished by introducing gaussian noise to the input, and using a loss function that encourages the cosine similarity between the two model outputs to be low. The results show that the defence can block the natural evolutionary strategies (NES) attack (Ilyas et al., 2018) to an ASR of 39.36%.

However, the authors did not evaluate the defence in the presence of EOTs. Since EOTs average predictions over the outputs from different queries, any gradients obfuscated by the usage of alternating models will be reduced in the presence of EOTs. Moreover, the defence adds computational overhead as it uses multiple models.

2.4 Conclusion

This chapter reviewed the defences proposed against query-based black-box attacks and organised them into the main strands of the literature. This chapter investigates the following research question: *What are the key approaches proposed for countering query-based black-box attacks?* Based on the existing literature, three major classes of defence are identified: input randomisation, stateful defence models (SDMs), and architecture-based defences. Among these, input randomisation and SDMs emerged as the two most established and influential lines of defence at the time this thesis began, which is why they form the central focus of the remainder of the thesis. These two approaches also represent fundamentally different mitigation strategies: randomisation attempts to disrupt the attacker's optimisation process,

whereas SDMs attempt to detect the attack through query history.

The literature review also showed why these two areas warrant closer scrutiny. Among the stateful detection techniques, recent OARS-resistant defences report extremely strong robustness claims, including 100% detection in some settings. Such claims demand careful evaluation, particularly because prior work in adversarial machine learning has repeatedly shown that apparently robust defences can fail under stronger adaptive attacks (Athalye et al., 2018). If these methods are to be relied upon in practice, their robustness must be tested under more demanding and defence-aware evaluation settings.

Likewise, the review identified randomisation-based defences as the leading mitigation strategy on the preprocessing side, with random noise defence (Qin et al., 2021) and randomised feature defence (Nguyen et al., 2024) standing out as key reference points. Architecture-based defences also emerged in the review, with recent work exploring model randomisation, alignment-based methods and retraining strategies. However, this line of work introduces additional computational overhead making deployment challenging. For that reason, the central focus of the remainder of this thesis is on evaluating the limitations and reliability of input-randomisation and OARS-resistant SDMs against query-based black-box attacks.

The investigation of processing defences reveals input and feature randomisation as the most effective strategies to mitigate black-box attacks. The performance of related processing strategies such as output-noise suffers in mitigating hard-label attacks. Therefore, we aim to improve over the state-of-the-art randomisation based methods in the next chapter.

Chapter 3

Random Transformation is more effective than random noise to mitigate query-based attacks

Randomisation has been observed as an effective technique to mitigate black-box attacks. However, determining the type of randomisation that is most effective at countering black-box attacks with minimal impact on clean performance is important in particularly sensitive domains. For example, in certain domains, random noise may not be a first choice as it may degrade image quality. Similarly, certain randomisation techniques may be more vulnerable to adaptive attackers. Therefore, it is important to investigate new randomisation strategies, determine whether they can outperform the state of the art, and stress test them under adaptive attackers.

This chapter addresses **RQ2**, "Can we use non-noise based inference-time randomisation to improve mitigation of query-based attacks?"

3.1 Introduction

In this chapter, we aim to explore whether new randomisation strategies can improve the mitigation of query-based black-box attacks beyond the current state-of-the-art. Earlier we

have discussed that randomisation can be utilised in various ways to mitigate black-box attacks. Here, we investigate whether randomised transformations can outperform existing defences and remain effective under demanding evaluation conditions.

Random Noise Defence (RND) (Qin et al., 2021) adds gaussian noise to incoming inputs with the aim to disrupt the attacker from gaining valuable attack optimisation directions. RND has emerged as the benchmark to improve upon in query-based black-box attacks, and we use it as the primary baseline for evaluating our proposed innovations (Zheng et al., 2023). The proposals to defend against black-box attacks that we test in this chapter are two randomised transformations: Random Resize-Cropping (RRC) and Random Rotation (RR). Such transformations are used to augment training data to improve the invariance of models to changes in the input images that do not change their classification, and hence improve model generalisation; however, while these transformations have been previously explored for defending against white-box attacks (Raff et al., 2019; Xie et al., 2017) on which they were unsuccessful (Sitawarin et al., 2022a), the contribution of this work is to investigate their integration with adversarial training and evaluate their effectiveness against query-based black-box attacks.

Despite the strong standing of RND, previous work shows that RND still leaves room for improvement because it has achieved limited robustness on popular datasets. For example, it achieves a robust accuracy of only 14.8% on Tiny ImageNet (Deng et al., 2009) with a PreActResNet18 model, and 53.5% and 18.1% robust accuracy with a ResNet-50 model on CIFAR-10 and ImageNet datasets, respectively. Therefore, we propose randomised transformations to improve over the RND defence. The defences are tested under a resourceful adaptive attacker (EOT) and the results show significant improvement of randomised transformations over noise-based defences.

Consequently, the results reported in this chapter are significant because it has been

argued that the more a model becomes invariant to randomisation the more adversarially vulnerable it becomes in white-box settings (Gao et al., 2022) and if the model is variant to the randomness of the defence then it achieves a lower clean accuracy. However, our results further emphasise the effectiveness of randomisation as a defence against query-based black-box attacks, as the trade-off between the drop in clean accuracy vs. a gain in robust accuracy is not symmetric (large gain in robust accuracy for a low drop in clean accuracy). These results are the main contributions of the chapter.

3.1.1 Contributions

To summarise, the main contributions of this chapter are as follows.

1. We find that randomised transformations are more effective than randomised noise at mitigating query-based black-box attacks. While these transformations have been studied for defending against white-box attacks (Raff et al., 2019; Xie et al., 2017) on which they were unsuccessful (Sitawarin et al., 2022a), their effectiveness when combined with adversarial training as a defence against query-based black-box attacks has not been systematically evaluated.

2. We demonstrate the efficacy of these transformations across three popular datasets, that is, ImageNet, Tiny ImageNet and CIFAR-10 across a variety of well-known image classification models.

3. We add to the evidence in the literature that randomisation based methods can favourably counter the drop in clean accuracy by disproportionately improving the robust accuracy; in fact, the increase in robust accuracy is much greater than the drop in clean accuracy.

4. We find that randomised transformations are more resilient than RND under EOT-50.

5. As another evidence of the efficacy of the proposed transformations in challenging circumstances we show that the randomised transformation defence effectively counters high-confidence adversarial examples unlike RND, which is vulnerable to them.

3.2 Black-box attacks

In this section, we briefly discuss the weakness of query-based black-box attacks that randomisation based strategies exploit and their adaptive measures.

3.2.1 Weakness of black-box attacks

Query-based black-box attacks are efficient at creating adversarial examples that can sometimes perform as well as white-box attacks (Andriushchenko et al., 2020). Many methods have been proposed such as one based on evolutionary strategies (Ilyas et al., 2018), genetic techniques (Wang et al., 2020) and random search methods (Andriushchenko et al., 2020). Black-box attacks are crucial as they help uncover gradient obfuscation (Athalye et al., 2018), which can give a false sense of adversarial robustness. Croce et al. (2020) have proposed Auto-attack (AA) - an ensemble of attacks that has become almost a standard to reliably test a model's adversarial robustness (Croce et al., 2020). The square attack is a part of AA, and it is widely regarded as one of the strongest query-based black-box attacks, which highlights the significance and esteem that it enjoys in the field (Croce et al., 2020; Qin et al., 2021).

However, recent work has shown that black-box attacks are vulnerable to Random Noise Defence (RND) (Qin et al., 2021). In this method, random noise is added to the images before processing (Qin et al., 2021; Byun et al., 2022). Black-box attacks are vulnerable to

adding random noise to input images at inference time. Combining RND with adversarial training has been shown to improve robustness against black-box attacks relative to models that are not adversarially trained (Qin et al., 2021).

While Random Noise Defence (RND) remains one of the strongest input randomisation strategies, subsequent work has proposed alternative processing-based defences that report improved robustness. For example, Chen et al. (2022) perturb the output logits to counter query-based black-box attacks, while Nguyen et al. (2024) introduce feature perturbations within the network to diminish the attacker's ability to estimate gradients. These methods claim to outperform RND under their respective evaluation settings. Consequently, Dong et al. (2020) and Lucas et al. (2023) highlight the need for more powerful query-based black-box attacks that remain effective in the presence of random perturbations, enabling a more rigorous evaluation of such defences.

3.2.2 Adaptive black-box attacks

Although RND poses a significant challenge to query-based black-box attacks, adaptive attacks have also been developed to circumvent defence mechanisms against black-box attacks. For example, the *OARS* attack (Feng et al., 2023a) has been shown to mitigate stateful defence models (SDMs) (Chen et al., 2020c; Choi et al., 2023; Juuti et al., 2019; Li et al., 2022).

Along a separate line of work, Sitawarin et al. (2022b) find that deterministic preprocessing techniques based on standard data augmentation operations, such as resizing, cropping, and quantisation, can mitigate or significantly weaken the performance of decision-based black-box attacks. However, they also show that these gains disappear once the preprocessing pipeline becomes known to the attacker, as the deterministic transformations can

be reverse-engineered and incorporated into the attack. This limitation arises because the same preprocessing operation is applied to every query, introducing no uncertainty during inference. Furthermore, their evaluation is limited to the ℓ_2 threat model and does not consider the more widely adopted ℓ_∞ constraint, which bounds the maximum perturbation applied to each pixel. Motivated by these limitations, this chapter evaluates randomised transformation-based preprocessing and compares its robustness against the strongest existing Random Noise Defence (RND) under the state-of-the-art *Square Attack*.

Motivated by the need for a more rigorous evaluation of processing-based defences under adaptive threat models, this chapter investigates the robustness of inference-time randomised techniques against a resourceful adversary operating under a high query budget. To lay the foundation for this investigation, the following sections first introduce the defence mechanism of Random Noise Defence (RND), which is widely regarded as one of the strongest defences against query-based black-box attacks, before describing the mechanism of the proposed randomised transformation-based defences.

3.3 Preliminaries

3.3.1 Threat Model

Consider a machine learning classifier f that classifies inputs $\mathbf{x} \in \mathbb{R}^d$ into labels $y_{\text{pred}} \in \mathbb{Z}_k$, where $f : \mathbb{R}^d \to \{1, \ldots, k\}$, d is the dimension of the input, and k is the number of classes. Let $y \in \mathbb{Z}_k$ denote the ground-truth label of a given input. The objective of the classifier is to correctly classify all input instances such that $f(\mathbf{x}) = y$. In contrast, the objective of the adversary is to construct an adversarial example $\mathbf{x}_{\text{adv}} = \mathbf{x} + \delta$ such that $f(\mathbf{x}_{\text{adv}}) \neq y$ (Carlini et al., 2019). The adversarial input $\mathbf{x}_{\text{adv}}$ is perturbed in such a way that it remains close to

the original input $\mathbf{x}$ according to an ℓ_p distance metric,

$$\|\mathbf{x} - \mathbf{x}_{\text{adv}}\|_p \leq \epsilon,$$

where p may correspond to ℓ_2, ℓ_0, or ℓ_∞. The perturbation budget ϵ constrains the maximum allowable change to the input while ensuring that the adversarial example remains sufficiently similar to the original sample yet causes the classifier to misclassify it.

Our focus in this chapter is on the ℓ_∞ distance metric and we assume that adversaries can access the score outputs (logits) of the model as in the conventional case of black-box attacks across this line of work.

3.3.2 Random Noise Defence

In an untargeted black-box attack (where the objective is only to make the model misclassify), the attack can be described as the following optimisation problem:

$$\min_{\mathbf{x}_{adv} \in \mathcal{A}_R(\mathbf{x})} f(\mathbf{x}_{adv}) = \min_{\mathbf{x}_{adv} \in \mathcal{A}_R(\mathbf{x})} (\mathcal{M}_y(\mathbf{x}_{adv}) - \max_{j \neq y} \mathcal{M}_j(\mathbf{x}_{adv})) \tag{3.1}$$

where $\mathcal{M}$ is the attacked model, $\mathcal{M}_j$ denotes the logit w.r.t class j, $\mathcal{A}_R(\mathbf{x}) = \{\mathbf{x}_{adv} |\ \|\mathbf{x} - \mathbf{x}_{adv}\|_p \leq \epsilon\}$ is used to represent an ℓ_p ball around $\mathbf{x}$. Thus, the attacker tries to minimise the difference of logit scores between the right and wrong classification. Note the attacker desires this difference to go below 0; therefore, the more negative the difference the greater is the confidence of the model in misclassifying the example.

To search for $\mathbf{x}_{adv}$ black-box attacks (Andriushchenko et al., 2020), (Ilyas et al., 2018), (Wang et al., 2020) use random perturbations $\boldsymbol{u}$ to alter the clean example $\boldsymbol{x}$. The search is supposed to be moving in the right direction if:

$$h(\mathbf{x}) = f(\mathbf{x} + \mu\mathbf{u}) - f(\mathbf{x}) < 0 \tag{3.2}$$

where μ is the perturbation size.

Random noise defence (RND) (Qin et al., 2021) and a closely related method called small noise defence (SND) (Byun et al., 2022) attempt to misguide a black-box attacker in search of an adversarial example by adding gaussian noise $\boldsymbol{\nu_1}, \boldsymbol{\nu_2} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, based on a factor ν, to attacker's queries to thwart the attacker's estimation of adversarial success as below:

$$g(\mathbf{x}) = f(\mathbf{x} + \mu\mathbf{u} + \nu\boldsymbol{\nu_1}) - f(\mathbf{x} + \nu\boldsymbol{\nu_2}) \tag{3.3}$$

The addition of randomised $\boldsymbol{\nu_1}$ and $\boldsymbol{\nu_2}$ misleads the attacker's estimate of a valuable attack direction. Qin et al. (2021) claim that the adversarial robustness (robustness gains) will be significantly higher when the defender's noise is larger than the attacker's noise (added to the clean example to find an adversarial example). However, the defender cannot arbitrarily increase ν excessively because it can decrease accuracy on the non-adversarial examples (natural accuracy or clean accuracy).

There are a couple of variations of RND based on where the noise is added. Random feature defence (Nguyen et al., 2024) adds noise to the inputs at the feature level to counter black-box attacks while AAA introduces noise on the output logits of the model.

These approaches have set the benchmarks in warding off black-box attacks; however, their performance can still improve. For example, RND achieves a robust accuracy of 14.8% on Tiny ImageNet (Deng et al., 2009) with a PreActResNet18 model, and 53.5% and 18.1% robust accuracy with a ResNet-50 model on CIFAR-10 and ImageNet datasets, respectively. Hence, we have proposed randomised transformations to improve mitigation of black-box

attacks.

3.4 Proposed Methodology

3.4.1 Randomised transformations

Randomised transformations during inference time were proposed by Xie et al. (2017), Guo et al. (2018), Raff et al. (2019) for countering adversarial attacks. Although they were later shown to be ineffective in preventing white-box attacks (Athalye et al., 2018), (Sitawarin et al., 2022a) their effectiveness for countering query-based black-box attacks was not explored. In this work, we propose Random Resize-Cropping (RRC) and Random Rotation (RR) to counter query-based black-box attacks and rigorously test their efficacy on well-established benchmark models and datasets.

Each time an inference is made, a randomised transformation randomly shifts the orientation or locations of the pixel values within the image, due to which the adversarial gradients are not estimated correctly. This is because the slight randomised variations in pixel locations lead to conflicting loss directions for each query, hence successfully misguiding the attacker. Moreover, we employ the proposed transformations with the adversarially trained models; this is based on the results of Qin et al. (2021), which showed that employing random transformation or random noise injection during inference time is most effective when combined with an adversarially robust model i.e. a model trained using adversarial retraining (Madry et al., 2018). Therefore, in this work, we only consider models that are trained adversarially and aim to improve their adversarial robustness to state-of-the-art black-box attack, that is, square attack. Therefore, randomisation is used to complement adversarially trained models.

Since randomised transformations cause significant changes to the input, we assess their performance by aggregating their predictions over multiple forward passes. The predictions from the multiple forward passes are combined using voting to form a final prediction. The label with the most votes becomes the classified label. We find that this leads to better performance when using randomised transformations than simply aggregating the logits over multiple predictions or using a single forward pass to form a prediction. Although the query time of the model is slightly increased, it comes with the benefit of better defence performance as will be shown in later sections.

As part of data augmentation, input data transformations that randomly change the input images are commonly used to augment the training dataset with additional data to improve model generalisation and invariance to visual artefacts, which are neutral to human observers but can mislead a neural network. However, here we propose that these transformations when used during inference time can also mislead the attacker trying to find adversarial directions. Under these random transformations, the equation that the attacker uses to find adversarial directions is given by:

$$m(\mathbf{x}) = f(t_{r_1}(\mathbf{x} + \mu\mathbf{u})) - f(t_{r_2}(\mathbf{x})) \tag{3.4}$$

where t_r is a random transformation, which could include Random Resize-Cropping (RRC), Random Rotation and Random Cropping-Padding. The transformation is applied during the inference stage. Random transformations mislead the attacker from finding useful adversarial directions. Therefore, one event of a transformation such as t_{r_1} might misguide the attacker by giving a higher loss value but another event of the same transformation may produce a completely incongruent outcome. The resulting confusion can effectively reduce the attack performance.

Next, we detail the two proposed transformations.

Random Resize-Cropping This transformation works by cropping a random size of an image, and resizing it to the original shape accepted by the model. We set the cropping limit to 80% of the image after which the image is resized to $32 \times 32 \times 3$ for the CIFAR-10 dataset, $64 \times 64 \times 3$ for Tiny ImageNet dataset and $224 \times 224 \times 3$ for the ImageNet dataset. We use bilinear interpolation for resizing. Based on preliminary experiments, we use the 80% limit because it better preserves clean accuracy while successfully mitigating black-box attacks. For visual demonstration, the original image of a Goldfish is shown in Figure 3.1 while its transformations are given in Figure 3.2.

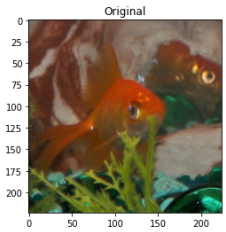


Figure 3.1: Original image of a Goldfish from the ImageNet dataset.

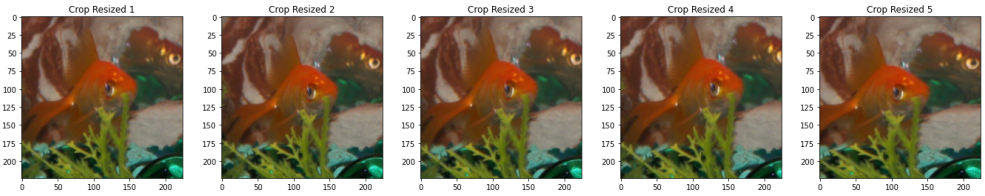


Figure 3.2: Random Resize-Cropping (RRC) variants of an image of a Goldfish from the ImageNet dataset. Our preliminary experiments revealed superior performance of 80% cropping value over smaller cropping values.

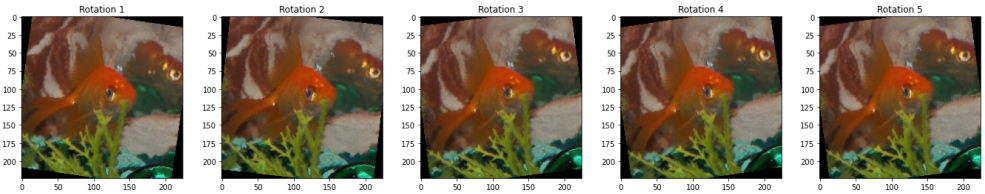


Figure 3.3: Random rotation variants of an image of a Goldfish from the ImageNet dataset.

Random Rotation. Random Rotation rotates the image using a random angle. We set the range of the random angle as $[-10^{\circ}, 10^{\circ}]$ and find it effective for the experiments conducted here. The parameter choice largely depends on the model's invariance to these transformations. Subsequent studies could aim to develop strategies for identifying the most resilient parameters, taking into account the characteristics of different models and datasets.

Note, in this transformation the original size of the image is maintained. The image is only rotated using the random angle during the inference stage and the randomised rotation misguides the attacker to incorrect directions instead of the adversarial directions, which compromises attack performance. An example of this transformation is given in Figure 3.3.

3.4.2 Evaluation

Attacker Capabilities. To ensure extraordinary numbers to stress-test the defences (Qin et al., 2021; Nguyen et al., 2024), we set the attacker's budget to 100,000 queries for EOT-10 attack and 500,000 queries for EOT-50 attack. These settings are deliberately intense so that the proposed methods are stress tested under strong attack conditions. All of our evaluations are made under the strong Expectation Over Transformation (EOT) (Athalye et al., 2018) attack strategy that averages the predictions before calculating attack directions. We set EOT to 10 or 50 to evaluate the defence performance under the square attack (Andriushchenko et al., 2020). Since Square Attack is widely credited as one of the best known query-based black-box attacks and is reported by Qin et al. (2021) as the best-performing attack against RND, it serves as a strong baseline adversary in our evaluation. The square attack is run for 10,000 iterations with default settings.

As discussed in Chapter 2, most of the output processing defences are vulnerable to hard-label attacks. Therefore, we also compare the performance of the randomised defences

under EOT-10 of a hard-label (decision-based) attack called the RayS attack Chen et al., 2020b. A hard-label attack is an attack that only uses the decision labels of the model to launch an attack without relying on the score probabilities.

For each EOT step, the prediction is received over a single random transform variant. However, by using 10 or 50 steps of EOT, the predictions are averaged out over 10 or 50 different random transformation variants, to intensify the attack and making it easier for the attacker. Therefore, for evaluation purposes, the adversary uses EOT to generate adversarial examples, and at each iteration of the attack, we test if the adversarial example generated at each iteration is successful at fooling the model over a model that uses voting over 10 different predictions. All the randomised defences are evaluated in the same manner to ensure reliable evaluation.

Algorithm 1 Confident Square Attack (C-SQA)

Require: f, $\mathbf{x}$, ϵ, N, M (EOT iterations), label $y_{true} \rightarrow \{1, ..., k\}$, image size w, number of colour channels c

$\mathbf{x}^{adv} \leftarrow init(\mathbf{x})$

$l^* \leftarrow L(f(\mathbf{x}), y)$

for all iterations in N **do**

  $h^{(i)} \leftarrow$ side length of the square to modify

  $\boldsymbol{\delta} \leftarrow \mathrm{P}(\epsilon, h^{(i)}, w, c, \mathbf{x}^{adv}, \mathrm{x})$

  $\mathbf{x}^{adv}_{new} \leftarrow$ Project $\mathbf{x}^{adv} + \boldsymbol{\delta}$ onto $\{\mathbf{z} \in \mathbb{R}^d : \quad \|\mathbf{z} - \mathbf{x}\|_p \leq \epsilon\} \cap [0, 1]^d$

  $l_{new} \leftarrow \frac{1}{M} \sum_{j=0}^{M} L(f(\mathbf{x}^{adv}_{new}), y)$

  if $l_{new} < l^*$ **then**

    $\mathbf{x}^{adv} \leftarrow \mathbf{x}^{adv}_{new}$

    $l^* \leftarrow l_{new}$

  end if

end for

return $\mathbf{x}^{adv}$

Confident Attack. Typically, black-box attacks stop once an adversarial example is found (Wu et al., 2021) to save query budget; however, under randomisation based testing this can produce examples that were simply *lucky* to be adversarial and fail to remain so when tested again. Therefore, to evaluate the defences, we also use confident black-box attacks as an additional measure during our experiments; these attacks keep optimising the adversarial loss until a specified number of queries has elapsed, even after a misclassification has occurred. Once the query budget is exhausted, the samples are passed through inference stage again to assess whether they remain misclassified.

Confident attacks aim to generate adversarial examples that continue to remain adversarial under random effects. In other words, the adversarial examples are more reliable. However, if a defence can stop the attacker from creating confident adversarial examples then it shows that the defence can truly mitigate attacks. The algorithm for the confident square attack is given in Algorithm 1; it is identical to the original square attack except for our additions on line 3 and line 7, which include running the algorithm until the complete number of iterations have exhausted and averaging of predictions for calculating the new loss value. Further details about the square attack algorithm are provided in the original work by Andriushchenko et al. (2020).

3.5 Experimental Settings

In this section, we introduce our choices for datasets, models, metrics and defences evaluated.

3.5.1 Datasets and Models

We evaluate the randomised defences on three datasets: CIFAR-10, Tiny ImageNet and ImageNet. We use the models and settings similar to the one used by Qin et al. (2021) that introduce RND and remains the reference due to its state-of-the-art performance. Therefore, we randomly sample 1,000 images from the CIFAR-10 test set and 1,000 images from the validation set provided by Huang et al. (2019) that were also used by Qin et al. (2021). To further improve the rigour of experiments, unlike Qin et al. (2021) we also evaluate our proposals on the Tiny ImageNet: we use 500 images randomly sampled from the test set. The models used for CIFAR-10 include a pre-trained ResNet-50 trained with ℓ_∞ distance of 8/255 by Robustness Library (Engstrom et al., 2019), a pre-trained Wide-ResNet-28-10 by Gowal et al. (2020), and a ResNet-18 adversarially trained by us; using an ℓ_∞ distance of 8/255 and TRADES (Zhang et al., 2019) as the loss function. We have at least matched the contemporary expectations (EOT is at least 10) and where computational allowances permit we have gone beyond it to EOT-50.

The model used for ImageNet is ResNet-50 trained with ℓ_∞ distance of 4/255 by (Engstrom et al., 2019). We train a PreActResNet18 model on Tiny ImageNet using adversarial training with TRADES (Zhang et al., 2019) as a loss function, and set the ℓ_∞ distance to 8/255.

The ℓ_∞ distance of the attack is set to 12.75/255 for all models and defences used in this work. As stated in section 3.4.2, square attack is the state-of-the-art black-box attack and is also part of the auto-attack ensemble (Croce et al., 2020). Therefore, we use the square attack with default settings, which gives the best attack performance (Qin et al., 2021).

Furthermore, the experiments are conducted on two machines, one with Intel Core i9

CPU, Nvidia GeForce RTX 3080 Ti GPU and 32GB of RAM, and the other with Intel core i7 CPU, Nvidia Quadro T1000 GPU and 16GB of RAM.

3.5.2 Data Augmentations

Typically, to improve model invariance and generalisation, data augmentations based on various transformations such as those proposed in this chapter are used during the training stage. We follow this good practice and use such well-trained models for our experiments.

This augmentation also makes sense for the defences we propose because a model untrained over these augmentations can simply misclassify even a clean example when it goes through the proposed transformations. Note, that this training is not a limitation of our method because, as discussed above, such invariance is desirable for a good model in any case. Moreover, previous work by Rebuffi et al. (2021) has already shown that data augmentation approaches can improve adversarial robustness.

The Robustness library employs different data augmentation strategies depending on the dataset. For the ImageNet model, it applies random resize-cropping, random horizontal flipping, and colour jitter, whereas the CIFAR-10 model is trained using random cropping, random horizontal flipping, colour jitter, and random rotation. In our experiments, we use the library's pretrained ResNet-50 models for both CIFAR-10 and ImageNet. In addition, we train two models locally: a PreActResNet-18 on TinyImageNet and a ResNet-18 on CIFAR-10. Both locally trained models are trained using random resize-cropping and random rotation as their data augmentation strategies, these are the same transformations we propose for defence.

3.5.3 Defences Evaluated

We evaluate our proposed randomised transformation based defences against the RND (Qin et al., 2021). Although RND, being so successful, is the main benchmark, we also report indicative results on its variation, random feature defence (RFD) (Nguyen et al., 2024) as discussed in section 3.3.2 and random cropping-padding (Xie et al., 2017), as discussed in section 3.4.1. AAA (Chen et al., 2022) was not evaluated as it is vulnerable to decision-based attacks as shown by Nguyen et al. (2024).

Therefore, we compare Random Resize-Cropping (RRC), and Random Rotation with RND (Qin et al., 2021) - the principal benchmark on all datasets for a comprehensive comparison. However, for comparisons with random cropping-padding (Xie et al., 2017) and randomised feature defence (Nguyen et al., 2024) we use the first 200 samples of the 1,000 samples from the ImageNet validation set.

Our experiments reveal that square attack gives the best attack performance, better than decision-based attacks, on defences that use input-level randomisation. This is supported by the results of (Qin et al., 2021).

3.5.4 Model Prediction

We modify the prediction of the models used for random transformation slightly. The final model prediction is taken after aggregation over 10 or 50 different variants of a random transformation for non-confidence increasing attacks. Under EOT-10 attack we take aggregation over 10 transformations whereas for EOT-50 attack we take aggregation over 50 transformations. We utilise this technique for all models used in this work that use randomisation during the inference stage (except for RND). This ensures that clean accuracy is not significantly affected. For random transformations and random feature defence (Nguyen et al.,

2024), we use voting to get the final prediction, whereas for RND we sum the logits over different predictions to select the predicted label since it is closer to the original work. To be complete, we tried RND with voting and it didn't improve the results and therefore we used the settings consistent with the original RND proposition. The illustration of model prediction is given in Figure 3.4.

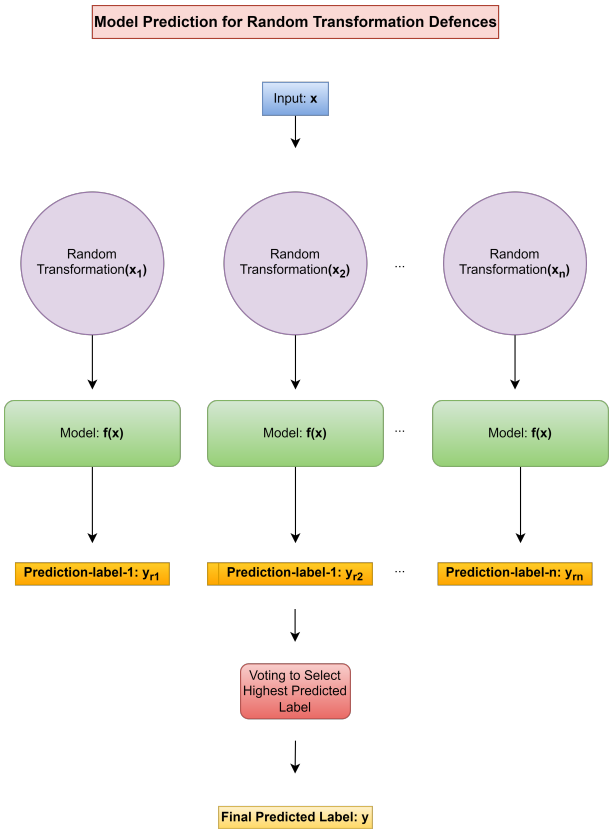


Figure 3.4: Illustration of how the model prediction is used for random transformation defences in this work.

Furthermore, the model prediction for confidence-increasing attacks uses aggregation over 100 transformations once the adversarial examples are generated after utilising the attack budget. During the adversarial example generation (attack), the attacker uses 10 EOT steps to average out the randomness.

3.5.5 Metrics

We report the results by using both robust accuracy and attack failures. To generate the results without confidence-based attacks, we used the code provided by (Qin et al., 2022). On the other hand, to generate the results with attacks that are modified to increase the confidence of the misclassification (confident-attack), we use the adversarial robustness toolbox (ART) (Nicolae et al., 2018). We also report the clean accuracy of the models, which is the prediction test accuracy on data that is not perturbed.

3.6 Experimental Results

In this section, we provide the experimental results for the randomised processing defences here.

3.6.1 CIFAR-10

We report the results on CIFAR-10 with EOT 10 and EOT-50 for both the regular and confidence increasing attacks. The EOT-10 and EOT-50 results are reported in Table 3.1 while the results under confidence increasing attack are reported in Table 3.2.

The results under EOT-10 show that the Random Rotation defence is the best-performing defence with a robust accuracy of 54.4%. Our proposed transformations give a competitive performance that improves over regular adversarially trained (AT) model with no defence by more than 21.0%. Additionally, under EOT-50 attacks the margin of improvement increases significantly. With Wide-ResNet-28-10, Random Rotation improves over RND by 4.6%, and Random Resize-Cropping (RRC) beats RND by 3.5% whereas

with ResNet-18 architecture, Random Rotation beats RND by 3.7% and Random Resize-Cropping (RRC) beats RND by 2.3%. The attack success rate graph for the defences are given in Figure 3.5.

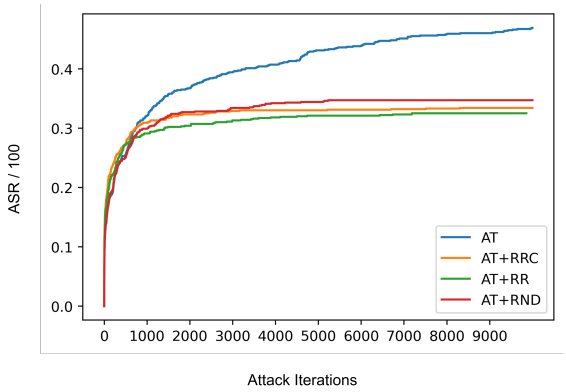


Figure 3.5: The attack success rate graph of the randomised defences on square attack with EOT-50 on ResNet-18 model architecture and CIFAR-10 dataset. The attack success rate is measured across the iterations of the attack.

Moreover, Table 3.2 shows the results under a confidence-increasing attack, where the full attack budget is utilised and confident adversarial examples are created. Our proposed random transformations outperform Random Noise Defence by more than 8.0%. This suggests that random transformations are better than Random Noise Defence at misguiding the attacker from finding adversarial regions to come up with confident adversarial examples. This demonstrates that random transformations are a more effective defence than noise-based defences.

3.6.2 ImageNet

The results on the ImageNet dataset are also reported on both the regular square attack and the confident square attack. We compare RND with our proposed transformations over 1,000 samples. Our proposed randomised cropping-resizing improves significantly over RND

Table 3.1: Performance of randomised defences under Square Attack (Andriushchenko et al., 2020) across CIFAR-10, ImageNet and Tiny ImageNet datasets. The best-performing defence is highlighted in bold. The second-best-performing defence is highlighted in bold italics.

Dataset	Test Samples	Model Architecture	Attack EOT Steps	Defence Technique	Clean Accuracy (%)	Attack Failures	Robust Accuracy (%)
CIFAR-10	1000	ResNet-50	10	AT	**88.7**	322	32.2
				AT+RND (Qin et al., 2021)	87.8	***535***	***53.5***
				AT+Resize-Cropping (Ours)	88.1	***535***	***53.5***
				AT+Rotation (Ours)	***88.7***	**544**	**54.4**
	1000	ResNet-18	50	AT	**82.7**	361	36.1
				AT+RND (Qin et al., 2021)	80.9	461	46.1
				AT+Resize-Cropping (Ours)	81.6	***484***	***48.4***
				AT+Rotation (Ours)	**82.2**	**498**	**49.8**
	1000	Wide-ResNet-28-10 (Zagoruyko et al., 2016)	50	AT	**89.9**	483	48.3
				AT+RND (Qin et al., 2021)	88.4	586	58.6
				AT+Resize-Cropping (Ours)	88.5	***621***	***62.1***
				AT+Rotation (Ours)	***88.8***	**632**	**63.2**
ImageNet	1000	ResNet-50	10	AT	**61.6**	98	9.8
				AT+RND (Qin et al., 2021)	56.7	181	18.1
				AT+Resize-Cropping (Ours)	***59.9***	**275**	**27.5**
				AT+Rotation (Ours)	59.1	***256***	***25.6***
Tiny ImageNet	500	PreActResNet-18	10	AT	***39.2***	43	8.6
				AT+RND (Qin et al., 2021)	36.0	74	14.8
				AT+Resize-Cropping (Ours)	**40.0**	**82**	**16.4**
				AT+Rotation (Ours)	37.4	***78***	***15.6***

by giving an improvement of 9.4%. Randomised rotation also improves over RND by 7.5%. The results are reported in Table 3.1.

Additionally, we compare the performance of RND and our proposed randomised transformations over the confident square attack in Table 3.2. The results show that the robust accuracy of all randomised defences is improved. The improvement is explained by the fact that finding confident adversarial examples is slightly more challenging than finding an input that causes misclassification by chance over a huge budget when the model is being tested under a stochastic technique. Our proposed randomised transformations are more effective at misguiding the attack from finding confident adversarial regions on both CIFAR-10 and ImageNet datasets. Randomised cropping-resizing gives a robust accuracy of 36.5% while RND gives a robust accuracy of only 19.5%.

Table 3.2: Performance of randomised defences on CIFAR-10 and ImageNet datasets under confident square attack (C-SQA). The best-performing defence is highlighted in bold. The second best-performing defence is highlighted in bold italics.

Dataset	**Test Samples**	**Model Architecture**	**Attack EOT Steps**	**Defence Technique**	**Attack Failures**	**Robust Accuracy** (%)
CIFAR-10	1000	ResNet-50	10	AT+RND (Qin et al., 2021)	543	54.3
				AT+Resize-Cropping (Ours)	**628**	**62.8**
				AT+Rotation (Ours)	***627***	***62.7***
ImageNet	200	ResNet-50	10	AT+RND (Qin et al., 2021)	39	19.5
				AT+Resize-Cropping (Ours)	**73**	**36.5**
				AT+Rotation (Ours)	***71***	***35.5***

3.6.3 Tiny ImageNet

We perform experiments on Tiny ImageNet to further establish the improvement of our proposed randomised transformations. The results on the Tiny ImageNet using a PreActResNet18 are given in Table 3.1. According to the results, RND gives a robust accuracy of 14.8% while Random Resize-Cropping (RRC) improves on it and takes it to 16.4%. On the other hand, Random Rotation also improves the robust accuracy by 0.8%. The results confirm the improvement of randomised transformations over RND.

3.6.4 Indicative Comparison with RFD and Random Padding-Cropping

Encouraged by previous results, we also present a preliminary comparison with RFD and randomised cropping-padding as mentioned in section 3.5.3. We use 200 samples for comparison as we use an expensive budget of 10,000 iterations combined with 10 EOT steps.

The different randomisation techniques to counter black-box attack are compared in Table 3.3. We use the RND (Qin et al., 2021) method with defender noise set to 0.05. For the RFD defence, a noise of 0.03 is introduced after each layer block of ResNet-50. The results show that random transformation defences outperform noise-defences such as RND

Table 3.3: Performance of additional randomised defences on first 200 samples from the 1000 random samples of the ImageNet validation set under square attack (Andriushchenko et al., 2020). The best-performing model is highlighted in bold. The second best-performing defence is highlighted in bold italics.

Dataset	**Test Samples**	**Model Architecture**	**Attack EOT Steps**	**Defence Technique**	**Clean Accuracy** (%)	**Attack Failures**	**Robust Accuracy** (%)
ImageNet	200	ResNet-50	10	AT+RND (Qin et al., 2021)	69.0	37	18.5
				AT+Crop-Pad (Xie et al., 2017)	69.0	56	28.0
				AT+RFD (Nguyen et al., 2024)	**75.0**	49	24.5
				AT+Resize-Cropping (Ours)	***72.5***	**63**	**31.5**
				AT+Rotation (Ours)	72.0	***58***	***29.0***

and RFD. Moreover, random resize-cropping outperforms all other defences by giving the highest robust accuracy of 31.5%.

3.6.5 Evaluation on hard-label attack

So far we have shown superior performance of RT over RND under a score-based attack (square attack). In this section, we show that a powerful decision-based attack (RayS) can also be vulnerable to input-randomisation defences. The results show that RRC is the best-performing defence on RayS, followed by RR. The results are shown in Table 3.4.

Table 3.4: Performance of randomised defences on CIFAR-10 under the hard-label RayS attack (Chen et al., 2020b).

Dataset	**Test Samples**	**Model Architecture**	**Attack EOT Steps**	**Defence Technique**	**Attack Failures**	**Robust Accuracy** (%)
CIFAR-10	1000	ResNet-50	10	AT	373	37.3
				AT+RND (Qin et al., 2021)	555	55.50
				AT+Resize-Cropping (Ours)	568	56.80
				AT+Rotation (Ours)	561	56.10

3.6.6 Ablation Study

We perform two ablation studies, the first on CIFAR-10 and ImageNet to show how accumulating the predictions over multiple randomised forward passes helps to increase the robustness of the model to query-based black-box attacks. We also conduct another ablation study to investigate the underlying reason behind the success of the proposed random transformations. We find that for these transformations to be successful randomisation plays a critical role as motivated in Section 3.4.1.

Table 3.5: Ablation comparing multipass and single-pass inference of randomised defences under Square Attack (Andriushchenko et al., 2020). ImageNet results use the first 200 of the 1000 random validation samples. The best-performing setting within each dataset is highlighted in bold, the second-best in bold italics.

Dataset	Test Samples	Model Architecture	Attack EOT Steps	Defence Technique	Inference	Clean Accuracy (%)	Attack Failures	Robust Accuracy (%)
CIFAR-10	1000	ResNet-18	50	AT	Deterministic	**82.7**	361	36.1
				AT+RND (Qin et al., 2021)	Multipass	80.9	461	46.1
					Single pass	81.5	374	37.4
				AT+Resize-Cropping (Ours)	Multipass	81.6	***484***	***48.4***
					Single pass	80.2	264	26.4
				AT+Rotation (Ours)	Multipass	***82.2***	**498**	**49.8**
					Single pass	79.4	390	39.0
ImageNet	200	ResNet-50	10	AT	Deterministic	**75.0**	21	10.5
				AT+RND (Qin et al., 2021)	Multipass	69.0	37	18.5
					Single pass	69.0	34	17.0
				AT+Resize-Cropping (Ours)	Multipass	***72.5***	**63**	**31.5**
					Single pass	71.5	48	24.0
				AT+Rotation (Ours)	Multipass	72.0	***58***	***29.0***
					Single pass	69.5	52	26.0

The results for the first ablation study are given in Table 3.5. We see that using a single forward pass significantly hinders the defence performance. On CIFAR-10, the robust accuracy of RND decreases from 46.1% to 37.4%, while the robust accuracy of Random Rotation decreases from 49.8% to 39.0%. The robust accuracy of Random Resize-Cropping (RRC) also decreases. The regular AT model is unaffected as no randomisation component affects its performance. On ImageNet, the robust accuracy of each defence also decreases on a single forward pass.

To explain what makes the randomised transformation work, we add a second ablation study to measure robust accuracy for randomised rotation defence by fixing the rotation angle to a specific value, instead of selecting from multiple random angles. Similarly, we measure the robust accuracy of Random Resize-Cropping (RRC) by fixing the cropping size to a specific value. The reason behind this ablation study is that one hypothesis for its success could be that as a result of these randomised transformations, parts of the images are cropped out; we want to understand whether it is the cropping alone that helps robustness by simply taking out the parts that the attacker might have perturbed or is it the mechanism of randomisation that plays a critical role.

The experiments are conducted using a ResNet-18 model on the CIFAR-10 dataset. The robust accuracies for rotation angles 1,5 and 10 are 36.1%, 31.7%, and 26.6%, respectively. In comparison, the robust accuracy of the AT model is 36.1%, which is significantly higher than the accuracies at fixed angles of 5 and 10 degrees. Similarly, the robust accuracies of resize-cropping when we fix the cropping scale to 80% and 90% are 32.3% and 31.6%, respectively. This shows that an attacker can circumvent cropped-out parts of images to find adversarial noise within other parts of the image. Consequently, it is the randomisation procedure (randomly choosing from multiple angles or cropping sizes) that results in higher robustness of the proposed pre-processing defences.

3.6.7 Increasing Attacker Noise

To evaluate the impact of increasing attacker noise (ϵ) on robustness, we perform additional experiments on CIFAR-10 with ResNet-18 model under EOT-50. (Qin et al., 2021) suggest that for RND to be effective, the defender noise must be increased as the attacker noise increases. However, this is a challenge as the attacker noise is unknown in all practical settings. In addition, adding higher defender noise can degrade image quality.

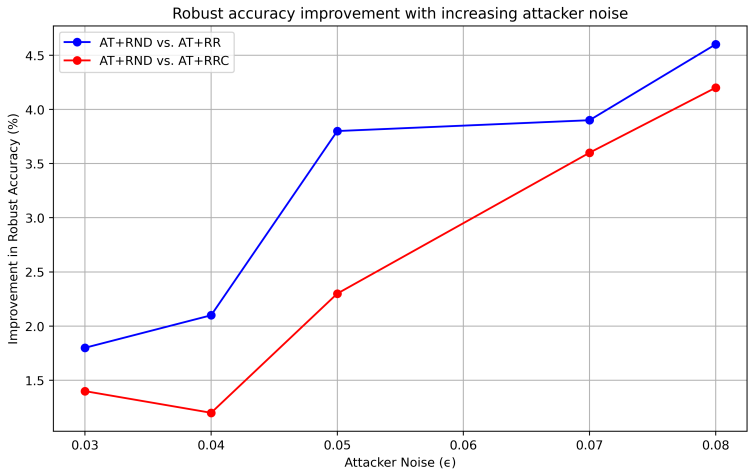


Figure 3.6: The improvement in robust accuracy of transformation based defences over RND as attacker noise is increased.

Therefore, we increase defender noise and keep it the same as the attacker noise in line with the experimental settings in the original work by Qin et al. (2021). On the other hand, our proposed transformations use the same parameter settings, which gives them the edge over RND. The results are reported in Figure 3.6. We observe a clear increase of robust accuracy improvement of transformation based defences over RND as attacker noise is increased. One possible explanation for the poor performance of RND is the higher degradation caused by pixel level noise added by RND.

To add further evidence to the image degradation hypothesis, we report the Structural Similarity Index Measure (SSIM) of the images (Wang et al., 2004) under various noise levels on randomisation techniques. We measure the SSIM between the image at ℓ_∞ distance of 0.05 and images at higher attacker and defender noise levels for both techniques.

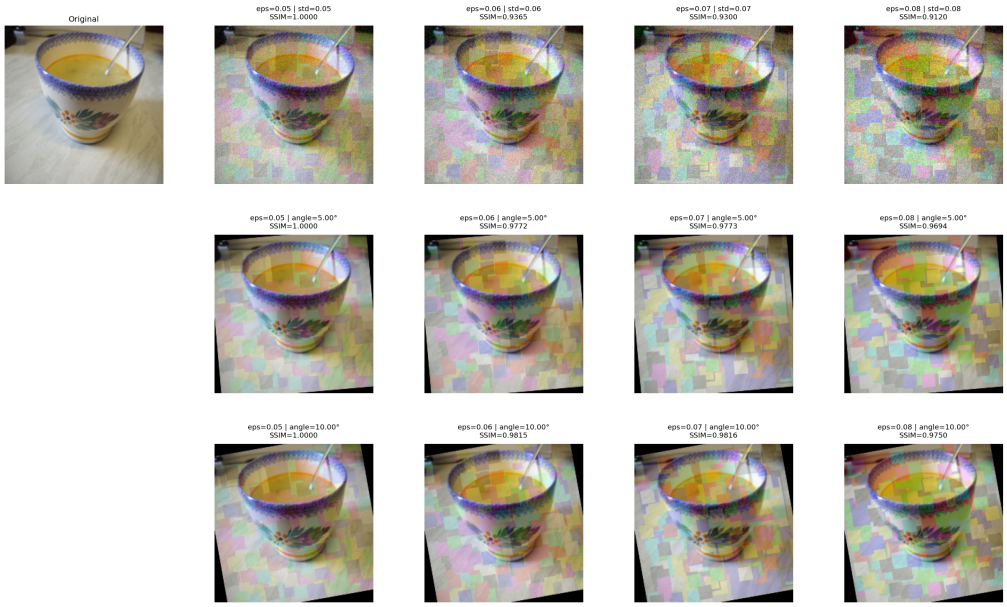


Figure 3.7: Representative SSIM example 1 under increasing attacker noise.



Figure 3.8: Representative SSIM example 2 under increasing attacker noise.



Figure 3.9: Representative SSIM example 3 under increasing attacker noise.

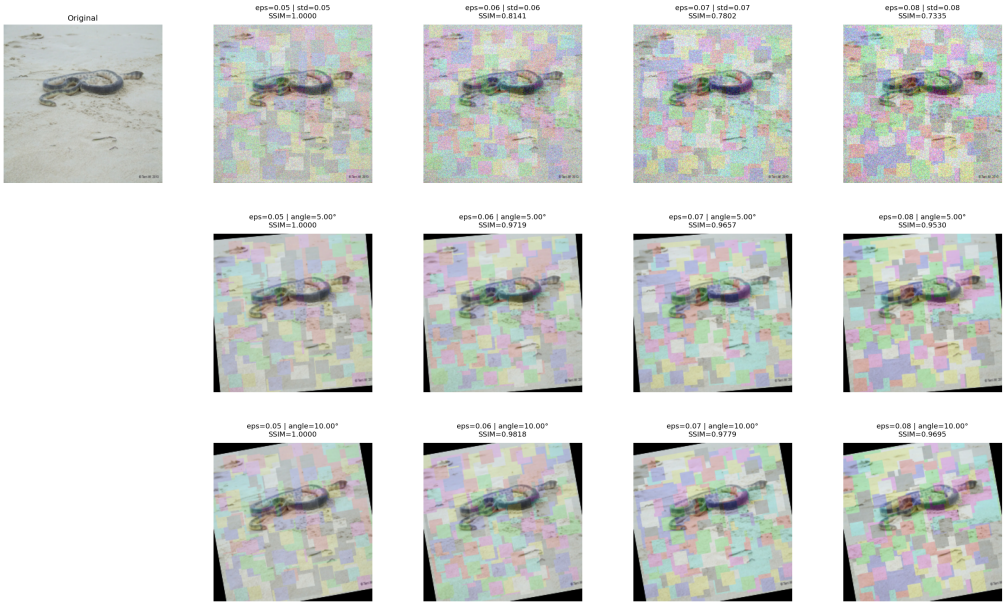


Figure 3.10: Representative SSIM example 4 under increasing attacker noise.



Figure 3.11: Representative SSIM example 5 under increasing attacker noise.

Table 3.6: SSIM comparison at the highest attacker-noise setting across five representative examples. Higher SSIM means less degradation.

Example	RND	RR (5°)	RR (10°)
Figure 3.7	0.9120	0.9694	**0.9750**
Figure 3.8	0.9209	0.9666	**0.9695**
Figure 3.9	0.9484	0.9824	**0.9855**
Figure 3.10	0.7335	0.9530	**0.9695**
Figure 3.11	0.9416	0.9771	**0.9795**

As seen qualitatively in Figures 3.7 to 3.11 and quantitatively in Table 3.6, RR preserves substantially higher structural similarity than RND because RND depends on increasing defender noise when attacker increases their noise.

3.6.8 Statistical Significance Test

To test the improvements for statistical significance, we calculated p-values on CIFAR-10 and ImageNet. For CIFAR-10 we used ResNet-18 model under EOT-50 on 1000 samples, whereas for ImageNet we used ResNet-50 under EOT-10 on the first 200 samples. To calculate p-values we have used McNemar test (McNemar, 1947), which can be used to measure the statistical significance between the accuracy calculations (Atik et al., 2021).

On CIFAR-10, the p-value for improvement of Random Rotation over RND is 1.03×10^{-7} while for Random Resize-Cropping (RRC) over RND it is 0.031. On ImageNet, the p-value for improvement of Random Rotation over RND is 2.10×10^{-7} while for Random Resize-Cropping (RRC) over RND it is 8.04×10^{-7}. In most of the cases the p-values are extremely low (of the order 10^{-7}); in all of the cases they are less than 0.05. Hence the significance of difference is clearly established.

3.6.9 Conclusion

This chapter has shown that randomised transformations are a strong alternative to noise-based randomisation for mitigating query-based black-box attacks. Across CIFAR-10, Tiny ImageNet, and ImageNet, the proposed Random Resize-Cropping (RRC) and Random Rotation defences consistently match or outperform RND, which remains the benchmark to improve upon in this line of work. On ImageNet (Russakovsky et al., 2015), under the EOT-10 Square Attack, inference-time Random Resize-Cropping (RRC) and Random Rotation

improve the robust accuracy of a ResNet-50 model (He et al., 2016) over RND (Qin et al., 2021) by 9.4% and 7.5%, respectively. On Tiny ImageNet (Deng et al., 2009), using a Pre-ActResNet18 model, Random Resize-Cropping (RRC) improves robust accuracy over RND by 1.6%, while Random Rotation improves it by 0.8%. On CIFAR-10 (Krizhevsky et al., 2009), using a ResNet-50 model, Random Rotation improves robust accuracy over RND by 0.9%, while Random Resize-Cropping (RRC) matches the performance of RND. These findings show that inference-time transformations provide a practically effective defence against query-based black-box attacks.

Additionally, to stress test the proposed methods under intense attack settings, previous studies (Qin et al., 2021), (Nguyen et al., 2024) test defences under an EOT-10 adversary, whereas we take the evaluation further by rigorously testing them against a strong EOT-50 adversary. The original RND work (Qin et al., 2021) suggests that using more than 10 EOT steps has a limited effect on the defence. However, our results suggest that RND is significantly vulnerable to larger EOT steps while random transformations are not. Under an EOT-50 adversary, the improvement is significantly higher as Random Rotation beats RND by 4.6% and Random Resize-Cropping (RRC) beats RND by 3.5% on ResNet-18 architecture.

Encouraged by these results, we also stress test the proposed transformations in more strenuous circumstances. We test the defences against high-confidence adversarial examples that are generated by decreasing the confidence on the correct label for the entire set of attack iterations unlike stopping the attack at the misclassification iteration. In a scenario where high-confidence adversarial examples determine attack success rate, randomised transformations outperform RND on CIFAR-10 by more than 8.0% and on ImageNet by around 16.0%.

The strength of this conclusion lies not only in the absolute gains, but also in how

the chapter structures the comparison and the evaluation. In particular:

- the proposed transformation-based defences are compared directly against RND, the established benchmark in query-based black-box robustness;
- both the benchmark defence and the proposed methods are evaluated against Square Attack (Andriushchenko et al., 2020), a widely recognised strong query-based black-box attack and the best-performing attack against RND reported by Qin et al. (2021);
- the defences are further stress tested under Expectation over Transformation (EOT) (Athalye et al., 2018), where averaging predictions over multiple forward passes provides the adversary with richer information than a single prediction and therefore creates a more demanding evaluation.

Furthermore, we have also compared our proposed transformations against the recently proposed variation of RND, Randomised feature defence (RFD) which adds gaussian noise in between the layers of the model (Nguyen et al., 2024). Our results show an improvement of randomised transformations over RFD.

Overall, the chapter adds to the evidence that randomisation can favourably trade a small drop in clean accuracy for a much larger gain in robust accuracy when the form of randomisation is chosen carefully and paired with suitably trained models. The results suggest that transformation-based randomness interferes with the attacker's search more effectively than additive noise and remains more resistant under adaptive and stress-tested conditions. At the same time, these findings do not imply that randomisation alone fully resolves the challenge of adaptive attackers. Rather, they motivate the next step of investigating whether transformation-based defences can be combined with stronger mechanisms, such as stateful defences, to provide more durable protection against query-based black-box attacks.

Chapter 4

Re-evaluating the robustness of Stateful Defence Models

In the previous chapters, we have shown that randomised transformations are more effective than random noise-based strategy at mitigating black-box attacks. However, the so-called stateful defence models that aim to detect query-based attacks also effectively mitigate black-box attacks. In fact, recent SDMs have claimed 100% detection rate. The claims of full robustness warrant a close and careful investigation as adaptive attackers can create techniques to bypass the defences such as the OARS attack (Feng et al., 2023b) discussed in Chapter 2.

This chapter addresses whether the claims of full robustness are valid and answers **RQ3**, "Can detection-based defences against query-based black-box attacks provide complete robustness, as suggested by prior work?"

4.1 Introduction

Although previously shown vulnerable to the so-called adaptive attacks (Feng et al., 2023b), recently the SDMs have presented some remarkable results in defence to an advanced adaptive attack called Oracle-Guided Adaptive Rejection Sampling (OARS) attack (Feng et al., 2023b). OARS is a state-of-the-art black-box attack that exposed the vulnerabilities of pre-

viously proposed SDMs considered to be secure (Chen et al., 2020c; Li et al., 2022; Choi et al., 2023; Esmaeili et al., 2022). However, recently proposed SDMs by Wang et al. (2024), Park et al. (2025), and Li et al. (2025) have claimed 0% attack success rate (ASR) against OARS, which is an extraordinary claim that merits further investigation.

Therefore, in this chapter, we investigate the recently proposed SDMs that claim 0% attack success rate; these are AdvQDet (Wang et al., 2024), GWAD (Park et al., 2025), and QPA (Li et al., 2025). The results presented later in this chapter show that GWAD (Park et al., 2025) and QPA (Li et al., 2025) are vulnerable to attacks under the ℓ_∞ distance metric. AdvQDet, although resistant to ℓ_∞ based attacks, can still be defeated by a new attack presented in this chapter termed **DazzlePatch attack**, which achieves an ASR of 75.50% with ℓ_∞ perturbation budget of 0.09 under 22 random restarts (attack seed initialisations). 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 (Wang et al., 2024) and QPA (Li et al., 2025).

- We show that incorporating input randomisation techniques (Qin et al., 2021; Ali et al., 2025) 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.

4.2 Related Work

Commonly effective defences against query-based attacks are input/output randomisation (Qin et al., 2021; Ali et al., 2025; Aithal et al., 2022b) or stateful detection (Chen et al., 2020c; Wang et al., 2024; Park et al., 2025; Li et al., 2025; Li et al., 2022). Randomisation based techniques perturb the inputs or outputs of the models to thwart the attack. On the other hand, Stateful Defence Models (SDMs), first proposed by Chen et al. (2020c), aim to detect adversarial examples during query-based black-box attacks by tracking the history of past queries.

The investigation by Chen et al. (2020c) 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 ℓ_∞ distance metric, a widely adopted and more stringent perturbation constraint than ℓ_2 under which achieving robustness is generally more challenging (Croce et al., 2021). They further assume that robustness under ℓ_∞ will naturally extend to other ℓ_p norms (Chen et al., 2020c).

Following their work, Blacklight (Li et al., 2022), Perceptual Image Hashing (PIHA) (Choi et al., 2023) and IIOT-SDA (Esmaeili et al., 2022) were 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 (Chen et al., 2020c). OARS changes the proposal distribution of the attack. For example, in the case of Square Attack (Andriushchenko et al., 2020) instead of sampling a single square, it samples multiple squares that lead to the least detection. OARS (Feng et al., 2023b) exposed multiple defences and showed that the stateful defences were "not yet secure".

OARS thus intensified the arms race and attracted subsequent works (Wang et al., 2024; Li et al., 2025; Park et al., 2025) on developing stateful detectors that are claimed to be invulnerable to the OARS attack, claiming a 0% attack success rate. This is an extraordinary claim that warrants a careful investigation. Although we have already introduced them in Chapter 2, here we briefly explain the working mechanics of OARS-resistant SDMs and discuss their limitations.

4.2.0.1 Gradient's Watch Adversarial Detection (GWAD)

GWAD (Park et al., 2025) detects attacks by training a classifier on the *Histogram of delta similarities (HoDS)*. The classifier detects attack by classifying a histogram of cosine similar-

ity, which is generated based on the perturbation differences (δ) between consecutive queries, capturing the signature of adversarial behaviour. Essentially, it analyses the query update patterns, that is HoDS, rather than the direct inputs themselves.

It maintains a bank of queries submitted and uses it to create HoDS, each attack has its own unique characteristic 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.

Furthermore, the authors show that GWAD is vulnerable to benign-query injection attacks, in which an adversary deliberately submits benign queries to displace adversarial queries from the query banks, thereby reducing the likelihood of detection. To mitigate this weakness, they propose a screening mechanism to propose an enhanced version called GWAD+, which identifies and filters benign queries before they are stored in the query banks.

Limitations:

- The defence has not been evaluated under the ℓ_∞ threat model. Moreover, we need to re-validate the claims under the ℓ_2 threat model, which we find the defence vulnerable under.
- It takes a certain number of queries to build HoDS and the mechanism to load HoDS is not addressed in Park et al. (2025). 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 only shown to generalise over different datasets and not on attacks that are unseen or have unique HoDS.
- The defence has not been evaluated under random restarts of the attack.

4.2.0.2 Query Provenance Analysis (QPA)

Another recent SDM that is resistant to OARS uses provenance analysis models relationships among system entities to detect and investigate intrusions (Han et al., 2020; King et al., 2003; Li et al., 2023). QPA (Li et al., 2025) 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 extracts a feature representation using a pretrained feature extractor and subsequently generates a perceptual hash using PIHA (Choi et al., 2023).

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 - graphs which may include nodes (image hashes) that are potentially a part of query-based attack. 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 ℓ_2 versions of the query-based black-box attacks. It has not been evaluated under the ℓ_∞ versions of most attacks except for Natural Evolutionary Strategies (NES) attack (Ilyas et al., 2018).
- The performance of the defence relies on the number of graphs stored in the cache. Detection efficiency can vary depending upon the number of graphs. However, authors find 20 graphs with 15 nodes each to be the best-performing.

- The authors assert that the query provenance graph cannot grow infinitely. Therefore, they reset the query provenance graph every 24 hours. The authors argue that in the event of resetting the database (bank), an adversary on average will require 333 days to launch an attack because average number of queries required for the attacks to succeed is 6,721. This depends on the defence's time-to-detect (TTD). TTD is the number of queries required to get the first detection. Therefore, 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.

4.2.0.3 AdvQDet

Unlike the previous SDMs resistant to OARS, AdvQDet boasts the ability of adversarial training. AdvQDet learns a robust similarity representation by fine-tuning the CLIP image encoder (Radford et al., 2021) using Adversarial Contrastive Prompt Tuning (ACPT). The fine-tuning serves as adversarial training and its goal is to encourage the encoder to produce similar embeddings for adversarial inputs.

To detect attacks, AdvQDet maintains a historical embedding bank containing the embeddings of previously submitted queries together with their predicted labels. Each incoming query is encoded using the fine-tuned CLIP encoder and compared against the embedding bank using cosine similarity. If the similarity exceeds a predefined threshold, the query is identified as part of an ongoing adversarial attack. Rather than forwarding the query to the target classifier, AdvQDet returns the cached prediction associated with the most similar historical embedding, thereby disrupting the attack optimisation process.

The authors report a Time-to-Detect (TTD) of three queries against query-based black-box attacks while demonstrating strong zero-shot generalisation across multiple datasets

and attack algorithms despite training only on ImageNet (Deng et al., 2009; Russakovsky et al., 2015). Furthermore, they show that the memory requirements remain practical, estimating that supporting one million users with a history of 100 queries each requires approximately 95 GB of storage using float16 precision. The framework is evaluated under the ℓ_∞ threat model, which is generally expected to transfer more effectively to other threat models than detectors trained under alternative perturbation constraints (Chen et al., 2020c).

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.

4.3 Evaluating Stateful Defence Models

4.3.1 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 ℓ_2, ℓ_0 or ℓ_∞. Only the *terminal* example is constrained under the ℓ_∞; how the attacker searches for x_{adv} (the intermediate queries) is unrestricted. This is the standard adversarial

robustness contract under ℓ_p (Carlini et al., 2019). We assume that the adversary can access the model's score outputs (logits) to accomplish this goal, while also considering the more restrictive hard-label-only setting, where only the model's predicted class labels are available to the adversary.

4.3.2 Evaluation Protocol

As discussed, although some SDMs demonstrate robustness against the advanced OARS attack, they continue to exhibit defence-specific weaknesses. Building on these observations, we propose a comprehensive evaluation protocol designed to systematically assess these weaknesses and provide a more rigorous basis for validating the robustness claims of stateful defence models.

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 ℓ_∞ threat model. Robustness under the ℓ_∞ threat model is generally harder to achieve than under ℓ_2 (Croce et al., 2021). By contrast, AdvQDet was evaluated under an ℓ_∞ 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 (Douceur, 2002) 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 ℓ_∞ threat model to ensure conclusions generalise beyond ℓ_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.

4.3.3 DazzlePatch Attack

The evaluation framework presented in the previous section provides a rigorous basis for assessing robustness. However, the effectiveness of such an evaluation ultimately depends on the strength of the attacks employed. As robust SDMs, such as AdvQDet, have become increasingly resistant to existing adaptive attacks, including OARS, there is a need for more effective query-based black-box attacks capable of challenging these defences. Motivated by this limitation, we propose DazzlePatch, a novel attack specifically designed to provide a stronger evaluation of SDM robustness.

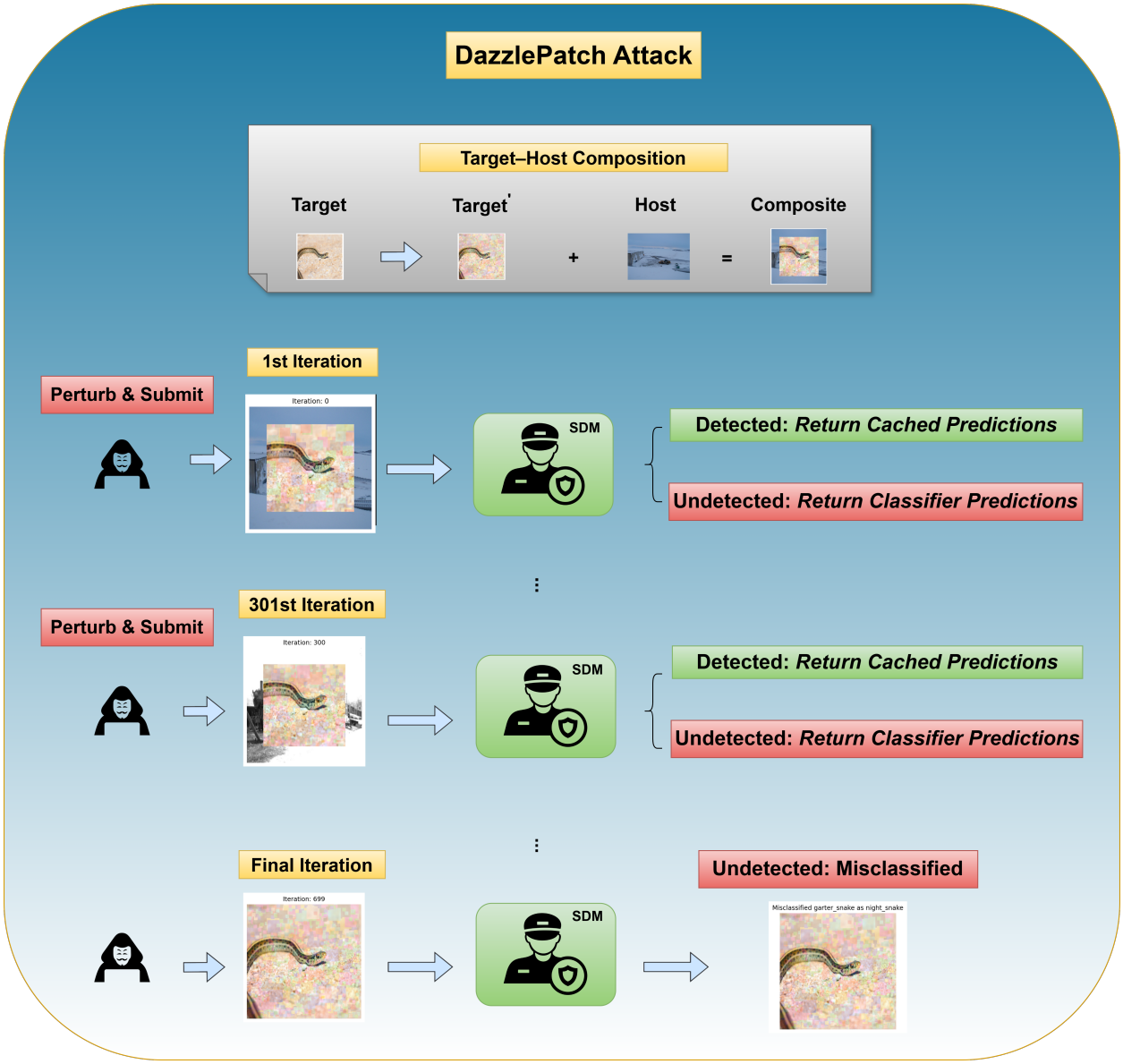


Figure 4.1: Proposed DazzlePatch replaces the border of a target image to avoid detection from SDMs while allowing Square Attack to make adversarial changes. In the final iteration, the target′ image is fully restored with its original borders to satisfy ℓ_p constraint.

Attack overview: Our method follows the spirit of query blinding (Chen et al., 2020c): 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

Algorithm 2 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}

a *uniquely sourced host* image, while the *host borders* (pixels outside the core) come from the host. To 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 Figure 4.1 and the pseudocode is given in Algorithm 2.

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 ℓ_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, after the first misclassification on the hosted composite, we continue with a *confidence-increasing* phase that explicitly maximises the incorrect-class confidence. This creates a robustness margin so that removing the host borders (de-hosting) minimises the chances of reverting the prediction into the correct class. Unlike in Chapter 3, where increasing confidence helped the defence, here it helps the attacker because the attacker seeks to ensure the image maintains its misclassification ability once original borders are restored.

Target Patch Size and Position:

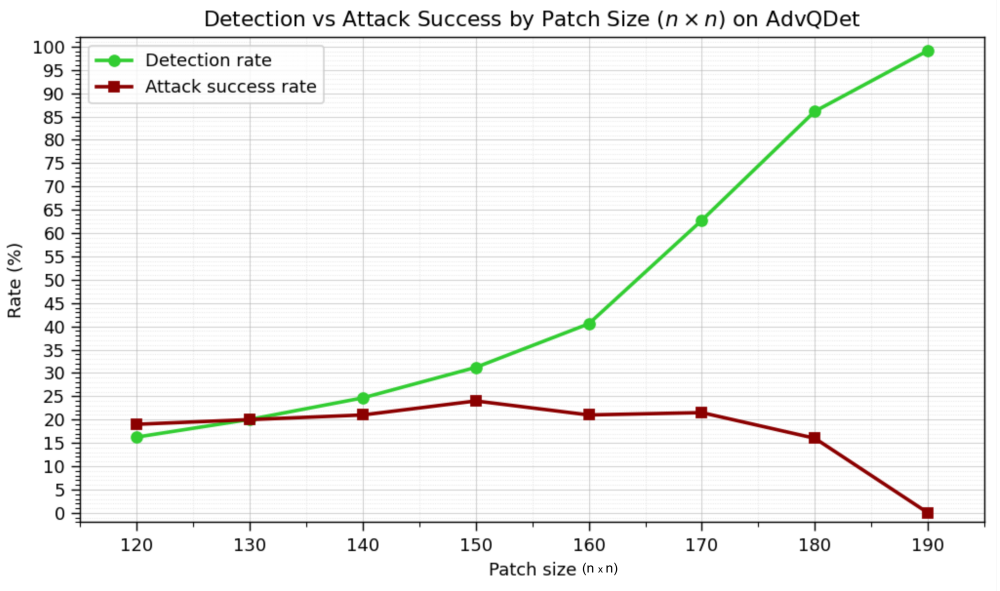


Figure 4.2: The relationship between detection rate and attack success rate on each patch size. Each patch size is evaluated on a single run of DazzlePatch Attack on AdvQDet.

The patch size of the target image plays a role in determining both the detection rate

and attack success rate as shown in Figure 4.2. 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 160×160 and 170×170 for the target core, out of a total image size of 224×224 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 centre and just change the borders around it. However, an alternative strategy can be to randomise the patch position. Previous theory (Qin et al., 2021) suggests that may instead weaken the attack.

4.3.4 Experimental Settings

Here, we discuss the experimental setup used to evaluate the SDMs. We replicate each defence in its original setting to ensure fidelity. The datasets and models reflect the selection in the original works. The summary of experimental settings is given in Table 4.1.

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: We replicate each method in its original setting to ensure

Table 4.1: Experimental settings used to evaluate the three stateful defence models. Each defence is evaluated under its original experimental setting to ensure fidelity.

Setting	**GWAD**	**QPA**	**AdvQDet**
Dataset	CIFAR-10	ImageNet	ImageNet
Evaluation Samples	1,000 test images	200 validation images	200 validation images
Backbone Classifier	ResNet-18	ResNet-152	ResNet-50
Square Strategy	Single-square	Multi-square	Multi-square
Query Budget	8,000	700	700
Perturbation Metric	ℓ_∞	ℓ_∞	ℓ_∞
Perturbation Budget(s)	0.01, 0.05	0.05, 0.07, 0.09	0.05, 0.07, 0.09

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 et al. (2019), 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. Moreover, we also use German Traffic Sign Recognition Benchmark (GTSRB) (Stallkamp et al., 2012), Chinese Traffic Sign dataset (Yemelyanov, 2020) and skin cancer dataset (Tschandl et al., 2018) for the false positive analysis of AdvQDet because these datasets contain similar looking images across different classes.

Additionally, the classifiers used in our work are based on ResNet architecture (He et al., 2016) as in the original works by Park et al. (2025), Li et al. (2025), and Wang et al.

(2024): ResNet-152 for QPA, ResNet-18 for GWAD, and ResNet-50 for AdvQDet.

All the experiments are performed on a single machine with Intel Core i9 CPU, Nvidia GeForce RTX 3080 Ti GPU and 32GB of RAM.

Attack configuration: Black-box attacks are usually run with a budget of 10,000 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 significantly smaller budget.

Furthermore, DazzlePatch attack perturbs the target image using a black-box optimisation procedure to induce misclassification under an ℓ_p metric. We find that the success of DazzlePatch-style attacks depends strongly on the optimisation recipe: for the Square Attack (Andriushchenko et al., 2020), inserting multiple squares per iteration substantially strengthens the attack against QPA and AdvQDet.

Additionally, 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. Our results indicate GWAD is already vulnerable to the original Square Attack under the ℓ_∞ 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 ℓ_2 based attacks and a single ℓ_∞ attack and restricts them under a

normalised budget of 0.05. Similarly, AdvQDet evaluates their defence on an ℓ_∞ budget of 0.05.

In this work, we focus on the ℓ_∞ metric. Particularly for the DazzlePatch attack, we use ℓ_∞ 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 (Li et al., 2022; Ding et al., 2018; Addepalli et al., 2022; De et al., 2025; Maung et al., 2020) have evaluated classifiers under higher budgets, even on low-resolution datasets such as CIFAR-10 (Krizhevsky et al., 2009). We show the impact of visual distortion when increasing the perturbation in Figure 4.3. The objects in the images are easily recognisable.

Moreover, since GWAD is notably susceptible to the standard Square Attack under the ℓ_∞ 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 ℓ_∞ 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 (Krasin et al., 2017; Google Research, 2021). 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 demonstrate that even this conservative attack setting remains effective against the evaluated defences. Moreover, our later analysis reveals that the presence of semantically meaningful border content, rather than random noise, is the key factor enabling the attack to evade the detector.

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%.

4.4 Experimental Results

We report the empirical results for each defence evaluated in this section. Before introducing the experimental results, we explain why the adaptive attacks are needed

The failure of OARS attack: AdvQDet, QPA, and GWAD all claim robustness against OARS (Feng et al., 2023b), 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 99.71% 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 attack fails on all three evaluated SDMs (Wang et al., 2024; Li et al., 2025; Park et al., 2025). Therefore, to bypass these defences, we modify the square attack to attack QPA and GWAD, whereas the adversarially trained AdvQDet is bypassed by using our novel attack.

4.4.1 GWAD

GWAD is evaluated on CIFAR-10 as the publicly available detection model by the authors was trained on CIFAR-10.

Table 4.2: 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 ℓ_∞	63.80	99.80
ASR under $\ell_2^{normalised} = 0.1$	63.70	14.90
Avg. hit rate	8.7	0.7

We use Square Attack under the more general ℓ_∞ ($\epsilon = 0.01, 0.05$) (Andriushchenko et al., 2020) 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 Table 4.2, with its HoDS loaded with initial queries. The optimisation of the attack starts only after 256 iterations (once the bank has loaded). At an ℓ_∞ budget of 0.05, the ASR reaches 99.80% with low detectability, while at 0.01 it drops to 63.80%. Notably, 63.70% of the 1000 samples also satisfy the original constraint by Park et al. (2025), producing visually imperceptible changes.

4.4.2 Query Provenance Analysis

QPA is evaluated on ImageNet as the publicly available detection model by the authors was trained on ImageNet.

Square attack: QPA (Li et al., 2025) does not assume a specific ℓ_p norm in the

Table 4.3: Robustness of QPA under adaptive attacks on 200 samples from ImageNet dataset. DazzlePatch achieves competitive ASR.

			DazzlePatch Attack (Ours)		**Square Attack**	
Restart	ϵ	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

threat model; however, we find that changing the norm makes substantial difference to the attack success rate. With ℓ_2, Square Attack has only 1.0% ASR against QPA with $\epsilon = 0.5$ and 1.0; however, with ℓ_∞ 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 Table 4.3. 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.

4.4.3 AdvQDet

AdvQDet is evaluated on ImageNet because the publicly available detection model provided by the authors was trained on this dataset. Furthermore, AdvQDet claims zero-shot generalisation, whereby a detector trained solely on ImageNet is expected to generalise to unseen datasets without additional training. Consequently, evaluating its false positive rate across multiple datasets is essential to verify the validity of this claim.

4.4.3.1 False Positive Rate

Table 4.4: 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

GTSRB: 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) (Stallkamp et al., 2012) it has a high false positive rate on speed limits as shown in Table 4.4. 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.

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 (Wang et al., 2024) exhibits a notably high false positive rate, as shown in Table 4.5. However, this behaviour is not consistent across all labels.

Table 4.5: 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

Chinese Traffic Signs: We also test the false positive rate of AdvQDet on Chinese Traffic Sign dataset (Yemelyanov, 2020). The false positive rate is high on most labels. The results are given in Table 4.6.

Skin Cancer: Similarly, on the skin cancer dataset (Tschandl et al., 2018), we find that for Basal Cell Carcinoma (BCC) label, the FP rate is 97.47% with non-cancerous labels. Skin cancer 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

Table 4.6: 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

attacks. Nevertheless, retraining the prompt token on these datasets may mitigate the FP rate.

4.4.3.2 Attack Results

Square attack: We apply the standard Square Attack to test the base performance of AdvQDet, and find that it 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 the first query, 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 (score-based): AdvQDet's minimal TTD (3) necessitates a novel bypass; we therefore propose DazzlePatch, which mounts attacks using Expectation Over Transformation (EOT) (Athalye et al., 2018) with three transformations. The loss is accumulated over every 3 queries for EOT, while remaining within the overall query budget

Table 4.7: Robustness of AdvQDet under adaptive attacks on 200 samples from the ImageNet dataset. DazzlePatch significantly outperforms Square Attack.

Attack			**Square**		**DazzlePatch (Ours)**					
Attack Type			**Score-based**		**Score-based**		**Hard-label-based**		**Score-based**	
Defence			**AdvQDet**		**AdvQDet**		**AdvQDet**		**AdvQDet + RRC**	
Restart	ϵ	No. of Squares	ASR (%)	Avg. Hit Rate (%)	ASR (%)	Avg. Hit Rate (%)	ASR (%)	Avg. Hit Rate (%)	ASR (%)	Avg. Hit Rate (%)
0	0.05	170	14.00	99.71	21.00	40.55	17.50	38.60	11.00	55.90
4	0.05	170	23.00	99.71	34.50	42.68	26.00	39.77	20.00	58.33
9	0.05	170	24.50	99.71	39.00	54.67	32.50	50.34	21.50	69.00
15	0.07	170	31.00	99.71	54.00	53.79	50.00	50.34	35.00	68.03
21	0.09	170	49.00	99.71	75.50	52.98	72.00	50.34	49.00	67.98

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 ℓ_∞ 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 significantly raises 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 Table 4.7.

DazzlePatch attack (hard-label): We also provide the results under the hard-label version of the square attack in Table 4.7. We design the hard-label version of DazzlePatch with square attack which starts with an epsilon of 0.2 on the central patch and gets decreased by 0.01 after every iteration until we reach the minimum epsilon. It works by maximising the number of votes for the incorrect label over 10 EOTs. The attack retains the sample that has most misclassifications over 10 iterations.

4.4.3.3 Combining Randomisation and AdvQDet

As proposed in Chapter 3, we use a random transformation-based defence against query-based black-box attacks called Random Resize-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 21 restarts is 49.0%; improving over vanilla AdvQDet by 26.50%.

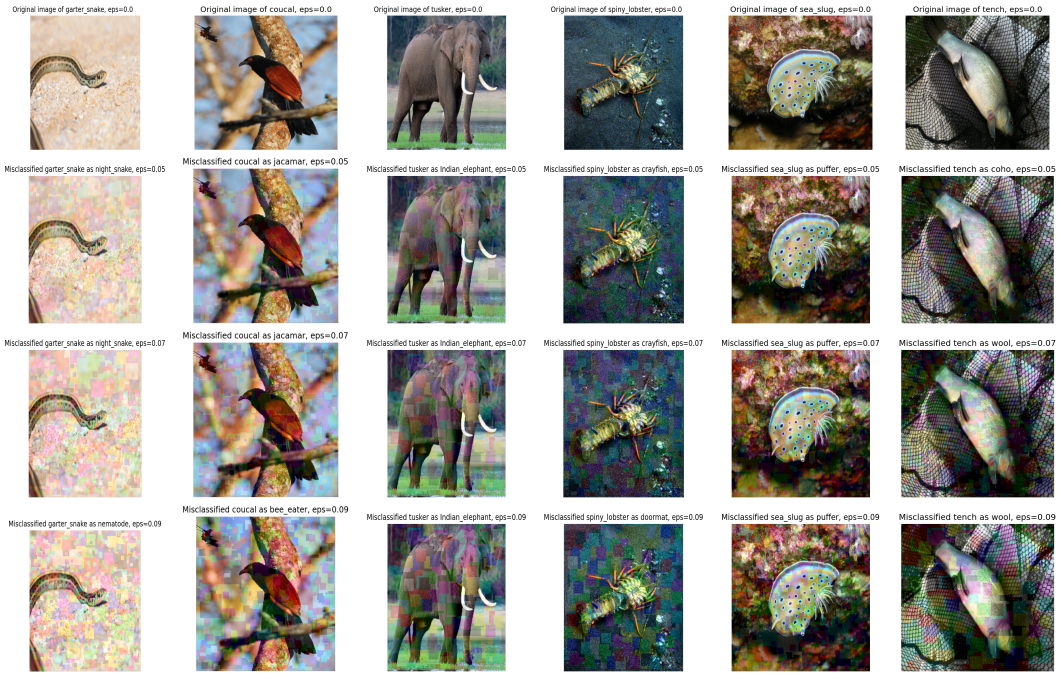


Figure 4.3: Visualisation of images distorted at various levels of perturbations using Square Attack with multiple squares.

4.4.4 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 ℓ_∞ 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 contrast, 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%. The crucial finding is that as long as host images are high-resolution and contain semantically meaningful information, the attack success will increase. If host images are low-resolution (CIFAR-10) or contain random patterns then it will make the composite vulnerable to detection.

4.5 Conclusion

This chapter aims to investigate the following research question **RQ3**, "Can detection-based defences against query-based black-box attacks provide complete robustness, as suggested

by prior work?" To answer this question, we evaluate the OARS-resistant SDMs.

According to our evaluation, QPA and GWAD can be compromised with only the Square Attack under ℓ_∞ 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. DazzlePatch 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. We believe that this work will motivate the development of attacks that produce valid adversarial examples with imperceptible distortions.

Moreover, we also show that integrating input randomisation techniques with SDMs can enhance the robustness. Although randomisation can partially mitigate the attacker, it is vulnerable to a resourceful attacker (high EOTs). Therefore, we aim to test the limitations of these defences under a resourceful attacker in the next chapter.

Chapter 5

Evaluating defences against black-box attacks under a resourceful attacker

In this chapter, we explore the limitations of two key defensive techniques used to mitigate query-based black-box attacks, namely, randomisation and stateful defence models under a resourceful attack.

This chapter addresses the following research question **RQ4**, "How do the defences against black-box attacks scale down as the attacker becomes significantly more resourceful?"

5.1 Introduction

We have so far learnt that randomised transformations enhance mitigation of query-based attacks and that detection of query-based attacks using stateful defences is not a foolproof method. Both defence methods are partially successful in mitigating query-based attacks. Moreover, combining the two defensive strategies yields superior defensive abilities to thwart attackers.

In this chapter, we seek to stress-test these defences under a resourceful attack. We first explore the limits of Random Noise Defence (RND) and Randomised Transformation defence (RT), particularly Random Rotation (RR) under a resourceful attack that can aver-

age output predictions using EOT-2000. Thereafter, we explore the impact of a resourceful attack on a defence that combines two black-box defences, namely, SDM and input randomisation, which has achieved state-of-the-art robust accuracy.

It is important to note that evaluating the limitations of input randomisation under such severe attack circumstances is computationally feasible on relatively small-scale datasets such as CIFAR-10. However, such computational resources may be readily available to well-resourced, government-level adversaries. In contrast, conducting the same experiments on ImageNet incurs substantially higher computational costs due to its larger image resolution and increased complexity. Consequently, CIFAR-10 is used to investigate the limitations of input randomisation with a 20 million query budget. Nevertheless, experiments involving the combination of SDM and randomisation are performed on ImageNet to demonstrate applicability in a large-scale setting, albeit under significantly reduced query budget of 5000 queries to maintain computational feasibility.

5.2 Input Randomisation under Resourceful Attacker

Input randomisation can mitigate query-based black-box attacks as seen in Chapter 3. However, experiments based on expectation-over-transformation (EOT) revealed that averaging over multiple outputs can reduce the effect of randomisation in mitigating query-based black-box attacks; we had used EOT-50 in Chapter 3. A question that naturally follows is whether higher EOTs can further reduce the impact of randomisation as a defence.

Therefore, in this section, we perform experiments by assuming a resourceful attacker with a high query-budget. We push the limits of randomisation by using an EOT of 2000. We perform these experiments on CIFAR-10 dataset for reasons outlined previously.

5.2.1 Experimental settings

We test Random Rotation (RR) and Random Noise Defence (RND) as part of the input randomisation defences under a resourceful attacker with the ability to launch EOT-2000 attack. The square attack is utilised for this purpose and the allowed iterations are 10,000. On each iteration, the attacker is further allowed 2000 queries to average the outputs to make EOT-2000 feasible.

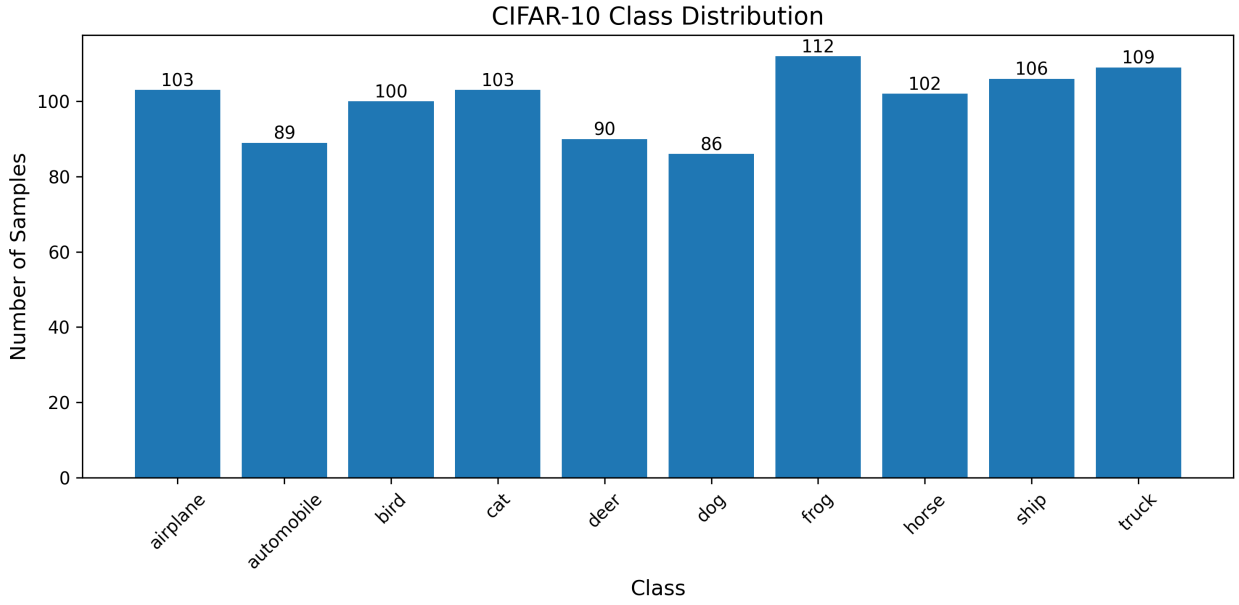


Figure 5.1: Per-class distribution of the randomly sampled 1000 images from the CIFAR-10 dataset.

We randomly sample 1,000 images from the CIFAR-10 dataset (the same samples as in Chapter 3). The class distribution is shown in Figure 5.1. The samples are almost evenly distributed except for deer, dog and automobile classes, which have samples below 100. However, in later sections, we will see that even though the class data is relatively evenly sampled, the robustness gains from input randomisation vary substantially across classes. Per-class analysis is therefore necessary to characterise this uneven improvement.

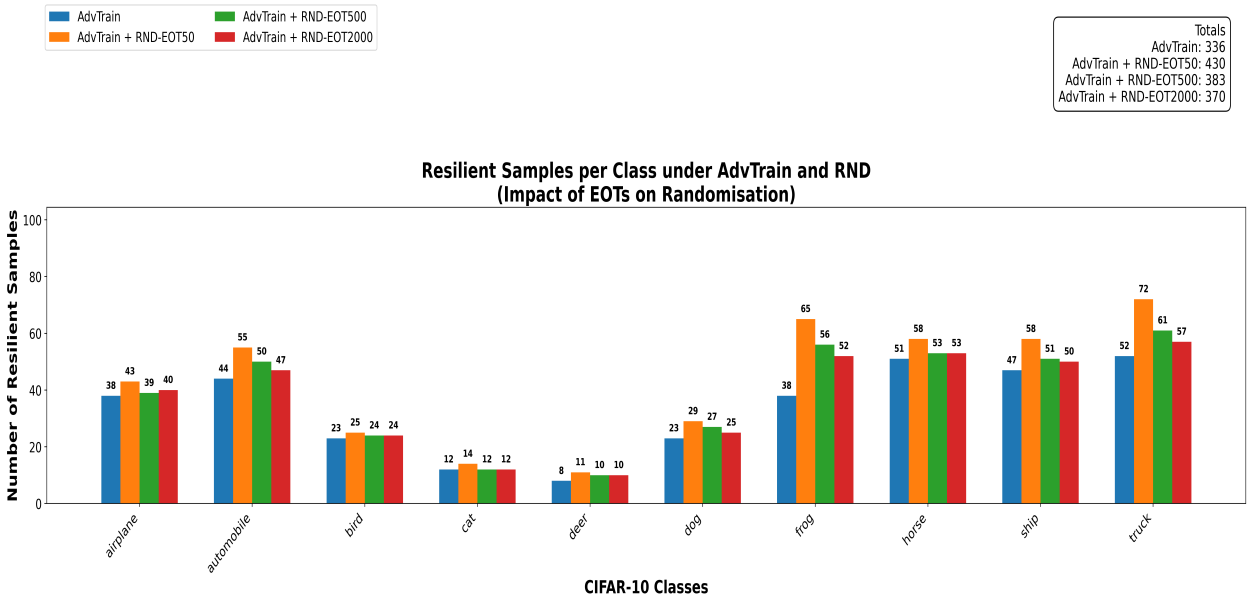
All the experiments are performed on a single machine with Intel Core i9 CPU, Nvidia GeForce RTX 3080 Ti GPU and 32GB of RAM. We perform the experiments on the adversarially trained ResNet-18 model, and use a perturbation budget of $\ell_\infty = 0.05$ with logit loss for evaluation.

5.2.2 Random Noise Defence under high EOTs

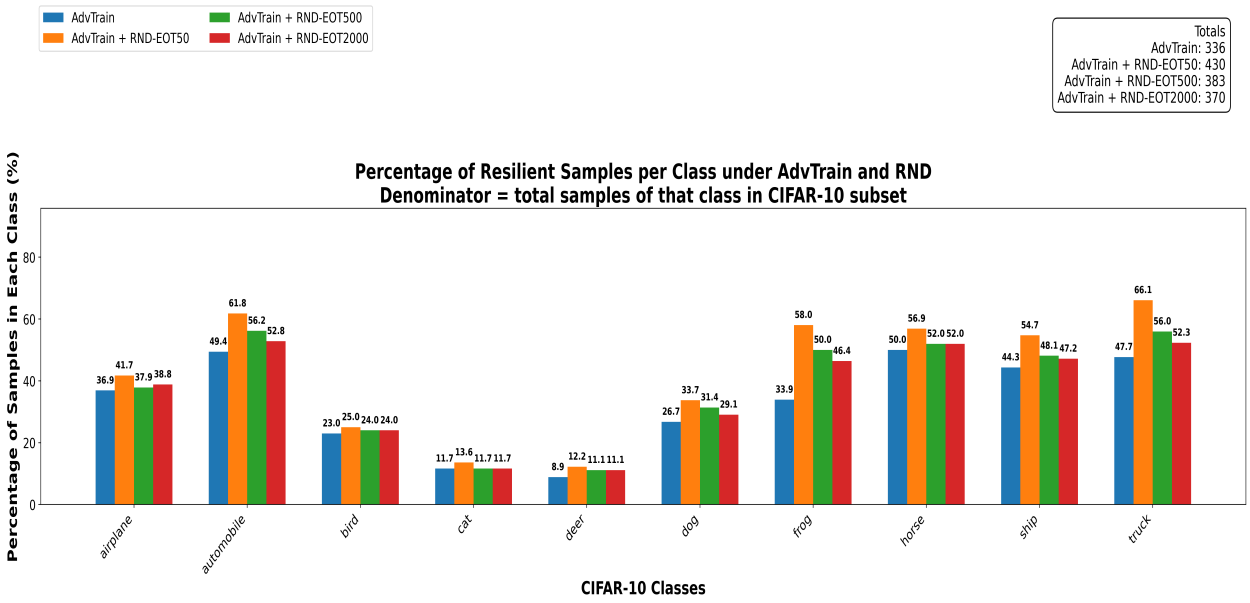
Random Noise Defence provides partial mitigation against query-based attacks. However, our results show that a sufficiently resourceful attacker can substantially diminish the effectiveness of this defence. As shown in Figure 5.2, RND+AT is only 3.4% better than AT alone under EOT-2000, whereas normally it was 9.4% better. This shows that high EOT is particularly detrimental to defence performances. Randomisation as a defence aims to make image samples maintain correct classification, which means the samples become resilient under the defence.

The results reveal that randomisation can substantially increase the number of resilient samples beyond those provided by adversarial training alone, but this improvement is highly dependent on both the attack strength and the class label. Under adversarial training alone, 336 samples remain resilient. Introducing RND with EOT-50 increases this number to 430 samples, representing the strongest overall performance. However, as the attacker increases the number of EOT samples from 50 to 500 and 2000, the number of resilient samples decreases to 383 and 370 respectively. This indicates that a stronger adaptive attacker is able to partially overcome the benefits of randomisation, although the defence still provides improvements over adversarial training alone.

A particularly notable finding is the strong class-dependent nature of resilience. Vehicle classes such as truck, automobile, and ship, together with the frog class, consistently



(a) Number of samples



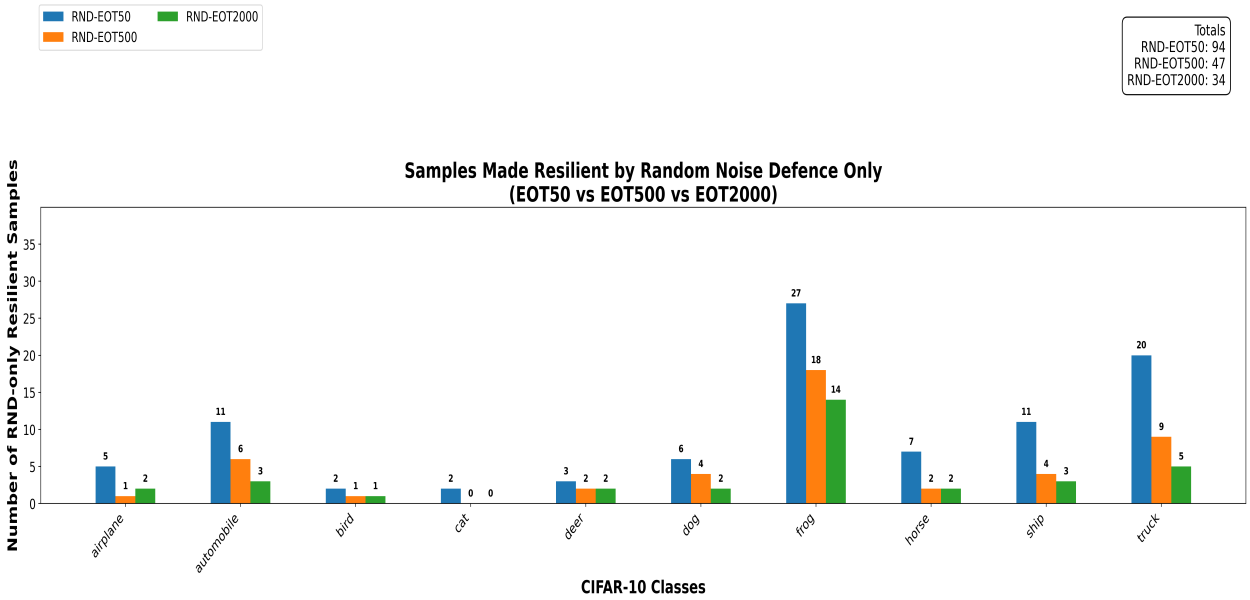
(b) Percentage of samples

Figure 5.2: Per-class samples (a) and per-class percentage of samples (b) made robust by defences such as adversarial training and randomisation. The defensive impact of randomisation reduces as the EOT increases.

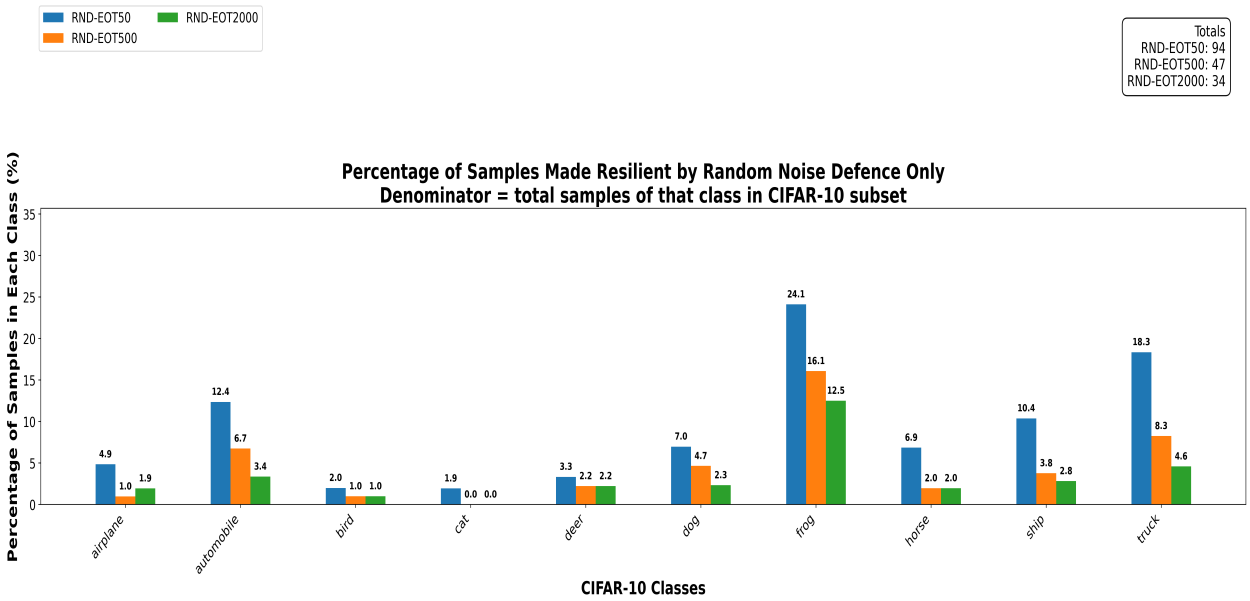
exhibit the highest numbers of resilient samples across all settings. For example, under AdvTrain + RND-EOT50, the truck class has 72 resilient samples, while frog has 65 and automobile has 55 as shown in Figure 5.2. In contrast, classes such as cat, deer, and bird remain considerably less resilient, with resilient counts generally below 25 samples. These observations suggest that certain classes are inherently more robust to adversarial perturbations and benefit more from randomisation-based defences.

When considering only the samples made resilient by Random Noise Defence beyond adversarial training as shown in Figure 5.3a and Figure 5.3b, the class imbalance becomes even more pronounced. RND-EOT50 contributes 94 additional resilient samples over robust samples from AT, whereas RND-EOT500 and RND-EOT2000 contribute only 47 and 34 respectively. The majority of these additional resilient samples over AT originate from a small number of classes. For RND-EOT50, the frog class alone contributes 27 additional resilient samples under EOT-50, while truck has 20 and automobile has 11. Conversely, classes such as cat, bird, and deer have only a handful of additional resilient samples. This demonstrates that the benefits of randomisation are not distributed uniformly across the dataset but are concentrated within specific classes.

The percentage-based analysis as shown in Figure 5.3b reinforces this observation. Under RND-EOT50, approximately 24.1% of all frog samples and 18.3% of all truck samples are newly protected by randomisation, compared with only 2.0% of bird samples and 1.9% of cat samples. As the attacker's EOT budget increases, these gains decline substantially. For example, the fraction of newly resilient frog samples falls from 24.1% under EOT-50 to 12.5% under EOT-2000, while truck drops from 18.3% to 4.6%. However, the class imbalance remains prominent even under EOT-2000.



(a) Number of samples



(b) Percentage of samples

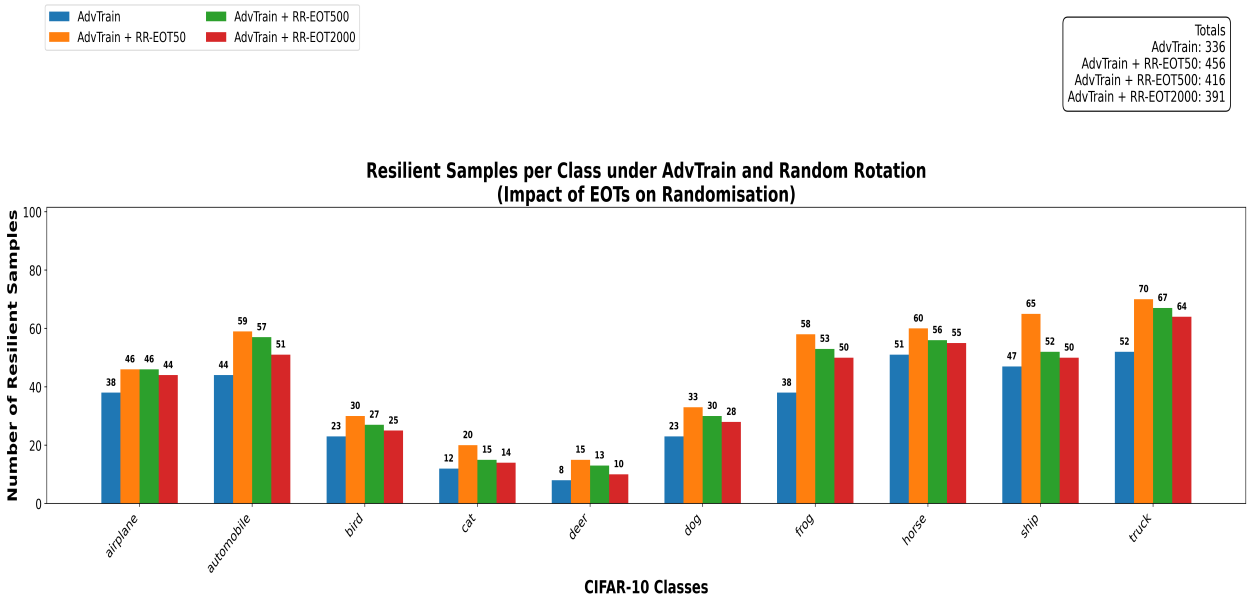
Figure 5.3: Per-class samples (a) and per-class % of samples (b) that remain resilient under various levels of EOTs (RND only). It is evident that randomisation disproportionately effects the classes. Frog and Truck samples are the most effected by randomisation.

5.2.3 Random Rotation Defence under high EOTs

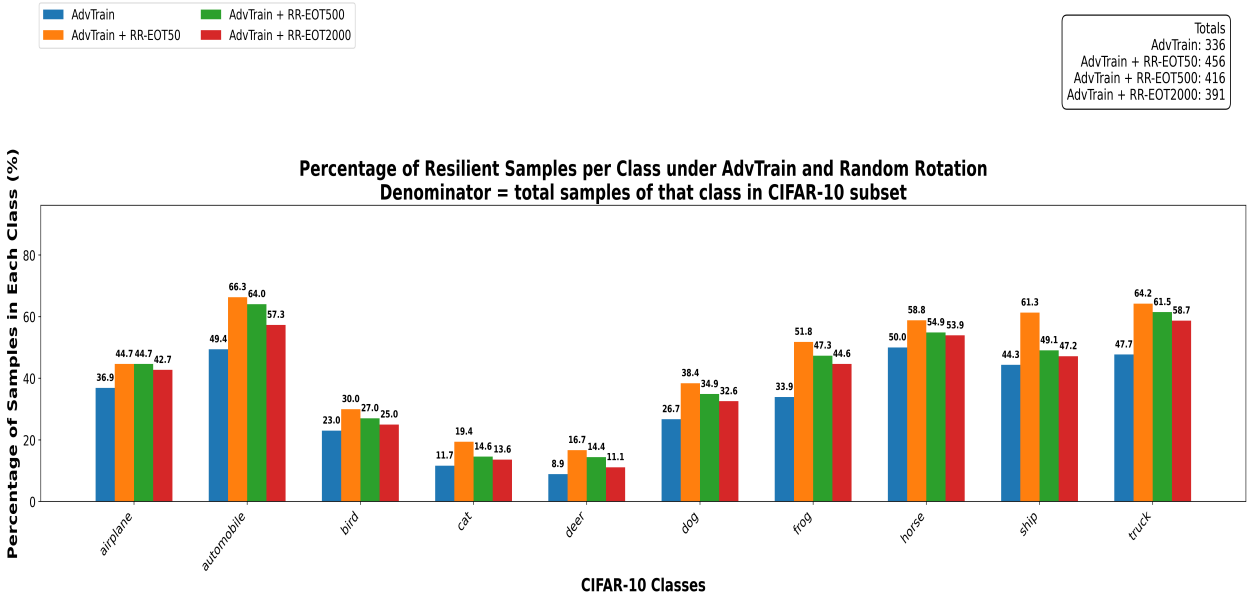
Similar to our evaluation on RND, we now evaluate the high EOTs on random rotation (RR). The results demonstrate that random rotation provides substantial robustness gains beyond adversarial training alone, although these gains decrease as the attacker employs larger EOT budgets. As shown in Figure 5.4a, adversarial training alone yields 336 resilient samples, whereas combining adversarial training with random rotation increases the number of resilient samples to 456 under EOT-50, 416 under EOT-500, and 391 under EOT-2000. While increasing the number of EOTs enables the attacker to better approximate the defence distribution and reduce the effectiveness of the randomisation, all rotation-based settings continue to outperform adversarial training alone. This suggests that random rotation remains a useful source of robustness even against adaptive attacks.

RR-EOT50 contributes 120 additional resilient samples over AT, substantially more than the gains observed under Random Noise Defence. This additional robustness decreases to 80 samples under RR-EOT500 and 55 samples under RR-EOT2000, confirming that stronger adaptive attacks reduce the effectiveness of the defence.

The percentage-based analysis in Figure 5.4b highlights the magnitude of class-specific gains. Under RR-EOT50, 16.9% of automobile samples, 17.9% of frog samples, 17.0% of ship samples, and 16.5% of truck samples become additionally resilient due to random rotation. In comparison, only 7.0% of bird samples and 7.8% of deer samples gain robustness, despite these classes initially exhibiting lower resilience. As the EOT budget increases, these gains steadily diminish. For example, automobile decreases from 16.9% under EOT-50 to 7.9% under EOT-2000, while ship decreases from 17.0% to only 2.8%. This trend demonstrates that adaptive attacks can significantly erode the protection provided by rotation-based randomisation when sufficient EOT samples are used.

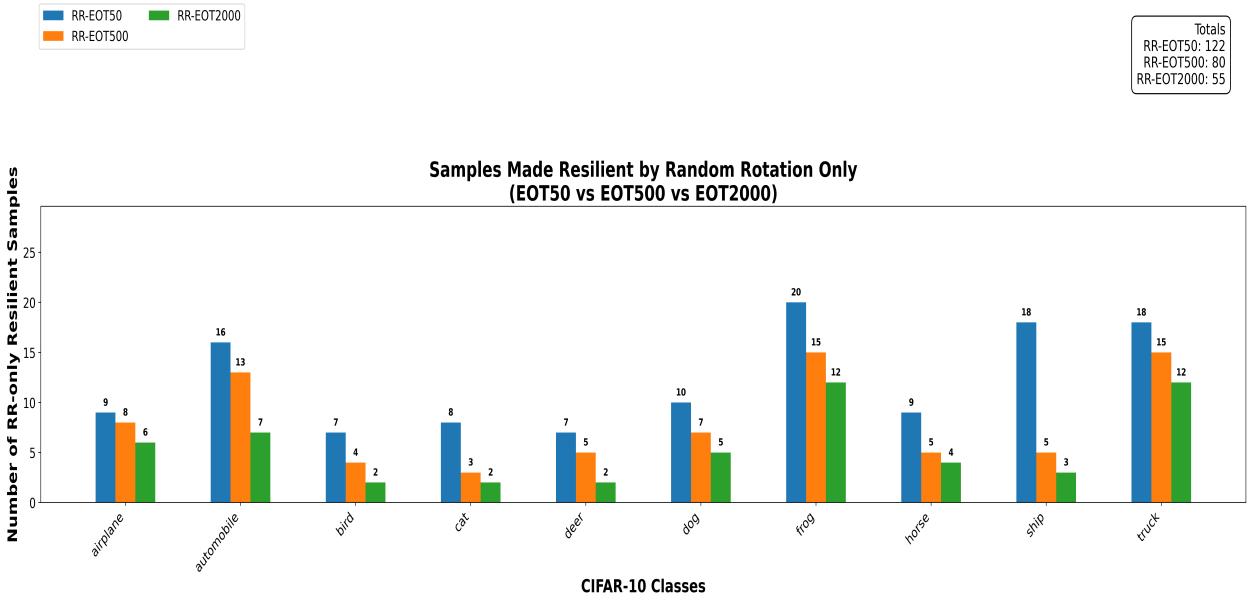


(a) Number of samples

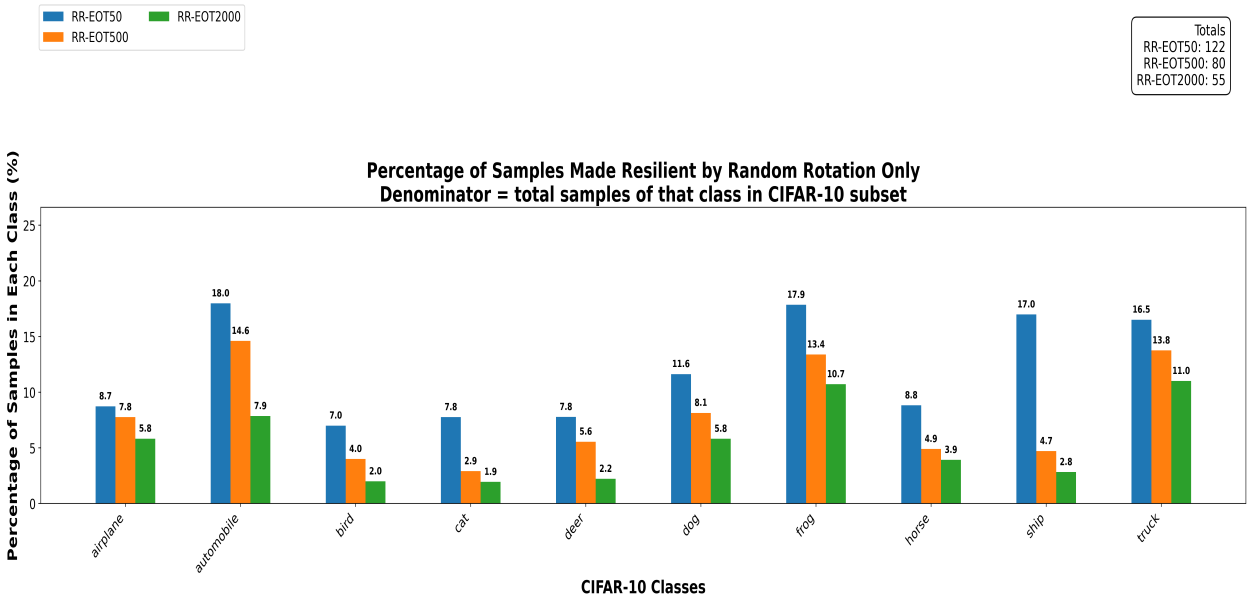


(b) Percentage of samples

Figure 5.4: Per-class samples (a) and per-class % of samples (b) made resilient by defences such as adversarial training and randomisation. The defensive impact of randomisation reduces as the EOT increases.



(a) Number of samples



(b) Percentage of samples

Figure 5.5: Per-class samples (a) and per-class % of samples (b) that remain resilient under various levels of EOTs (only RR improved samples). It is evident that randomisation disproportionately effects the classes. Frog and Truck samples are the most effected by randomisation.

A striking observation is the strong influence of class labels on resilience. Similar to the random noise experiments, classes such as automobile, truck, ship, and frog consistently exhibit the highest numbers of resilient samples. As shown in Figure 5.4a, under RR-EOT50, automobile reaches 59 resilient samples, truck reaches 70, ship reaches 65, and frog reaches 58. In contrast, classes such as cat, deer, and bird remain considerably more vulnerable, with resilient counts generally below 30 samples even under the strongest defence setting. These results suggest that the benefits of random rotation are not distributed uniformly across the dataset but instead depend heavily on individual classes.

Examining only the samples made resilient by random rotation beyond adversarial training in Figure 5.5a and Figure 5.5b reveals even stronger class-dependent behaviour. The largest gains occur for frog, truck, ship, and automobile, which contribute 20, 18, 18, and 16 additionally resilient samples respectively under RR-EOT50. Conversely, classes such as cat, deer, and bird contribute relatively few additional resilient samples. Under EOT-2000 Frog and Truck each have 12 resilient samples while the rest are under 7. These findings indicate that random rotation primarily benefits a subset of classes rather than improving robustness uniformly across the entire dataset.

5.2.4 Highly confident misclassified samples under RND

Simply introducing randomisation during inference time increases robust accuracy which a resourceful attacker can reduce by using high EOTs. In the following experiments, we investigate the source of the observed robustness gains in the presence of randomisation. Specifically, we examine whether the improved robustness arises because input randomisation misguides the attacker during optimisation, or whether it is an inherent property of the input itself. In the latter case, certain images or classes may be more susceptible to randomisation, such that even slight perturbations introduced by the defence alter the

model's decision boundary and change the predicted label independently of the attacker's optimisation process.

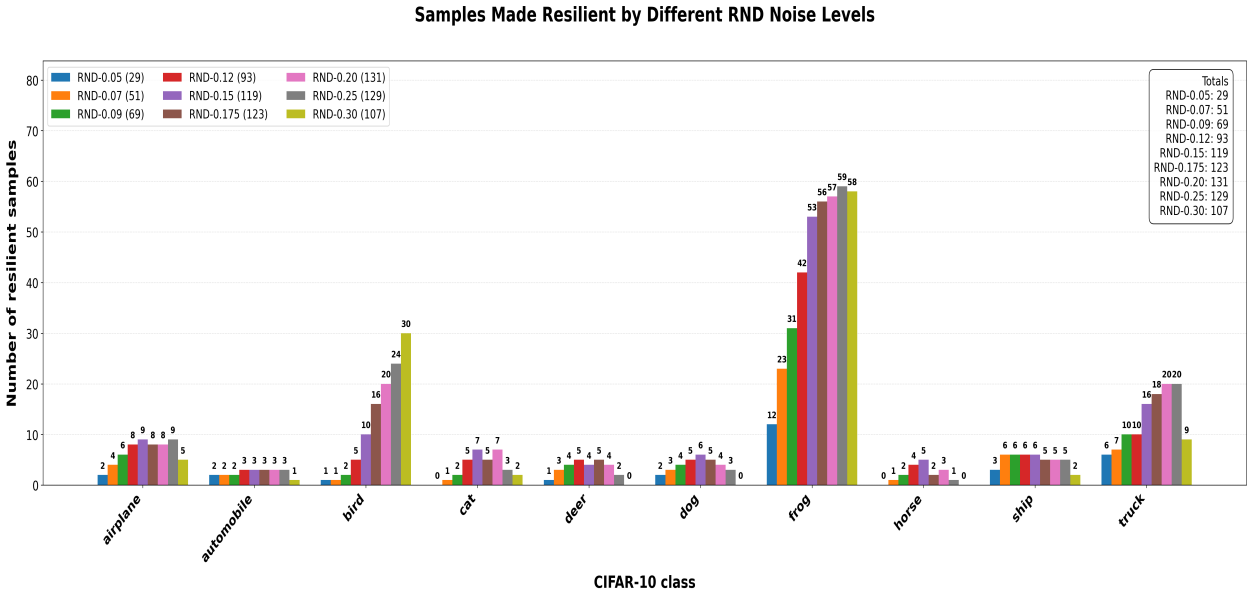


Figure 5.6: Per-class % of samples that remain resilient under various levels of RND magnitudes. It is evident that randomisation disproportionately effects the classes.

To get some clues, we launch an experiment where square attack is used to generate highly confident misclassified samples. The samples are generated without using randomisation in the forward passes because randomisation can disrupt the black-box attacker from confident adversarial regions as shown in Table 3.2 of Chapter 3. Upon misclassification, the confidence in misclassification is increased so randomisation is less likely to flip the classification into the true class. Randomised transformations and noise can flip the classification of samples generated through black-box attacks. The effect is more prominent for black-box attacks that stop optimisation upon misclassification - leaving the sample near the boundary. Confidence increasing attacks aim to reduce the flipping by increasing the misclassification confidence after misclassification occurs. Once the highly confident misclassified samples are generated using square attack, we apply RND on the samples to check how many flip from the wrong class into the correct class.

As shown in Figure 5.6, with a simple gaussian noise magnitude of 0.05, 2.9% of the 1000 samples flip into the correct class. This flipping increases as the noise magnitude increases but peaks at magnitude of 0.25. However, even in this case there is a significant imbalance of the resilient samples at the class-level. The Frog class leads to the highest gains along with truck. Surprisingly, Bird class is also seen to make significant gains due to the noise. The results suggest that there is something with the nature of frog images, which randomisation exploits that leads to robustness gains.

Furthermore, input randomisation methods as described in Chapter 3, aim to increase the defensive abilities of the model by disrupting or misguiding the attack optimisation process of query-based attacks. Qin et al. (2021) asserts that the success of input randomisation comes from misguiding the attacker. However, even under a resourceful attacker (EOT-2000), there are still some samples that remain resilient under RND (34 samples). This raises the question: what causes these samples to remain resilient? Therefore, in the following experiments, we investigate the factors underlying this robustness gain.

To do this, we launch multiple attacks. First, we get the number of samples that were resilient after an EOT-2000 square attack with RND as the inference-time input randomisation defence, which we name RND-EOT2000. Additionally, we generate high confidence adversarial examples using the confident square attack without inference time randomisation in the forward pass. Unlike Chapter 3, where introducing randomisation during inference stage led to robustness gains, here we aim to reduce the robustness gains from randomisation. Therefore, not using randomisation during inference stage allows the attack to create highly confident adversarial examples. The entire purpose of generating these examples is to reduce the chance of a sample flipping classification into the correct class under randomisation. Once such examples are generated we apply random perturbations on them such as RND with noise $\delta = 0.05$ or Random Rotation (RR-10) to see if randomisation causes the samples to flip into the correct class.

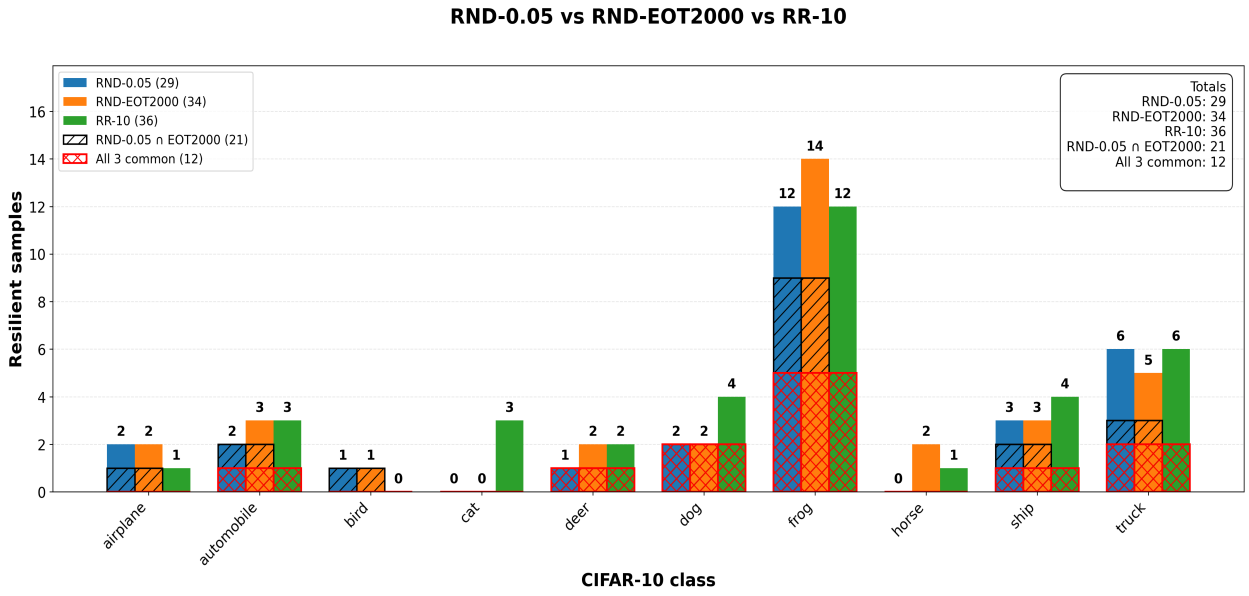


Figure 5.7: Per-class resilient number of samples under various randomisation techniques.

The results are presented in Figure 5.7. A sample is counted only if it was misclassified under adversarial training (AT) but became resilient when randomisation was applied. Among all classes, the frog class exhibits the highest number of resilient samples. Interestingly, approximately 21 of the 34 resilient samples identified under RND-EOT2000 are also resilient under standard RND ($\delta = 0.05$). This substantial overlap suggests that the robustness of many of these samples cannot be attributed solely to disrupting the attacker's optimisation process. Instead, it indicates that, for a significant subset of samples, randomisation consistently alters the model's prediction, potentially due to intrinsic properties of the input images and their interaction with the model's decision boundaries.

5.2.5 Practical Feasibility of EOT-2000 Attacks

The EOT-2000 configuration used in this study incurs a substantial computational cost. Specifically, the attack performs 10,000 optimisation iterations, with each iteration evaluating 2,000 transformed samples, resulting in approximately 20 million model evaluations.

Although such a query budget is practical in an offline research setting, it represents a significant computational burden in real-world deployment scenarios. Nevertheless, this level of computation may still be feasible for highly resourceful adversaries, such as government-sponsored actors or other well-funded organisations with access to large-scale computing infrastructure. Accordingly, to evaluate the defence under a stringent threat model, we assume a similarly well-resourced defender, representative of large technology companies, that can sustain the computational overhead required to provide model predictions while deploying computationally intensive defence mechanisms.

5.3 SDM+Randomisation under Resourceful Attacker

We have seen in Chapter 4 that stateful detection when coupled with input randomisation induces synergistic effects on robustness gains. However, in the previous experimental settings in Chapter 4, we assumed a small query budget of 700 with an EOT of 3. In comparison to traditional black-box attack budgets in the literature, this is a relatively small query budget.

Therefore, in this section, we assess the limitations of combining SDM+randomisation as a defensive strategy under increased query budget. To do this, we sample 5000 images randomly from *Open Images* v7 release (Krasin et al., 2017; Google Research, 2021). Furthermore, we increase the EOT from 3 to 20 and the central patch size is reduced to 130 to ensure detection rates are low and optimisation succeeds.

Based on the experiments, we find that 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. Increasing the attacker query budget is more impactful for adversarial success. However, the attacker must

Table 5.1: Comparison of low and high query budget settings for evaluating SDM+Randomisation (RRC) under DazzlePatch attack.

Parameter	**Low Budget Setting**	**High Budget Setting**
Number of Host Images	700	5,000
EOT Samples	3	20
Query Budget	700	5,000
Patch Size	160×160	130×130
Perturbation Budget (ϵ)	0.05	0.05
Attack Success Rate (ASR)	11.00%	**17.00**%
Hit rate	55.90%	**16.87**%

expend a high number of queries to achieve it. These results support the previous results on randomisation i.e. randomisation can be used to make the job of the attacker harder. However, if the attacker is able to expend resources then they are likely to be successful.

5.4 Conclusion

Overall, this chapter investigates the impact of increasing the attacker's query budget to evaluate the limitations of black-box defences and address **RQ4** of this thesis, "How do the defences against black-box attacks scale down as the attacker becomes significantly more resourceful?" The results demonstrate that larger query budgets substantially improve attack success rates against these defences. However, this increased effectiveness comes at the cost of requiring significantly more queries, making such attacks computationally expensive.

These findings reveal a limitation of randomisation-based defences: the robustness

gains they provide can be substantially reduced by a sufficiently resourceful adaptive attacker. As a result, while randomisation may increase the cost of mounting a successful attack, it does not provide reliable or guaranteed security against adversaries with sufficient query budgets.

Chapter 6

Conclusions

This thesis investigated how machine learning systems can be defended against query-based black-box attacks, where the adversary has no access to model parameters but can iteratively refine an attack through repeated queries. The central aim of the thesis is to investigate the challenges associated with mitigating query-based black-box attacks.

To investigate the challenges, the thesis examined the two most established black-box defence families identified in the literature: processing-based input randomisation and detection-based stateful defence models (SDMs). It then evaluated both their practical benefits and their failure modes under stronger adaptive attacks. The overall conclusion is that transformation-based randomisation is more effective than noise-based randomisation, and that combining randomisation with SDMs can further improve robustness in some settings. However, neither randomisation nor SDMs, alone or in combination, provide complete security against a resourceful adaptive attacker.

6.1 Answers to the Research Questions

RQ1: What are the key approaches proposed for countering query-based black-box attacks?

The systematic literature review showed that the most established defence approaches

proposed against query-based black-box attacks are processing defences and detection-based defences, with architecture-based approaches emerging more recently as a third line of work. Randomisation-based processing methods aim to disrupt the attacker's optimisation process by introducing variability into inputs, intermediate representations, or outputs. Detection defences instead attempt to detect attack behaviour by analysing query history and identifying suspicious patterns of similarity or progression. This taxonomy provides the conceptual foundation of the thesis and explains why the empirical investigation focused on pre-processing based randomisation and detection-based stateful defence models as the two dominant established paradigms.

RQ2: Given that randomisation is intended to mislead attackers and reduce the effectiveness of their queries, can we use non-noise based inference-time randomisation to improve mitigation of query-based attacks?

Among the methods evaluated, transformation-based randomisation, especially Random Rotation (RR) and Random Resize-Cropping (RRC), performs better than Random Noise Defence (RND) under the tested budgets and datasets. On ImageNet (Russakovsky et al., 2015), under the EOT-10 Square Attack, inference-time Random Resize-Cropping (RRC) and Random Rotation improve the robust accuracy of a ResNet-50 model (He et al., 2016) over RND (Qin et al., 2021) by 9.4% and 7.5%, respectively. On Tiny ImageNet (Deng et al., 2009), using a PreActResNet18 model, Random Resize-Cropping (RRC) improves robust accuracy over RND by 1.6%, while Random Rotation improves it by 0.8%. On CIFAR-10 (Krizhevsky et al., 2009), using a ResNet-50 model, Random Rotation improves robust accuracy over RND by 0.9%, while Random Resize-Cropping (RRC) matches the performance of RND.

Even under strong attack conditions with higher EOT on CIFAR-10, the proposed transformations outperform RND by around 3.5%. Therefore, the answer to RQ2 is that

transformation-based randomisation is the most effective inference-time randomisation strategy evaluated in this thesis.

RQ3: Can detection-based defences against query-based black-box attacks provide complete robustness, as suggested by prior work?

The answer is no. Recent claims of 100% robustness provided by SDMs offer a false promise. We found that modifying square attack can easily bypass two of the defences. However, the adversarially trained SDM cannot be bypassed by simple modifications to existing attacks. Therefore, we proposed DazzlePatch attack and showed that robust SDMs can still be bypassed. These findings show that detection-based defences remain dependent on assumptions about how attacks traverse the input or representation space, and those assumptions can be targeted once the adversary is explicitly optimising for evasion.

The thesis further showed that combining SDMs with randomisation improves robustness relative to either defence alone. However, as the attacker is allowed larger query budgets and stronger EOT settings, the robustness gains of SDM+randomisation diminish.

RQ4: How do the defences scale down as the attacker becomes significantly more resourceful?

The limitation identified by this thesis is that existing defences do not provide foolproof security; they only interfere with how efficiently the attacker can achieve their objective. Randomisation attempts to corrupt the optimisation signal, while SDMs attempt to detect the optimisation process through repeated queries. In both cases, a sufficiently adaptive attacker can recover ground by increasing query budgets and using larger EOT values. We find that high-query adaptive attacks substantially diminish the robustness gains provided by input randomisation, reducing the improvement from 9.4% to 3.4%. In contrast, adaptive attacks increase the attack success rate (ASR) against the combined SDM+randomisation

defence by approximately 6.0%.

6.2 Contributions to Knowledge

This thesis makes four principal contributions to knowledge.

First, it provides a systematic literature review and taxonomy of defences against query-based black-box attacks, clarifying how the field is organised and highlighting the relative maturity of randomisation-based and detection-based approaches.

Second, it demonstrates empirically that transformation-based randomisation is a stronger and more practical mitigation strategy than additive noise-based randomisation across the evaluated settings.

Third, it contributes a new adaptive attack, DazzlePatch, for evaluating recent stateful defence models. This attack challenges the claim that OARS-resistant SDMs provide complete robustness and shows that even advanced detectors remain vulnerable under stronger adaptive evaluation. We find that as the defence methods become sophisticated, there may not be a universally common attack that beats them all. However, by presenting the DazzlePatch attack, we show that crafty attackers can find loopholes in defences evaluated only against the popular attack methods.

Fourth, it advances the thesis-level understanding of black-box robustness by showing that current defences should not be understood as foolproof defences, but as mechanisms that trade off attacker cost. It shows that highly resourced adversaries can diminish the benefits of randomisation by using high EOTs.

6.3 Limitations

The findings of this thesis should be interpreted in light of several limitations and threats to validity. A first limitation concerns the computational scope of the evaluations. The adaptive evaluations considered in this thesis employ demanding attack settings, including different EOT configurations (EOT-10, EOT-50, and EOT-2000), which require substantial query budgets and computational resources. These settings were selected to provide a rigorous assessment of defence robustness by stress-testing the evaluated methods under strong attacker capabilities. However, the associated computational cost limits the extent to which such experiments can be repeated multiple times. Consequently, the evaluations are based on single experimental runs.

The second limitation is that some evaluations were conducted using subsets of the available datasets, with 200 and 1,000 samples selected depending on the experimental setting. These sample sizes were chosen to enable computationally feasible evaluations while maintaining challenging attack scenarios under large query budgets. In particular, for DazzlePatch, a subset of 200 samples was used due to the high computational cost associated with evaluating multiple attack initialisations, where each sample was tested using 21 random restarts. Therefore, the objective was not merely to measure average-case performance, but to rigorously stress-test the evaluated defences under demanding attacker settings. Nevertheless, evaluations on larger portions of the datasets would provide further evidence regarding the generalisability of the observed trends.

A third limitation relates to the breadth of the experimental evaluation. While this thesis considers CIFAR-10, Tiny ImageNet, ImageNet, and additional datasets for SDM false-positive analysis, these benchmarks represent only a subset of the environments in which adversarial attacks and defences may ultimately be deployed. These datasets were chosen because the original papers proposing the defences (Qin et al., 2021; Wang et al.,

2024; Park et al., 2025) also evaluated their methods on these datasets, thereby enabling direct comparison with prior work. Nevertheless, results obtained on these benchmarks may not fully capture behaviour in practical deployment scenarios, where image distributions, model architectures, and query patterns may differ substantially from those considered in standard benchmark evaluations.

In addition, the benchmark defence methods were not evaluated under identical experimental settings. Specifically, QPA and AdvQDet were evaluated on the ImageNet dataset under the proposed DazzlePatch attack, whereas GWAD was evaluated on CIFAR-10 using the standard Square Attack. Consequently, direct quantitative comparisons between these methods should be interpreted with caution, as differences in dataset characteristics, including image complexity, class diversity, and model architecture, can influence both attack success and defence robustness.

Despite these differences, the qualitative trends observed are expected to extend beyond the datasets considered, as the underlying attack mechanisms are not inherently dataset-specific. However, this hypothesis should be validated through experiments on a broader range of datasets and experimental settings. In particular, evaluating QPA and AdvQDet on CIFAR-10 and assessing DazzlePatch on datasets beyond ImageNet would provide stronger evidence for the generalisability of the proposed attack while enabling more comprehensive and directly comparable evaluations across benchmark defences.

6.4 Future Work

Several directions for future work follow directly from these findings.

First, as argued in Chapter 4, a rigorous standard of adaptive evaluation must be

adopted, so that robustness claims can be assessed under a more unified protocol.

Second, while this thesis evaluated a subset of randomised transformations, including random rotation, random resizing and padding, and random resize-cropping, other transformations such as zoom, warping, brightness, contrast, and saturation warrant further investigation. We validated our choice of transformations by demonstrating that they do not degrade image quality to the same extent as RND. Our hypothesis was that the selected random transformations, such as random resize-cropping and random rotation, preserve image quality better than RND while still providing defensive benefits. This work paves the way for exploring additional transformations for defence purposes; however, we leave such investigations to future work.

Third, there is a need for defences that provide stronger guarantees than optimisation disruption alone. This includes more principled combinations of adversarial training and randomisation with certified defences that guarantee robustness without relying purely on obscuring the attack signal.

Fourth, as the defences improve, the attacker has to rely on generating adversarial examples that do not remain imperceptible to be successful. Therefore, an interesting direction would be to develop attacks that generate adversarial examples that remain imperceptible and evade detection, even by a human observer.

Finally, the strong class-dependent effects observed in this thesis suggest a deeper question about why some samples benefit disproportionately from randomisation and adversarial training. Investigating this phenomenon could lead to class-aware defences that are more principled than uniform randomisation.

6.5 Implications for Knowledge and Practice

This thesis has significant implications for both knowledge and practice in adversarial machine learning. For knowledge, it advances the field's understanding of black-box robustness by empirically demonstrating that transformation-based randomisation (e.g., Random Rotation and Random Resize-Cropping) is a more effective mitigation strategy than additive noise-based approaches such as RND, and by introducing DazzlePatch as a novel adaptive attack that disproves the prevailing claim that stateful defence models (SDMs) can offer complete, foolproof robustness against query-based black-box attacks. More fundamentally, it reframes how robustness should be conceptualised: rather than treating randomisation and detection-based defences as absolute safeguards, the thesis shows that they function as cost-increasing and success-delaying mechanisms whose benefits are highly class-dependent and diminish substantially once an adversary becomes sufficiently resourceful (e.g., via larger query budgets and higher EOT values).

For practice, these findings carry direct implications for practitioners and organisations deploying machine learning systems as a service: defences should not be marketed or relied upon as guarantees of security, evaluation protocols must incorporate rigorous, standardised adaptive and high-query-budget testing before deployment claims are made, and practical deployments would benefit from combining randomisation with stateful detection while still assuming that a well-resourced attacker (e.g., a government-level or well-funded adversary) can eventually erode these protections. Consequently, practitioners should prioritise raising attacker cost and reducing attack reliability over pursuing unattainable complete robustness, while committing to continuous, adaptive re-evaluation of deployed defences as attack methodologies evolve.

6.6 Final Remarks

In summary, this thesis shows that defending against query-based black-box attacks remains an open and difficult problem. However, the thesis advances the state-of-the-art in these defences by proposing transformation-based randomisation as a more effective defence than additive noise-based randomisation. Modern stateful defence models can improve robustness and, when combined with randomisation, can further raise attacker cost, yet they still do not provide complete protection against a sufficiently resourceful adversary.

The broader thesis-level conclusion is therefore clear: the field has made meaningful progress in increasing the difficulty for query-based black-box attacks, but current methods do not yet justify claims of complete robustness. The most defensible interpretation of the evidence is that these defences increase cost, delay success, and sometimes reduce attack reliability, but they do not fundamentally resolve the security problem. Developing defences that remain reliable under adaptive evaluation while preserving practical usability remains a central challenge for future research.

References

Addepalli, Sravanti, Samyak Jain, Gaurang Sriramanan, and R Venkatesh Babu (2022). "Scaling adversarial training to large perturbation bounds". In: *European Conference on Computer Vision*. Springer, pp. 301–316.

Aithal, Manjushree B and Xiaohua Li (2022a). "Boundary defense against black-box adversarial attacks". In: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 2349–2356.

— (2022b). "Mitigating black-box adversarial attacks via output noise perturbation". In: *IEEE Access* 10, pp. 12395–12411.

Alfarra, Motasem, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem (2022). "Combating adversaries with anti-adversaries". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 6, pp. 5992–6000.

Ali, Ziad Tariq Muhammad, R Muhammad Atif Azad, Muhammad Ajmal Azad, James Holyhead, Iain Rice, and Ali Shariq Imran (2025). "Random transformations to improve mitigation of query-based black-box attacks". In: *Expert Systems with Applications* 264, p. 125840.

Ali, Ziad Tariq Muhammad, Raja Muhammad Atif Azad, Muhammad Ajmal Azad, Iain Rice, Umar Daraz, Ali Shariq Imran, and James Holyhead (2026). "On Evaluating Stateful Defence Models against Query-Based Black-Box Attacks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 809–818.

Alshanqiti, Abdullah, Ali Ismail Dumlu, Abdullah Ahmed, and Aeshah Alsughayyir (2022). "Protecting Image-classifiers from Adversarial Attacks using an Adaptive Clustering

Buffer-based Detection". In: *2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*. IEEE, pp. 94–99.

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

Athalye, Anish, Nicholas Carlini, and David A. Wagner (2018). "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*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 274–283.

Atik, Saziye Ozge and Cengizhan Ipbuker (2021). "Integrating convolutional neural network and multiresolution segmentation for land cover and land use mapping using satellite imagery". In: *Applied Sciences* 11.12, p. 5551.

Bokobza, Roey and Yisroel Mirsky (2025). "Counter-Samples: A Stateless Strategy to Neutralize Black-Box Adversarial Attacks". In: *ACM Transactions on Intelligent Systems and Technology* 16.4, pp. 1–23.

Byun, Junyoung, Hyojun Go, and Changick Kim (2022). "On the Effectiveness of Small Input Noise for Defending Against Query-based Black-Box Attacks". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3051–3060.

Carlini, Nicholas, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin (2019). "On evaluating adversarial robustness". In: *arXiv preprint arXiv:1902.06705*.

Carmon, Yair, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi (2019). "Unlabeled data improves adversarial robustness". In: *arXiv preprint arXiv:1905.13736*.

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

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

Chen, Sizhe, Zhehao Huang, Qinghua Tao, Yingwen Wu, Cihang Xie, and Xiaolin Huang (2022). "Adversarial attack on attackers: Post-process to mitigate black-box score-based query attacks". In: *Advances in Neural Information Processing Systems* 35, pp. 14929–14943.

Chen, Steven, Nicholas Carlini, and David Wagner (2020c). "Stateful detection of black-box adversarial attacks". In: *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pp. 30–39.

Choi, Seok-Hwan, Jinmyeong Shin, and Yoon-Ho Choi (2023). "PIHA: Detection method using perceptual image hashing against query-based adversarial attacks". In: *Future Generation Computer Systems* 145, pp. 563–577.

Croce, Francesco, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein (2021). "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*. Ed. by Joaquin Vanschoren and Sai-Kit Yeung. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/a3c65c2974270fd093ee8a9bf8ae7d0b-Abstract-round2.html.

Croce, Francesco and Matthias Hein (2020). "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks". In: *International conference on machine learning*. PMLR, pp. 2206–2216.

De, Palma Alessandro, Durand Serge, Chihani Zakaria, Terrier François, and Caterina Urban (2025). "On Using Certified Training towards Empirical Robustness". In: *Transactions on Machine Learning Research Journal*.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Ding, Gavin Weiguang, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang (2018). "Mma training: Direct input space margin maximization through adversarial training". In: *arXiv preprint arXiv:1812.02637*.

Dong, Yinpeng, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu (2020). "Benchmarking adversarial robustness on image classification". In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 321–331.

Douceur, John R (2002). "The sybil attack". In: *International workshop on peer-to-peer systems*. Springer, pp. 251–260.

Engstrom, Logan, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras (2019). *Robustness (Python Library)*. URL: https://github.com/MadryLab/robustness.

Esmaeili, Bardia, Amin Azmoodeh, Ali Dehghantanha, Hadis Karimipour, Behrouz Zolfaghari, and Mohammad Hammoudeh (2022). "IIoT deep malware threat hunting: from adversarial example detection to adversarial scenario detection". In: *IEEE Transactions on Industrial Informatics* 18.12, pp. 8477–8486.

Farshadfar, Elahe, Amir Mahdi Sadeghzadeh Mesgar, and Rasool Jalili (2025). "Divergent Twins Fencing: Protecting Deep Neural Networks Against Query-based Black-box Adversarial Attacks." In: *ISeCure* 17.2.

Feng, Ryan, Ashish Hooda, Neal Mangaokar, Kassem Fawaz, Somesh Jha, and Atul Prakash (2023a). "Investigating Stateful Defenses Against Black-Box Adversarial Examples". In: *arXiv preprint arXiv:2303.06280*.

Feng, Ryan, Ashish Hooda, Neal Mangaokar, Kassem Fawaz, Somesh Jha, and Atul Prakash (2023b). "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*, pp. 786–800.

Fu, Yanzhang, Zizheng Guo, and Jizhou Luo (2026). "Dashed Line Defense: Plug-And-Play Defense Against Adaptive Score-Based Query Attacks". In: *arXiv preprint arXiv:2602.08679*.

Gao, Yue, Ilia Shumailov, Kassem Fawaz, and Nicolas Papernot (2022). "On the Limitations of Stochastic Pre-processing Defenses". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. URL: https://openreview.net/forum?id=P_eBjUlzlV.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2015). "Explaining and Harnessing Adversarial Examples". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Google Research (2021). *Open Images V7*. https://storage.googleapis.com/openimages/web/index.html. Dataset with 9 million images annotated with image-level labels, object bounding boxes, segmentation masks, and visual relationships.

Gowal, Sven, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli (2020). "Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples". In: *arXiv preprint arXiv:2010.03593*. URL: https://arxiv.org/pdf/2010.03593.

Gu, Jindong, Fangyun Wei, Philip Torr, and Han Hu (2023). "Exploring Non-additive Randomness on ViT against Query-Based Black-Box Attacks". In: *arXiv preprint arXiv:2309.06438*.

Guo, Chuan, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten (2018). "Countering Adversarial Images using Input Transformations". In: *International Conference on Learning Representations*.

Han, Xueyuan, Thomas Pasquier, Adam Bates, James Mickens, and Margo Seltzer (2020). "Unicorn: Runtime provenance-based detector for advanced persistent threats". In: *arXiv preprint arXiv:2001.01525*.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Huang, Zhichao and Tong Zhang (2019). "Black-box adversarial attack with transferable model-based embedding". In: *arXiv preprint arXiv:1911.07140*.

Ilyas, Andrew, Logan Engstrom, Anish Athalye, and Jessy Lin (2018). "Black-box adversarial attacks with limited queries and information". In: *International conference on machine learning*. PMLR, pp. 2137–2146.

Jia, Wei, Zhaojun Lu, Runze Yu, Liaoyuan Li, Haichun Zhang, and Zhenglin Liu (2023). "AND: Adaptive Noise to Defend Query-Based Black-Box Attacks Without Affecting Normal Inference". In: *Available at SSRN 4596212*.

Juuti, Mika, Sebastian Szyller, Samuel Marchal, and N Asokan (2019). "PRADA: protecting against DNN model stealing attacks". In: *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, pp. 512–527.

Khan, U, U Nawaz, A Qayyum, S Ashraf, Y Xie, MH Khan, M Bilal, and J Qadir (2026). *Surgical scene understanding in the era of foundation AI models: a comprehensive review (2025)*.

King, Samuel T and Peter M Chen (2003). "Backtracking intrusions". In: *Proceedings of the nineteenth ACM symposium on Operating systems principles*, pp. 223–236.

Krasin, Ivan, 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 (2017). *OpenImages: A public dataset for large-scale*

multi-label and multi-class image classification. Dataset release, https://storage.googleapis.com/openimages/web/index.html.

Krizhevsky, Alex and Geoffrey Hinton (2009). *Learning Multiple Layers of Features from Tiny Images*. Tech. rep. University of Toronto.

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

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

Li, Shaofei, Ziqi Zhang, Haomin Jia, Yao Guo, Xiangqun Chen, and Ding Li (2025). "Query Provenance Analysis: Efficient and Robust Defense against Query-based Black-box Attacks". In: *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 1641–1656.

Lucas, Keane, Matthew Jagielski, Florian Tramèr, Lujo Bauer, and Nicholas Carlini (2023). "Randomness in ML Defenses Helps Persistent Attackers and Hinders Evaluators". In: *arXiv preprint arXiv:2302.13464*.

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu (2018). "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*.

Maho, Thibault, Teddy Furon, and Erwan Le Merrer (2021). "Surfree: a fast surrogate-free black-box attack". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10430–10439.

Maung, Maung, April Pyone, and Hitoshi Kiya (2020). "Encryption inspired adversarial defense for visual classification". In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 1681–1685.

McNemar, Quinn (1947). "Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika* 12.2, pp. 153–157.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement". In: *Bmj* 339.

Nguyen, Hung Quang, Yingjie Lao, Tung Pham, Kok-Seng Wong, and Khoa Doan (2024). "Understanding the robustness of randomized feature defense against query-based adversarial attacks". In: *International Conference on Learning Representations*. Vol. 2024, pp. 54019–54039.

Nicolae, Maria-Irina, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards (2018). "Adversarial Robustness Toolbox v1.2.0". In: *CoRR* 1807.01069. URL: https://arxiv.org/pdf/1807.01069.

Nirala, Ashutosh Kumar, Jin Tian, Olukorede Fakorede, and Modeste Atsague (2025). "AlignFix: Fixing Adversarial Perturbations by Agreement Checking for Adversarial Robustness against Black-box Attacks". In: *Transactions on Machine Learning Research*.

Pang, Ren, Xinyang Zhang, Shouling Ji, Xiapu Luo, and Ting Wang (2020). "Advmind: Inferring adversary intent of black-box attacks". In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1899–1907.

Park, Jeonghwan, Niall McLaughlin, and Ihsen Alouani (2025). "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*, pp. 10235–10243.

Qin, Zeyu, Yanbo Fan, Hongyuan Zha, and Baoyuan Wu (2021). "Random noise defense against query-based black-box attacks". In: *Advances in Neural Information Processing Systems* 34, pp. 7650–7663.

Qin, Zeyu, Xuanchen Yan, and Baoyuan Wu (2022). *BlackboxBench (Python Library)*. URL: https://github.com/SCLBD/BlackboxBench.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763.

Raff, Edward, Jared Sylvester, Steven Forsyth, and Mark McLean (2019). "Barrage of random transforms for adversarially robust defense". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6528–6537.

Rashid, Aqib and Jose Such (2023). "Malprotect: Stateful defense against adversarial query attacks in ml-based malware detection". In: *IEEE Transactions on Information Forensics and Security* 18, pp. 4361–4376.

Rebuffi, Sylvestre-Alvise, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann (2021). "Fixing data augmentation to improve adversarial robustness". In: *arXiv preprint arXiv:2103.01946*.

Rice, Leslie, Eric Wong, and Zico Kolter (2020). "Overfitting in adversarially robust deep learning". In: *International conference on machine learning*. PMLR, pp. 8093–8104.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3, pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

Schmidt, Ludwig, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry (2018). "Adversarially Robust Generalization Requires More Data". In: *NeurIPS*, pp. 5019–5031.

Sitawarin, Chawin, Zachary J Golan-Strieb, and David Wagner (2022a). "Demystifying the adversarial robustness of random transformation defenses". In: *International Conference on Machine Learning*. PMLR, pp. 20232–20252.

Sitawarin, Chawin, Florian Tramèr, and Nicholas Carlini (2022b). "Preprocessors Matter! Realistic Decision-Based Attacks on Machine Learning Systems". In: *arXiv preprint arXiv:2210.03297*.

Stallkamp, Johannes, Marc Schlipsing, Jan Salmen, and Christian Igel (2012). "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition". In: *Neural networks* 32, pp. 323–332.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus (2014). "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler (2018). "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific data* 5.1, pp. 1–9.

Tuna, Ömer Faruk, Leyli Karaçay, and Utku Gülen (2024). "A Novel Method to Mitigate Adversarial Attacks Against AI-as-a-Service Functionality". In: *2024 IEEE Middle East Conference on Communications and Networking (MECOM)*. IEEE, pp. 181–186.

Vo, Quoc Viet, Bao Gia Doan, Ehsan Abbasnejad, and Damith Ranasinghe (2025). "Can Model Randomization Offer Robustness Against Query-Based Black-Box Attacks?" In: URL: https://openreview.net/forum?id=DpnY7VOktT.

Wang, Kaijie, Yingwen Wu, Jie Yang, and Xiaolin Huang (2025a). "Towards Robust Synthetic Aperture Radar Classification: Counteracting Black-Box Adversarial Attacks". In: *IET Radar, Sonar & Navigation* 19.1, e70062.

Wang, Lina, Kang Yang, Wenqi Wang, Run Wang, and Aoshuang Ye (2020). "MGAAttack: Toward More Query-efficient Black-box Attack by Microbial Genetic Algorithm". In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2229–2236.

Wang, Xin, Kai Chen, Xingjun Ma, Zhineng Chen, Jingjing Chen, and Yu-Gang Jiang (2024). "AdvQDet: Detecting query-based adversarial attacks with adversarial contrastive prompt tuning". In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6212–6221.

Wang, Yongzhi, Ahsan Habib, Likhitha Reddy Kesara, Brahmarshi Jasti, Renjie Hu, and Tejasv Singh (2025b). "Detecting Black-box Model Probing Attacks Through Probability Scores". In: *IEEE Access*.

Wang, Zekai, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan (2023). "Better diffusion models further improve adversarial training". In: *arXiv preprint arXiv:2302.04638*.

Wang, Zhou, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli (2004). "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4, pp. 600–612.

Wohlin, Claes (2014). "Guidelines for snowballing in systematic literature studies and a replication in software engineering". In: *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pp. 1–10.

Wu, Boxi, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu (2021). "Attacking adversarial attacks as a defense". In: *arXiv preprint arXiv:2106.04938*.

Xie, Cihang, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille (2017). "Mitigating adversarial effects through randomization". In: *arXiv preprint arXiv:1711.01991*.

Yemelyanov, Dmitry (2020). *Chinese Traffic Signs*. https://www.kaggle.com/datasets/dmitryyemelyanov/chinese-traffic-signs. Accessed: 2025-11-21.

Zago, João G, Eric A Antonelo, Fabio L Baldissera, and Rodrigo T Saad (2020). "It is double pleasure to deceive the deceiver: disturbing classifiers against adversarial attacks". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 160–165.

Zagoruyko, Sergey and Nikos Komodakis (2016). "Wide residual networks". In: *arXiv preprint arXiv:1605.07146*.

Zhang, Hengyu, Guang Cheng, Yuyang Zhou, and Zongyao Chen (2025). "MalShield: Enhancing Android Malware Detection with Stateful Defense against Query Attacks". In: *2025 IEEE 24th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, pp. 681–689.

Zhang, Hongyang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan (2019). "Theoretically principled trade-off between robustness and accuracy". In: *International Conference on Machine Learning*. PMLR, pp. 7472–7482.

Zhang, Jie, Pura Peetathawatchai, Florian Tramèr, and Avital Shafran (2026). "Laundering AI Authority with Adversarial Examples". In: *arXiv preprint arXiv:2605.04261*.

Zhang, Yimeng, Yuguang Yao, Jinghan Jia, Jinfeng Yi, Mingyi Hong, Shiyu Chang, and Sijia Liu (2022). "How to robustify black-box ml models? a zeroth-order optimization perspective". In: *arXiv preprint arXiv:2203.14195*.

Zheng, Meixi, Xuanchen Yan, Zihao Zhu, Hongrui Chen, and Baoyuan Wu (2023). "BlackboxBench: A Comprehensive Benchmark of Black-box Adversarial Attacks". In: *arXiv preprint arXiv:2312.16979*.

Zhu, Ziqi, Bin Zhu, Huan Zhang, Yu Geng, Le Wang, Denghui Zhang, and Zhaoquan Gu (2022). "Defense against query-based black-box attack with small Gaussian-noise". In: *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*. IEEE, pp. 249–256.

Zimmer, Pascal, Sébastien Andreina, Giorgia Azzurra Marson, and Ghassan Karame (2024). "Closing the Gap: Achieving Better Accuracy-Robustness Tradeoffs Against Query-Based Attacks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 19, pp. 21859–21868.