Expert Systems With Applications 264 (2025) 125840

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Random transformations to improve mitigation of query-based black-box attacks

Ziad Tariq Muhammad Ali [a,*], R. Muhammad Atif Azad [a], Muhammad Ajmal Azad [a], James Holyhead [b], Iain Rice [a], Ali Shariq Imran [c]

[a] School Of Computing And Digital Technology, Birmingham City University, Birmingham, United Kingdom
[b] Covatic, Birmingham, United Kingdom
[c] Department of Computer Science, Norwegian University of Science and Technology, Norway

ARTICLE INFO

Dataset link: https://github.com/ziyadt1995/RT_to_mitigate_QBBAs/

Keywords:
Black-box attacks
Adversarial examples
Randomised defences
Neural networks

ABSTRACT

This paper proposes methods to upstage the best-known defences against query-based black-box attacks. These benchmark defences incorporate gaussian noise into input data during inference to achieve state-of-the-art performance in protecting image classification models against the most advanced query-based black-box attacks. Even so there is a need to improve upon them; for example, the widely benchmarked Random noise defense (RND) method has demonstrated limited robustness – achieving only 53.5% and 18.1% with a ResNet-50 model on the CIFAR-10 and ImageNet datasets, respectively – against the square attack, which is commonly regarded as the state-of-the-art black-box attack. Therefore, in this work, we propose two alternatives to gaussian noise addition at inference time: random crop-resize and random rotation of the input images. Although these transformations are generally used for data augmentation while training to improve model invariance and generalisation, their protective potential against query-based black-box attacks at inference time is unexplored. Therefore, for the first time, we report that for such well-trained models either of the two transformations can also blunt powerful query-based black-box attacks when used at inference time on three popular datasets. The results show that the proposed randomised transformations outperform RND in terms of robust accuracy against a strong adversary that uses a high budget of 100,000 queries based on expectation over transformation (EOT) of 10, by 0.9% on the CIFAR-10 dataset, 9.4% on the ImageNet dataset and 1.6% on the Tiny ImageNet dataset. Crucially, in two even tougher attack settings, that is, high-confidence adversarial examples and EOT-50 adversary, these transformations are even more effective as the margin of improvement over the benchmarks increases further.

1. Introduction

Adversarial examples are carefully crafted inputs that fool machine learning (ML) models into misclassifying the data (Biggio et al., 2013; Szegedy et al., 2014). The methods for fooling or attacking ML classifiers vary depending on the capabilities of the adversary and can be categorised as either white-box or black-box. With a white-box method an adversary has complete access to the parameters, architecture and the training process of the model, while with a black-box method, the adversary has no such knowledge (Yuan et al., 2019). Depending on the level of opacity in a black-box method, the adversarial insight may be limited to only accessing model outputs, such as predicted probabilities or class labels. In more transparent scenarios, adversaries might even access the underlying training dataset (Carlini et al., 2019).

Although white-box attacks are the most harmful they are not always practical because they require access to the model parameters either directly or through parameter extraction attacks (Carlini et al., 2020); black-box methods are thus typically the first line of attack for most practical settings. Therefore, based on the significance of black-box attacks, this paper is concerned with improving the state-of-the-art defences against the black-box methods.

Various approaches have been in practice to blunt the adversarial attacks. They were partially countered by the development of techniques such as *adversarial training* (AT) (Goodfellow et al., 2015), which was later improved upon by Aleks et al. (2018); it remains the method of choice for achieving inherent adversarial robustness (Athalye et al., 2018). Other defence methods proposed are based on input preprocessing (Aprilpyone & Kiya, 2021; Guo et al., 2018; Kurakin et al., 2016)

* Corresponding author.
E-mail addresses: ziad.ali@mail.bcu.ac.uk (Z.T.M. Ali), atif.azad@bcu.ac.uk (R.M.A. Azad), muhammadajmal.azad@bcu.ac.uk (M.A. Azad), james@covatic.com (J. Holyhead), iain.rice@bcu.ac.uk (I. Rice), ali.imran@ntnu.no (A.S. Imran).

https://doi.org/10.1016/j.eswa.2024.125840
Received 2 February 2024; Received in revised form 9 November 2024; Accepted 16 November 2024
Available online 23 November 2024

and detection of adversarial examples (Chakraborty et al., 2021; Xu et al., 2017). Most input preprocessing techniques were later shown to be ineffective by the development of adaptive attacks by Ali et al. (2022), Athalye et al. (2018) and Tramer et al. (2020). This is because these defences often rely on gradient obfuscation to hinder attack methods; however, adaptive attacks can circumvent this obfuscation, thereby undermining the defences' effectiveness.

A more successful class of input preprocessing defences against adversarial attacks alters the input data stochastically by adding random noise; some variations of these defences add random noise during training (Cohen et al., 2019), whereas other works only add it during inference (Byun et al., 2022; Gu et al., 2023; Guo et al., 2018; Jia et al., 2023; Nguyen et al., 2023; Qin et al., 2021; Xie et al., 2017). While the efficacy of these methods as a defence against white-box attacks has been challenged (Gao et al., 2022; Gnecco-Heredia et al., 2023; Lucas et al., 2023), against black-box attacks stochastic methods such as the *Random Noise defense (RND)* (Qin et al., 2021) and *Adversarial Attack on Attackers (AAA)* (Chen et al., 2022) have been shown to be effective. Later, AAA technique, which introduces noise at the output logits, was shown to be vulnerable to decision-based attacks (Nguyen et al., 2023). However, the RND method has enjoyed some popularity and become the state-of-the-art benchmark to improve upon in this line of work (Zheng et al., 2023).

Despite achieving top-standing, previous work shows that RND needs further 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.

This paper aims to improve upon the best-known defence against the best query-based black-box attack under intense attack-settings. Therefore:

- we compare our proposed alternative methods against RND;
- we test all the newly proposed as well as the benchmark defences against the state-of-art in query-based black-box attacks i.e. the square attack (Andriushchenko et al., 2020; Qin et al., 2021);
- and to intensify the attack and stress-test the defences, we modify square attack by averaging model-predictions over multiple forward passes (an approach termed *Expectation over Transformation* or EOT (Athalye et al., 2018)) to make the classification decision; this averaging provides greater information to the adversary than a single prediction and allows it a greater chance to beat the defences (Nguyen et al., 2023; Qin et al., 2021).

The proposals to defend against black-box attacks that we test in this work are two randomised transformations: random resized cropping and random rotation. 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, we show for the first time that they can also effectively defend the models against query-based black-box attacks.

We find that randomised transformations are more effective than randomised noise defence at mitigating query-based black-box attacks.

On ImageNet dataset (Russakovsky et al., 2015) we show that the robust accuracy of a ResNet-50 model (He et al., 2016) with inference-time random cropping-resizing and random rotation is 9.4% and 7.5% better respectively than that of RND (Qin et al., 2021) under the EOT-10 square attack.

On Tiny ImageNet (Deng et al., 2009) with a PreActResNet18 model randomised crop-resize improves robust accuracy over RND by 1.6%, while random rotation improves the robust accuracy by 0.8% under EOT-10 square attack.

On CIFAR-10 dataset (Krizhevsky & Hinton, 2009) we find that on ResNet-50 random rotation beats RND by 0.9%, while random cropping-resizing achieves the same performance as RND under the EOT-10 square attack.

Encouraged by these results, we also explored 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%.

Addtionally, to ensure a robust evaluation of defences, previous studies (Nguyen et al., 2023; Qin et al., 2021) test the defences under EOT-10 adversary. We take it a step further and rigorously test defences against a strong adversary of EOT-50. 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 cropping-resizing beats RND by 3.5% on ResNet-18 architecture.

Furthermore, we have also compared our proposed transformations against the recently proposed variation of RND, Randomised feature defense (RFD) which adds gaussian noise in between the layers of the model (Nguyen et al., 2023). Our results show an improvement of randomised transformations over RFD.

Consequently, the results reported in this paper 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 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 paper.

To summarise, the main contributions of the paper are as follows.

1. We find that randomised transformations are more effective than randomised noise at mitigating query-based black-box attacks. The impact of these transformations as a defence against query-based black-box attacks has not been previously explored.
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 RND is significantly vulnerable to EOT-50 adversary while randomised transformations are not.
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.

The rest of the paper is outlined as follows. In Section 2, work related to black-box attacks and defences against black-box attacks is presented, Section 3 introduces preliminary information for understanding the aim of defences against black-box attacks, our proposed methodology is presented in Section 4 and the experimental results are discussed in Section 5. Lastly, the study's limitations are reviewed, followed by the concluding remarks in Section 6.

2. Related work

In this section, we introduce black-box attacks, defences that were used to counter black-box attacks, and the latest attacks that aim to bypass a few of the defences.

2.1. Black-box attacks and their weakness

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 and Hein (2020) have proposed Auto-attack (AA), which is an ensemble of attacks that has become almost a standard to reliably test a model's adversarial robustness (Croce & Hein, 2020). The square attack is a part of AA; this highlights the significance and esteem that it enjoys in the field.

However, recent work has shown that black-box attacks are vulnerable to small perturbations to input data, also called Random Noise defense (RND). Therefore, Byun et al. (2022) and Qin et al. (2021) have shown that black-box attacks are vulnerable to adding random noise to input images at inference time. The RND method when combined with adversarial training has been claimed to give superior robustness against black-box attacks as compared to models that are not adversarially trained. The work in Chen et al. (2022) perturbs the output logits to counter black-box attacks while that in Nguyen et al. (2023) perturbs the features in the model to diminish the black-box attack capabilities. Therefore, Dong et al. (2020) and Lucas et al. (2023) have called for the development of powerful black-box attacks that can remain effective under random noise. Even so, while RND remains the defence to beat, some other defences have been breached, which we summarise next.

2.2. Adaptive black-box attacks

Although the RND poses a challenge to black-box attacks, adaptive black-box attacks have been developed to counter some other defences that attempt to mitigate black-box attacks. For example, they have been able to overcome (Feng et al., 2023) the so-called Stateful defence models (SDM) (Chen et al., 2020; Choi et al., 2023; Juuti et al., 2019; Li et al., 2022) that work by rejecting queries that are similar based on comparing past queries using a defence specific metric. These attacks counter the defence by using rejection sampling to pick queries that will not cause a model to detect collision.

Along a separate line of work (Sitawarin et al., 2022) have shown that using model-specific (non-randomised) resizing, cropping, or quantisation as part of data pre-processing can weaken the attack performance of decision-based black-box attacks but only until these modifications are unknown to the attacker. Therefore, to counter these preprocessors, they develop attacks that first reverse-engineer the preprocessors and then develop adversarial examples. However, their work does not consider the widely used l_∞ distance metric to limit the maximum adversarial perturbation and uses only the l_2 distance metric. Furthermore, their work also does not consider defences that use randomisation during the inference stage.

Since RND method is currently considered the best defence against black-box attacks, in the following section we introduce the defence mechanism of random noise defences against black-box attacks. Moreover, we also describe the mechanism of randomised transformations.

3. Preliminaries

3.1. Threat model

Consider a machine learning classifier C that classifies inputs $\mathbf{x} \in \mathbb{R}^d$ into labels $y_{pred} \in \mathbb{Z}_k$ as $C : \mathbb{R}^d \rightarrow \{1, \ldots, k\}$ where d is the dimension of the input and k is the number of classes. $y \in \mathbb{Z}_k$ is the true label or class of a given input. The goal of the classifier is to classify the inputs as $C(\mathbf{x}) = y$ for all input instances. In contrast, the objective of the adversary is to alter the input instance as $\mathbf{x}_{adv} = \mathbf{x} + \delta$ so that $C(\mathbf{x}_{adv}) \neq y$ (Carlini et al., 2019). The adversarial input $\mathbf{x}_{adv}$ is perturbed in such a way that it is close to the original instance x according to an l_p distance metric of the form $\|\mathbf{x} - \mathbf{x}_{adv}\|_p \leq \epsilon$ where p can be any metric l_2, l_0 or l_∞, but is still misclassified by the classifier. The distance metric ensures that no input feature or combination of features is changed more than the allowed budget, and this is given by ϵ.

Our focus in this work 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.

3.2. Random noise defense

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:

$$\min_{\mathbf{x}_{adv} \in \mathcal{A}_R(\mathbf{x})} f(\mathbf{x}_{adv}) = \min_{\mathbf{x}_{adv} \in \mathcal{A}_R(\mathbf{x})} (\mathcal{M}_y(\mathbf{x}_{adv}) - \max_{j \neq y} \mathcal{M}_j(\mathbf{x}_{adv})) \tag{1}$$

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:

$$h(\mathbf{x}) = f(\mathbf{x} + \mu\mathbf{u}) - f(\mathbf{x}) < 0 \tag{2}$$

where μ is the perturbation size.

Random noise defense (RND) (Qin et al., 2021) and small noise defense (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:

$$g(\mathbf{x}) = f(\mathbf{x} + \mu\mathbf{u} + \nu\boldsymbol{\nu_1}) - f(\mathbf{x} + \nu\boldsymbol{\nu_2}) \tag{3}$$

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., 2023) 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.

3.3. Random transformation defense

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 emphasise that these transformations when used during inference time 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:

$$m(\mathbf{x}) = f(t_{r_1}(\mathbf{x} + \mu\mathbf{u})) - f(t_{r_2}(\mathbf{x})) \tag{4}$$

where t_r is a random transformation, which could include random cropping-resizing, random rotation, random cropping-padding, etc. The transformation is applied during the inference stage. Random transformations misleads 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.

4. Proposed methodology

4.1. Randomised transformations

Randomised transformations during inference time were proposed by Guo et al. (2018) and Xie et al. (2017) for countering adversarial attacks. Although they were later shown to be ineffective in preventing white-box attacks (Athalye et al., 2018) their effectiveness for countering query-based black-box attacks was not explored. In this work, we propose random cropping-resizing and random rotation 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 (Aleks 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.

Next, we detail the two proposed transformations.

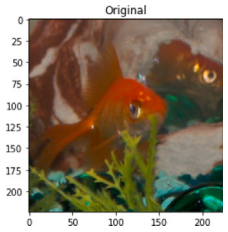

Fig. 1. Original image of a Goldfish from the ImageNet dataset.

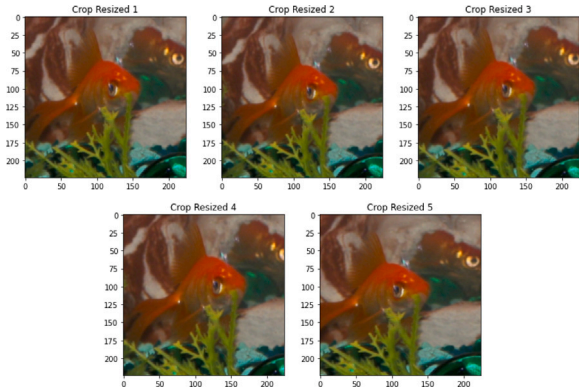

Fig. 2. Random cropping-resizing variants of an image of a Goldfish from the ImageNet dataset. Our preliminary experiments revealed superior performance of 80% cropping value over smaller cropping values.

4.1.1. Random cropping-resizing

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 Fig. 1 while its transformations are given in Fig. 2.

4.1.2. 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 Fig. 3.

4.2. Evaluation

4.2.1. Attacker capabilities

To ensure that the defences are tested reliably, we set the attacker's budget to 100,000 queries for EOT-10 attack and 500,000 queries for EOT-50 attack. 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

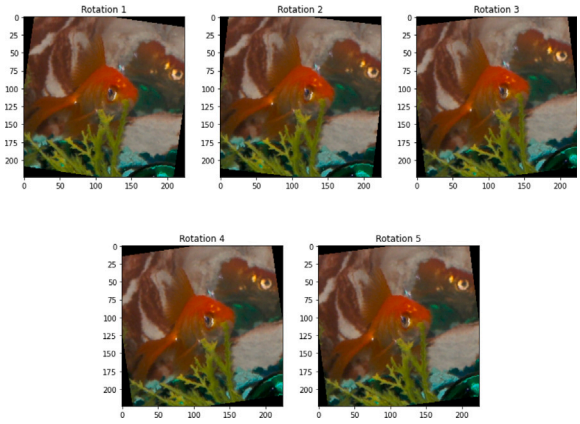

Fig. 3. Random rotation variants of an image of a Goldfish from the ImageNet dataset.

set EOT to 10 or 50 to evaluate the defence performance under the square attack (Andriushchenko et al., 2020). The square attack is a powerful black-box attack, evident from the fact that it is the best-performing attack against RND as reported by Qin et al. (2021). The square attack is run for 10,000 iterations with default settings.

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, ensuring a reliable evaluation as that is close to how the model will be used in practical settings. 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

1: $\mathbf{x}^{adv} \leftarrow init(\mathbf{x})$
2: $l^* \leftarrow L(f(\mathbf{x}), y)$
3: **for all** iterations in N **do**
4: $h^{(i)} \leftarrow$ side length of the square to modify
5: $\boldsymbol{\delta} \leftarrow \mathrm{P}(\epsilon, h^{(i)}, w, c, \mathbf{x}^{adv}, \mathbf{x})$
6: $\mathbf{x}^{adv}_{new} \leftarrow$ Project $\mathbf{x}^{adv} + \boldsymbol{\delta}$ onto $\{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z} - \mathbf{x}\|_p \leq \epsilon\} \cap [0, 1]^d$
7: $l_{new} \leftarrow \frac{1}{M} \sum_{j=0}^{M} L(f(\mathbf{x}^{adv}_{new}), y)$
8: **if** $l_{new} < l^*$ **then**
9: $\mathbf{x}^{adv} \leftarrow \mathbf{x}^{adv}_{new}$
10: $l^* \leftarrow l_{new}$
11: **end if**
12: **end for**
13: return $\mathbf{x}^{adv}$

4.2.2. 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.

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 (Andriushchenko et al., 2020).

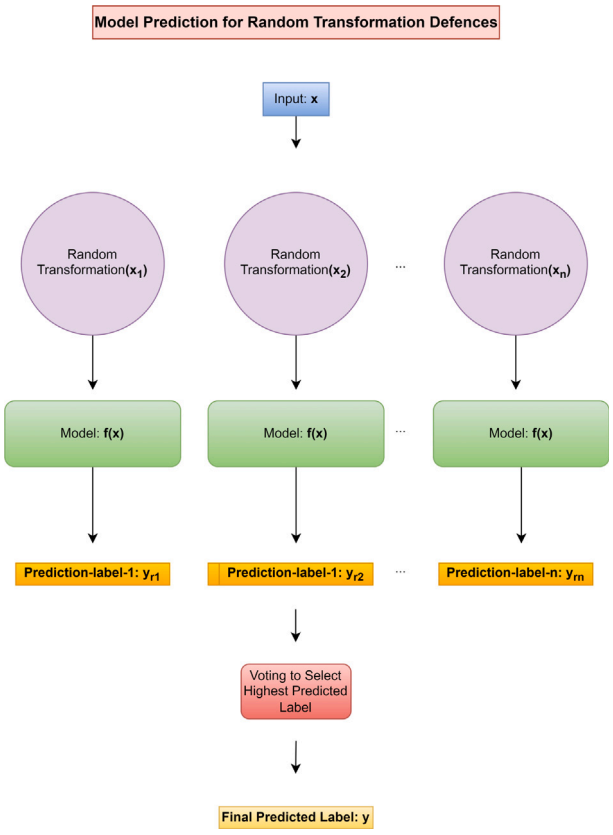

Fig. 4. Illustration of how the model prediction is used for random transformation defences in this work.

5. Experimental results

5.1. Experimental settings

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). Therefore, we randomly sample 1000 images from the CIFAR-10 test set and 1000 images from the validation set provided by Huang and Zhang (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. The ResNet-50 is used for EOT-10 attack experiments while Wide-ResNet-28-10 and ResNet-18 are used for EOT-50 attack experiments.

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.

Table 1
Performance of randomised defences on CIFAR-10, ImageNet and Tiny ImageNet datasets under square attack (Andriushchenko et al., 2020). 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	Defense 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+Crop-Resize (Ours)	88.1	***535***	***53.5***
				AT+Rotation (Ours)	***88.7***	**544**	**54.4**
ImageNet	1000	ResNet-50	10	AT	**61.6**	98	9.8
				AT+RND (Qin et al., 2021)	56.7	181	18.1
				AT+Crop-Resize (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+Crop-Resize (Ours)	**40.0**	**82**	**16.4**
				AT+Rotation (Ours)	37.4	***78***	***15.6***

The l_∞ distance of the attack is set to 12.75/255 for all models and defences used in this work. We use the square attack with default settings as it is a powerful attack giving the best attack performance (Qin et al., 2021) and part of the auto-attack ensemble (Croce & Hein, 2020).

Furthermore, the experiments are conducted on two machines, one with Intel core i9 CPU, Nvidia GeForce RTX 3080 Ti GPU and 32 GB of RAM, and the other with Intel core i7 CPU, Nvidia Quadro T1000 GPU and 16 GB of RAM.

5.1.1. Data augmentations

Typically, to improve model invariance and generalisation, data augmentations based on various transformations such as those proposed in this paper 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 (Engstrom et al., 2019) uses random resized cropping, random horizontal flip and colour jitter for the ImageNet model, while it uses random cropping, random horizontal flip, colour jitter and random rotation for the CIFAR-10 model. Furthermore, we use random resized cropping and random rotation to train the PreActResNet18 model on TinyImageNet and ResNet-18 model on CIFAR-10.

5.2. 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 defense (RFD) (Nguyen et al., 2023) as discussed in Section 3.2 and random cropping-padding (Xie et al., 2017), as discussed in Section 4.1. AAA (Chen et al., 2022) was not evaluated as it is vulnerable to decision-based attacks as shown by Nguyen et al. (2023).

Therefore, we compare random cropping-resizing, and random rotation with RND (Qin et al., 2021) 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., 2023) we use the first 200 samples of the 1000 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). For this reason, we evaluate the defences using square attack.

5.2.1. 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., 2023), 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. The illustration of model prediction is given in Fig. 4.

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.

5.3. 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, 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. To measure the clean accuracy, we take a majority vote over 50 forward passes as per Fig. 4 for every randomised defence technique except in the case of ablation study where we use a single forward pass.

5.4. Results

5.4.1. CIFAR-10

We report the results on CIFAR-10 under the regular square attack with EOT-10 and EOT-50, along with the confidence-increasing square attack. The EOT-10 results are reported in Table 1, while the EOT-50 results are reported in Table 4. The results under confidence increasing attack are reported in Table 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 cropping-resizing beats RND by 3.5% whereas with ResNet-18 architecture, random rotation beats

Table 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	Defense technique	Attack failures	Robust accuracy (%)
CIFAR-10	1000	ResNet-50	10	AT+RND (Qin et al., 2021)	543	54.3
				AT+Crop-Resize (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+Crop-Resize (Ours)	**73**	**36.5**
				AT+Rotation (Ours)	***71***	***35.5***

Table 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	Defense 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., 2023)	**75.0**	49	24.5
				AT+Crop-Resize (Ours)	***72.5***	**63**	**31.5**
				AT+Rotation (Ours)	72.0	***58***	***29.0***

Table 4
Performance of randomised defences on randomly sampled 1000 samples from the CIFAR-10 dataset 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	Defense technique	Clean accuracy (%)	Attack failures	Robust accuracy (%)
CIFAR-10	1000	ResNet-18	50	AT	**82.7**	361	36.1
				AT+RND (Qin et al., 2021)	80.9	461	46.1
				AT+Crop-Resize (Ours)	81.6	***484***	***48.4***
				AT+Rotation (Ours)	**82.2**	**498**	**49.8**
		Wide-ResNet-28-10 (Zagoruyko & Komodakis, 2016)	50	AT	**89.9**	483	48.3
				AT+RND (Qin et al., 2021)	88.4	586	58.6
				AT+Crop-Resize (Ours)	88.5	***621***	***62.1***
				AT+Rotation (Ours)	***88.8***	**632**	**63.2**

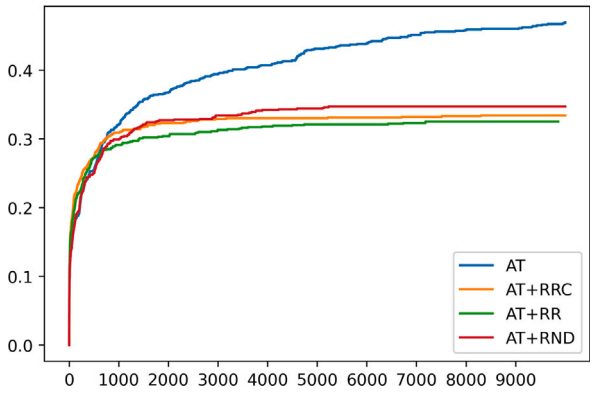

Fig. 5. The attack success rate graph of the randomised defences on square attack with EOT-50 on ResNet-18 model architecture and CIFAR-10 dataset. The attack success rate is measured across the iterations of the attack.

RND by 3.7% and random cropping-resizing beats RND by 2.3%. The attack success rate graph for the defences are given in Fig. 5.

Moreover, the results in Table 2 show that under a confidence-increasing attack, where the full attack budget is utilised and strong 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.

5.4.2. 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 1000 samples. Our proposed randomised cropping-resizing improves significantly over RND by giving an improvement of 9.4%. Randomised rotation also improves over RND by 7.5%. The results are reported in Table 1.

Additionally, we compare the performance of RND and our proposed randomised transformations over the confident square attack in Table 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%.

5.4.3. 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 1. According to the results, RND gives a robust accuracy of 14.8% while random crop-resize improves on it and takes it to 16.4%. On the other hand, random rotation also improves the robust accuracy by 0.6%. The results confirm the improvement of randomised transformations over RND.

5.4.4. 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 5.2. 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. We use the RND (Qin et al., 2021)

Table 5
Ablation study results to evaluate the performance of randomised defences with a single forward pass 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	Defense technique	Clean accuracy (%)	Attack failures	Robust accuracy (%)
CIFAR-10	1000	ResNet-18	50	AT	**82.7**	361	36.1
				AT+RND (Qin et al., 2021)	***81.5***	***374***	***37.4***
				AT+Crop-Resize (Ours)	80.2	264	26.4
				AT+Rotation (Ours)	79.4	**390**	**39.0**
ImageNet	200	ResNet-50	10	AT	**75.0**	21	10.5
				AT+RND (Qin et al., 2021)	69.0	34	17.0
				AT+Crop-Resize (Ours)	***71.5***	***48***	***24.0***
				AT+Rotation (Ours)	69.5	**52**	**26.0**

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 random noise defence such as RND and RFD by 3.5%. Moreover, randomised cropping-resizing outperforms all other defences by giving the highest robust accuracy of 31.5%.

5.4.5. 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 3.3.

The results for the first ablation study are given in Table 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 cropping-resizing 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 resized-cropping 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 natural (undefended) model is 36.1%, which is significantly higher than the accuracies at fixed angles of 5 and 10 degrees. Similarly, the robust accuracies of resized-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.

5.4.6. 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. 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 Fig. 6. We observe a clear increase of robust accuracy improvement of transformation based defences over RND as attacker noise is increased. We attribute the poor performance of RND to the higher degradation caused by pixel level noise added by RND. We also hypothesise that on datasets more sensitive to noise, the performance of RND might suffer considerably in comparison to transformation based defences.

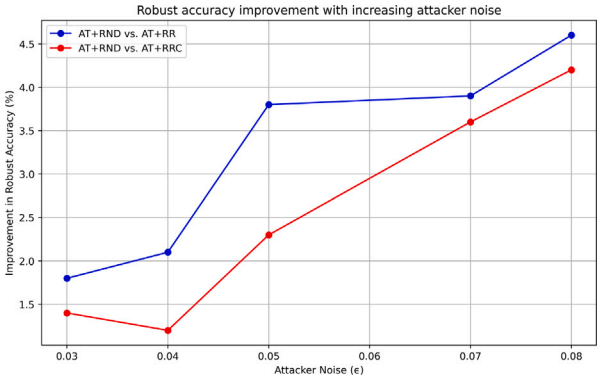

Fig. 6. The improvement in robust accuracy of transformation based defences over RND as attacker noise is increased.

5.4.7. 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 & Ipbuker, 2021). On CIFAR-10, the p-value for improvement of random rotation over RND is 1.03×10^{-7} while for random resized-cropping 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 resized-cropping 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.

6. Limitations, future work and conclusion

6.1. Limitations and future work

Nguyen et al. (2023) and Qin et al. (2021) have pointed out that randomised defences work to mitigate query-based black-box attacks,

and their performance against mitigating transferability attacks remains an open problem. Additionally, all randomised defences may also be vulnerable to an attacker with higher number of EOTs (>500). Although we have tried to evaluate performance on EOT-50, their robustness might drop further against EOT > 500 at a cost of a higher number of queries. The effect of higher EOTs and its potential to be applied in a realistic setting along with transferability attack evaluation is left for future investigation.

6.2. Conclusion

In this work, we have proposed random transformations such as random resized cropping and random rotation as a more effective defence to mitigate black-box attacks. Experimental results demonstrate the improvement over previously proposed SOTA methods such as RND. We observe that data augmentations during training help with the defencive abilities of the proposed transformations. Furthermore, we also observe that randomised transformations such as random rotation and random resized-cropping perform better than noise-based defences. While RND adds pixel-level noise on top of the attacker's noise, which might significantly degrade image quality resulting in lower robustness, the proposed transformations crop out the image and then extrapolate, which is a fundamentally different operation.

It is also noteworthy that against confidence-based attacks, the improvement over the RND increases even further across all the tests. This further reinforces the potential of the proposed randomised transformations. Furthermore, we find that it is not only the transformations that improve the adversarial robustness, but their combinations with randomisation that plays a critical role.

Despite these improvements, it is important to note that these transformations can also be used together with RND and/or in concatenation with each other. However, this investigation is left for future.

CRediT authorship contribution statement

Ziad Tariq Muhammad Ali: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Writing – original draft. **R. Muhammad Atif Azad:** Conceptualization, Validation, Investigation, Writing – original draft, Resources, Supervision. **Muhammad Ajmal Azad:** Validation, Investigation, Writing – original draft, Supervision. **James Holyhead:** Validation, Writing – review & editing, Supervision. **Iain Rice:** Validation, Investigation, Writing – original draft, Supervision. **Ali Shariq Imran:** Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and code are available at https://github.com/ziyadt1995/RT_to_mitigate_QBBAs/.

References

Ali, Z. T. M., Mohammed, A., & Ahmad, I. (2022). Evaluating adversarial robustness of secret key-based defenses. *IEEE Access*, *10*, 34872–34882.

Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020). Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision* (pp. 484–501). Springer.

Aprilpyone, M., & Kiya, H. (2021). Block-wise image transformation with secret key for adversarially robust defense. *IEEE Transactions on Information Forensics and Security*, *16*, 2709–2723.

Athalye, A., Carlini, N., & Wagner, D. A. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In J. G. Dy, & A. Krause (Eds.), *Proceedings of machine learning research: vol. 80, Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (pp. 274–283). PMLR.

Atik, S. O., & Ipbuker, C. (2021). Integrating convolutional neural network and multiresolution segmentation for land cover and land use mapping using satellite imagery. *Applied Sciences*, *11*(12), 5551.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., & Roli, F. (2013). Evasion attacks against machine learning at test time. In *Machine learning and knowledge discovery in databases* (pp. 387–402). Berlin, Heidelberg: Springer Berlin Heidelberg.

Byun, J., Go, H., & Kim, C. (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, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., & Kurakin, A. (2019). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.

Carlini, N., Jagielski, M., & Mironov, I. (2020). Cryptanalytic extraction of neural network models. In *Annual international cryptology conference* (pp. 189–218). Springer.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, *6*(1), 25–45.

Chen, S., Carlini, N., & Wagner, D. (2020). Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM workshop on security and privacy on artificial intelligence* (pp. 30–39).

Chen, S., Huang, Z., Tao, Q., Wu, Y., Xie, C., & Huang, X. (2022). Adversarial attack on attackers: Post-process to mitigate black-box score-based query attacks. *Advances in Neural Information Processing Systems*, *35*, 14929–14943.

Choi, S.-H., Shin, J., & Choi, Y.-H. (2023). PIHA: Detection method using perceptual image hashing against query-based adversarial attacks. *Future Generation Computer Systems*, *145*, 563–577.

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *International conference on machine learning* (pp. 1310–1320). PMLR.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Dong, Y., Fu, Q.-A., Yang, X., Pang, T., Su, H., Xiao, Z., & Zhu, J. (2020). Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 321–331).

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., & Tsipras, D. (2019). Robustness (Python library).

Feng, R., Hooda, A., Mangaokar, N., Fawaz, K., Jha, S., & Prakash, A. (2023). Investigating stateful defenses against black-box adversarial examples. arXiv preprint arXiv:2303.06280.

Gao, Y., Shumailov, I., Fawaz, K., & Papernot, N. (2022). On the limitations of stochastic pre-processing defenses. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems*.

Gnecco-Heredia, L., Chevaleyre, Y., Negrevergne, B., Meunier, L., & Pydi, M. S. (2023). On the role of randomization in adversarially robust classification.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (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*.

Gowal, S., Qin, C., Uesato, J., Mann, T., & Kohli, P. (2020). Uncovering the limits of adversarial training against norm-bounded adversarial examples. arXiv preprint arXiv:2010.03593.

Gu, J., Wei, F., Torr, P., & Hu, H. (2023). Exploring non-additive randomness on ViT against query-based black-box attacks. arXiv preprint arXiv:2309.06438.

Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. In *International conference on learning representations*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, Z., & Zhang, T. (2019). Black-box adversarial attack with transferable model-based embedding. arXiv preprint arXiv:1911.07140.

Ilyas, A., Engstrom, L., Athalye, A., & Lin, J. (2018). Black-box adversarial attacks with limited queries and information. In *International conference on machine learning* (pp. 2137–2146). PMLR.

Jia, W., Lu, Z., Yu, R., Li, L., Zhang, H., & Liu, Z. (2023). AND: Adaptive noise to defend query-based black-box attacks without affecting normal inference. Available at SSRN 4596212.

Juuti, M., Szyller, S., Marchal, S., & Asokan, N. (2019). PRADA: protecting against DNN model stealing attacks. In *2019 IEEE European symposium on security and privacy* (pp. 512–527). IEEE.

Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images*. Citeseer.

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.

Li, H., Shan, S., Wenger, E., Zhang, J., Zheng, H., & Zhao, B. Y. (2022). Blacklight: Scalable defense for neural networks against {Query-Based}{Black-Box} attacks. In *31st USeNIX security symposium* (pp. 2117–2134).

Lucas, K., Jagielski, M., Tramèr, F., Bauer, L., & Carlini, N. (2023). Randomness in ML defenses helps persistent attackers and hinders evaluators. arXiv preprint arXiv:2302.13464.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (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*.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.*, *12*(2), 153–157.

Nguyen, Q. H., Lao, Y., Pham, T., Wong, K.-S., & Doan, K. D. (2023). Understanding the robustness of randomized feature defense against query-based adver-sarial attacks. arXiv preprint arXiv:2310.00567.

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I., & Edwards, B. (2018). Adversarial robustness toolbox v1.2.0. CoRR 1807.01069.

Qin, Z., Fan, Y., Zha, H., & Wu, B. (2021). Random noise defense against query-based black-box attacks. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems*. URL https://openreview.net/forum?id=ZPSD4xZc6j8.

Qin, Z., Yan, X., & Wu, B. (2022). BlackboxBench (Python library).

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. (2021). Fixing data augmentation to improve adversarial robustness. arXiv preprint arXiv:2103.01946.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

Sitawarin, C., Tramèr, F., & Carlini, N. (2022). Preprocessors matter! Realistic decision-based attacks on machine learning systems. arXiv preprint arXiv:2210.03297.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2014). Intriguing properties of neural networks. In Y. Bengio, & Y. LeCun (Eds.), *2nd international conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, conference track proceedings*.

Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On adaptive attacks to adversarial example defenses. arXiv preprint arXiv:2002.08347.

Wang, L., Yang, K., Wang, W., Wang, R., & Ye, A. (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).

Wu, B., Pan, H., Shen, L., Gu, J., Zhao, S., Li, Z., Cai, D., He, X., & Liu, W. (2021). Attacking adversarial attacks as a defense. arXiv preprint arXiv:2106.04938.

Xie, C., Wang, J., Zhang, Z., Ren, Z., & Yuille, A. (2017). Mitigating adversarial effects through randomization. arXiv preprint arXiv:1711.01991.

Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155.

Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2805–2824, IEEE.

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. arXiv preprint arXiv:1605.07146.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning* (pp. 7472–7482). PMLR.

Zheng, M., Yan, X., Zhu, Z., Chen, H., & Wu, B. (2023). BlackboxBench: A comprehensive benchmark of black-box adversarial attacks. arXiv preprint arXiv:2312.16979.