



When zero query meets adversarial attacks: Feature maps as an offensive and defensive approach ^{☆,☆☆}

Joana Cabral Costa ^{ID}*, Tiago Roxo ^{ID}, Hugo Proença ^{ID}, Pedro R.M. Inácio ^{ID}

Instituto de Telecomunicações, Universidade da Beira Interior, R. Marquês de Ávila e Bolama, 6201-001, Covilhã, Portugal

ARTICLE INFO

Dataset link: ZQBA: Zero Query Black-box Adversarial Attack (Original data)

Keywords:

Adversarial attack
Adversarial training
Black-box
Feature maps
Robustness
Transferability

ABSTRACT

Current black-box adversarial attacks either require multiple queries or diffusion models to produce adversarial samples that can impair the target model performance. However, these methods have limited applicability in real-world settings, where the attacker does not have unlimited time to interact with the target model. The Zero-Query Black-box Adversarial (ZQBA) attack introduces a feature-map-based strategy that allows adversarial examples to be pre-computed during an offline setup phase and later deployed in time-constrained environments, benefiting from high transferability across architectures and domains. Building upon ZQBA, this paper proposes using feature-aware perturbations from ZQBA complemented with Adversarial Training (AT) to improve the robustness across diverse attack scenarios. Our results show that incorporating ZQBA perturbations into AT improves robustness on white-box, black-box and transfer-based settings, suggesting that these perturbations provide complementary regularization effects that reduce sensitivity to small input variations. Overall, our findings highlight the potential of feature-aware perturbations as an efficient and query-free component for strengthening model robustness. All the source code is available at <https://github.com/Joana-Cabral/ZQBA>.

1. Introduction

Deep Neural Networks (DNNs) have achieved remarkable performance in numerous fields, ranging from Natural Language Processing [1,2] and Medical Imaging [3,4], to Active Speaker Detection [5–7], strongly contributing to widespread integration and acceptance of Artificial Intelligence in everyday life. Inevitably, this also led to an increased appeal for attackers to target vulnerabilities in DNNs, and, as demonstrated by Szegedy et al. [8], DNNs may fail to generalize reliably and can be manipulated by imperceptible perturbations [9,10], leaving them susceptible to adversarial attacks.

Adversarial attacks are typically categorized into *white-box* approaches, where the attacker can access model gradients, and *black-box* approaches, where no gradient information is available, but the attacker can perform repeated querying of the target model. In the field of image analysis, recent works regarding *black-box* scenarios use only input images and predictions to estimate the gradients of a target model [11] or train generative models [12,13] to create strong perturbations. These either depend on thousands of queries

to the target model to recreate its gradients or on training generative models specifically designed to produce adversarial samples that strongly depend on the attack used. Zero Query Black-box Adversarial (ZQBA) [14] is an attack that generates adversarial perturbations using feature maps extracted from different DNNs, without the previously mentioned limitations. This attack has demonstrated transferability across different networks and domains, motivating a deeper analysis into feature-aware perturbations as a potential defense. Consequently, we employ feature-map-based Adversarial Training (AT) to enhance the generalization capacity of the models.

Building upon the Zero Query Black-box Adversarial (ZQBA) attack [14], this work investigates the value of integrating ZQBA-generated feature-aware perturbations into AT as a means to improve model robustness across diverse attacks thereby posing as an extension for ZQBA. An overview of the proposed methodology is depicted in Fig. 1, illustrating the integration of feature map-derived perturbations into AT and flexibility of inclusion with existing methods. Furthermore, the results suggest that including these perturbations during training phase

[☆] This article is part of a Special issue entitled: 'Security-AI' published in Image and Vision Computing.

^{☆☆} This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., and, when eligible, co-funded by EU funds under project/support UID/50008/2025 – Instituto de Telecomunicações, with DOI identifier <https://doi.org/10.54499/UID/50008/2025>; and by the FCT Doctoral Grant 2020.09847.BD and Grant 2021.04905.BD.

* Corresponding author.

E-mail addresses: joana.cabral.costa@ubi.pt (J.C. Costa), tiago.roxo@ubi.pt (T. Roxo), hugomcp@ubi.pt (H. Proença), prmi@ubi.pt (P.R.M. Inácio).

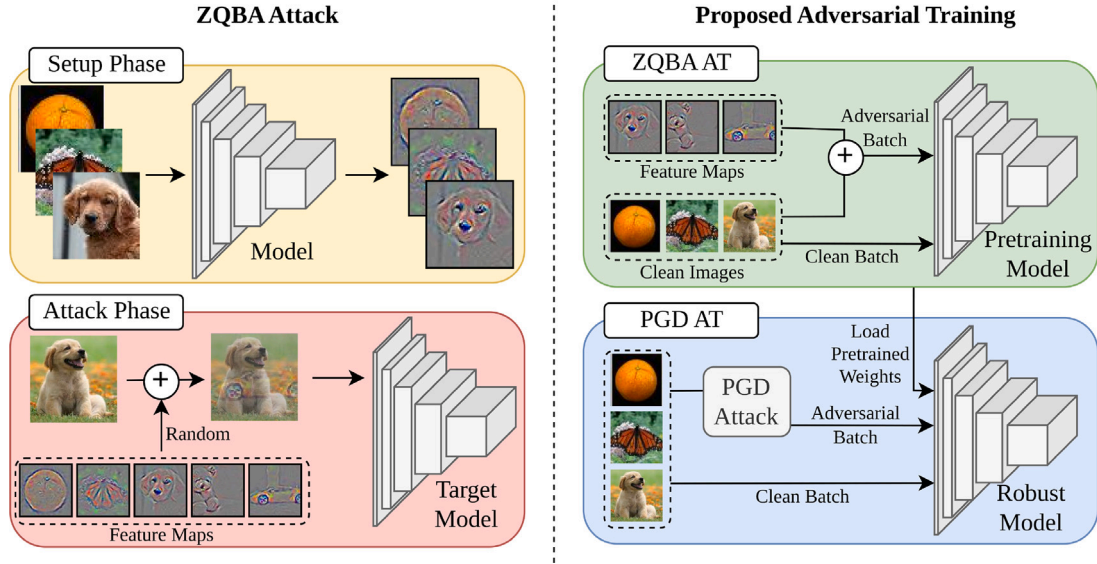


Fig. 1. Overview of Zero-Query Black-box Attack (ZQBA) and the proposed Adversarial Training (AT) approach. ZQBA has two phases: (1) *Setup*, where all the feature maps (perturbations) are obtained; and (2) *Attack*, where previously obtained perturbations are added to the image to be attacked, without querying the target model. In ZQBA AT, the feature maps obtained during the *Setup* phase are randomly superimposed onto clean training samples during the pretraining stage. The resulting model weights are subsequently transferred to the final robust model, which is trained under a standard AT procedure using PGD as the adversarial attack.

helps models to better generalize across different attack scenarios and supports the development of new approaches considering this type of perturbations. Overall, the key contributions presented in this paper can be summarized as follows:

- We extend the use of ZQBA, an attack that leverages the feature map representations to create perturbations, toward improving model adversarial robustness in white-box, black-box, and gray-box scenarios;
- We propose the inclusion of feature-aware perturbations during the training phase, which we observed that act as loss regularizer and, as a result, using these weights as a pretrain to improve model generalization;
- We present results for quantitative analysis in white-box, black-box, and gray-box (transfer attacks) showing consistent improvements in robustness for the proposed AT methodology in relation to standard AT, in our settings. Visualization techniques corroborate these results, suggesting better clustering capacity and smaller loss values.

The remaining of the paper is structured as follows: Section 2 discusses the related works in white-box and black-box scenarios and AT; Section 3 describes the attack scenario, the ZQBA strategy, and the theoretical foundations and methodology for ZQBA AT; Section 4 reports the experimental setup, ablation studies, and performance analysis, accompanied by a discussion; finally, Section 5 concludes the paper.

2. Related work

White-box Attacks rely on accessing model gradients to create adversarial samples. Several methods have been proposed in this for more than a decade. Fast Gradient Sign Method (FGSM) [15] is a one-step method that finds adversarial samples using the model loss function. Jacobian-based Saliency Maps (JSM) [16] directly maps the input to the output and only alters small fractions of the image. DeepFool [17] uses iterative linearization to create minimal perturbations that cross the decision boundary. Projected Gradient Descent (PGD) [18] uses saddle point formulation to find a strong perturbation

through multiple iterations. SmoothFool [19] uses smooth adversarial perturbations to enhance transferability and stealthiness. Auto-Attack (AA) [20] combines four attacks (the majority are white-box) to evaluate the robustness of a defense. The requirement of white-box attacks to access model gradients for creating adversarial samples limits their practicality in real-world scenarios, which deviates from the intended scope of ZQBA.

Black-box Attacks either require a significant amount (e.g., in the order of thousands) of queries to the target model or the training of diffusion models to produce adversarial examples. Zeroth Order Optimization (ZOO) [11] directly estimates the gradients of the target model without training, achieving better transferability. Square Attack [21] perturbs the images in localized square-shaped regions in random positions. Ray Searching (RayS) [22] is a hard-label attack that uses discrete problem optimization to fasten searches while being effective. Park et al. [23] uses the gradients of a surrogate model to achieve higher query efficiency. DiffAttack [13] uses both textual and visual data to train a diffusion model to generate human-insensitive perturbations. TAIGen [24] uses both diffusion models and Class Activation Maps (CAM) to create adversarial examples. ScoreAdv [25] leverages interpretable saliency maps to produce realistic adversarial images using diffusion models. In the case of ZQBA, the adversarial samples are generated without interaction with the target model (0 queries) and do not rely on generative models.

Adversarial Training is a machine learning technique used to make models more robust by training them on intentionally perturbed or examples, namely adversarial samples from white-box attacks, since these adapt to the model gradients during training and help increasing its robustness. Adversarial Weight Perturbation (AWP) [26] uses a double-perturbation mechanism that disturbs both inputs and weights and can be applied to existing AT methods. Semantic Constraint Adversarial Robust Learning (SCARL) [27] incorporates semantic information into the model training process and maximizing mutual information between the visual and textual representations. Core Feature-aware Adversarial Training (COFAT) [28] selectively preserves benign sample regions on feature maps with high-weight and suppresses those with low weights, ensuring the model focuses on core features. AT-PR [29] identifies the region with the highest proportion of Adversarial Examples in the loss landscape, rather than the Adversarial Example that maximizes the loss.

While most AT approaches modify the loss function or the training procedure, this paper shows that loading a previously trained model with ZQBA samples helps improve robustness across different attack scenarios.

3. Preliminaries and methodology

3.1. Attack scenario

This paper considers a threat model in which the adversary aims to degrade the classification performance of a target Deep Learning model by introducing imperceptible perturbations to input images. As such, a successful attack is characterized by a measurable decrease in classification accuracy, coupled with a high SSIM score [30], ensuring that the adversarial perturbations remain imperceptible to human observers. Depending on the amount of information, there are two types of attacks: (1) *white-box*, which require access to the gradients of the target model; and (2) *black-box*, which demand multiple queries to the target model to simulate its behavior. While *white-box* access may be difficult to obtain in realistic scenarios, *black-box* attacks can be costly or infeasible within short time frames due to requiring thousands of queries, suggesting that more realistic attack scenarios may be necessary [31]. With that purpose in mind, we explored the creation of adversarial attacks without querying the target model by adding the representation of a model different from the target one, while balancing the imperceptibility and effectiveness of the perturbation. The attack assumes the training of dedicated models to retrieve the feature maps, which introduce minimal computational costs. The threat models considered in our experiments are described as follows: (1) in the *white-box* setting, the attacker is assumed to have complete knowledge of and access to the target model, including its gradients, architecture, and training data; (2) in the *gray-box* setting, the attacker knows architecture of the model and training dataset but does not have access to gradient information; and (3) the *black-box* setting assumes that the attacker can only interact with the model through its outputs, namely logits or labels, and has no knowledge of the model architecture or the data used during training.

3.2. Image perturbation for attacks

Obtaining Feature Maps. Our approach involves training a model and obtaining its representations for different classes by extracting feature maps from layers that come before the classification layers (i.e., fully connected layers) using backpropagation. To obtain these feature maps, we use Guided Backpropagation [32], which is a gradient-based approach that retrieves the gradient of images when backpropagating through the Rectified Linear Unit (ReLU) activation function, where only the flow of positive gradients is allowed by changing the negative gradient values to zero. Fig. 2 summarizes how the forward and backward passes for Guided Backpropagation are used to obtain the feature maps. This approach is typically used to display the features of the input image that maximize the activation of the feature maps, resembling the features that more closely influence the model prediction. Other possible approaches to obtain the visual representation of models could include GradCAM and its variants. However, Guided Backpropagation demonstrated greater potential for applicability since it clearly represented the object, rather than attention areas of the model.

Creating Adversarial Images. The idea behind ZQBA is to maintain the original image representation while adding perturbations that do not significantly alter this image, while impairing the classification of a model. Therefore, after obtaining the feature maps for every image in a dataset, ZQBA conjugates a randomly selected feature map with a clean image, according to the following equation:

$$\mathbf{X}_{adv} = \alpha \cdot (\nabla_{\mathbf{X}} \cdot f(\mathbf{X})) + \mathbf{X}, \quad (1)$$

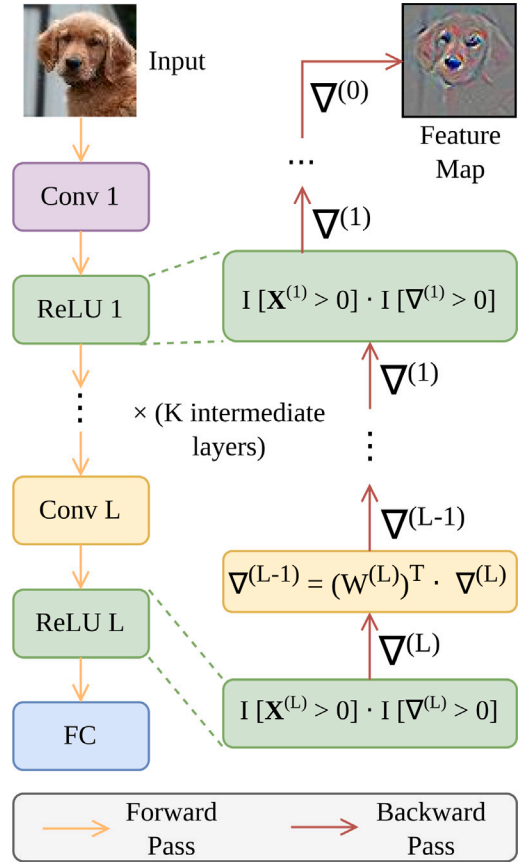


Fig. 2. Overview of Guided Backpropagation feature map extraction, starting from the layer prior to the classification layers (i.e., FC), as defined in ZQBA. The ReLU layers only let positive activations and positive gradients pass. FC refers to Fully Connected, Conv refers to Convolutional, and ∇ refers to the backpropagated gradients.

where $\mathbf{X}_{adv}$ is the perturbed image, α is the perturbation weight, $\nabla_{\mathbf{X}}$ is the gradient with respect to the input image pixels, $f(\mathbf{X})$ is the logits corresponding to the model predicted class, and $\mathbf{X}$ is the original image. ZQBA considers $\alpha = 0.4$ since it achieved the best results in performance and image quality, as shown in the Ablation Studies (Section 4.3).

Feature Map Selection. The selection of the feature maps could either be based on the similarity between the clean image and the image used to obtain the feature map, or by choosing a random feature map, which ultimately increases the viability of the attack and its application range, as shown in the Ablation Studies (Section 4.3). Usually, when perturbations are applied to an image, there is a maximum threshold to limit the amount of manipulation in the clean image ($\epsilon = 8/255$ for CIFAR-10 and CIFAR-100 and $\epsilon = 4/255$ for Tiny ImageNet). In the ZQBA context, this threshold does not directly apply since it is used to limit the amount of noise generated based on the gradients of the model for *white-box* or based on the number of queries for *black-box*. To ensure that the attacked images created by ZQBA remain imperceptible, we assess the quality of the generated images using Structural Similarity (SSIM) [30], which evaluates the images according to the perception of the human eye, and ensure that SSIM remains above or equal to 0.95 (SSIM ≥ 0.95), at which it is reported that the human observer cannot detect manipulations to the image [33].

3.3. Adversarial training setup

Motivation and Theoretical Formulation. When considering feature learning theory, standard training leads models to learn non-robust

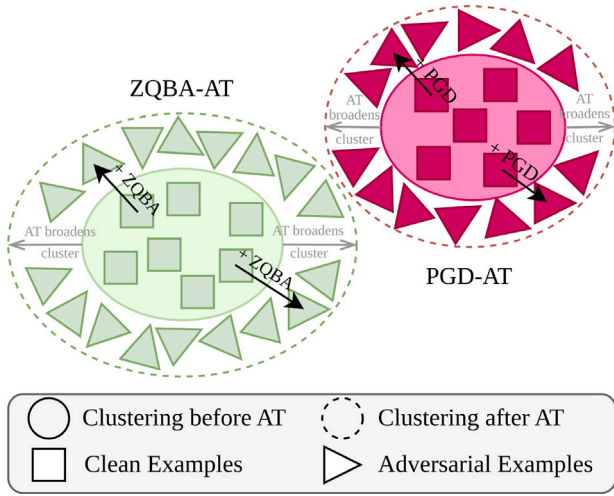


Fig. 3. Illustration of the clusters before and after applying AT (i.e., when trained only on clean examples and when adversarial samples are added).

features that are susceptible to small perturbations. As such, a perturbation depicting only the flow of positive gradients, isolating the features that most strongly activate the model, is sufficient to fool multiple networks, as we show with ZQBA. AT prevents this effect by adding adversarial examples during training, ensuring a more robust feature learning, and it is innate to think that ZQBA could also contribute to increase AT performance. Fig. 3 illustrates the concept of performing AT. A model that was standardly trained creates clusters only with clean samples (smaller circle), then the perturbations force these samples out of this cluster, and when we include the adversarial examples in the training phase, it creates the broader cluster (bigger circle). This makes it much harder for other adversarial attacks to create examples outside the bigger circle.

The standard AT is highly oriented toward exploring the worst-case scenarios, but this information is highly dependent on the attack used. Namely, the standard AT is trying to minimize the loss function of the model, when presented with worst-case adversarial examples, defined as:

$$\min_{\theta} \mathbb{E}_{(\mathbf{X}, Y)} \left[\max_{\|\delta\| \leq \epsilon} \mathcal{L}(f_{\theta}(\mathbf{X} + \delta), Y) \right], \quad (2)$$

where $\mathbb{E}$ is the CE function, $\mathbf{X}$ is the input image, Y is the label corresponding to $\mathbf{X}$, δ is the perturbation, $\mathcal{L}$ is the loss function, ϵ is the maximum perturbation value, and f_{θ} are the model logits. Solving this optimization problem often leads to overfitting to the attack, usually PGD, and lesser domain generalization. As such, we use a mechanism that considers the directions aligned with the gradients of the model, but only uses the features that positively contribute to the model prediction (i.e., the most relevant features). With this approach we are creating samples outside the hyperplane for a specific class, which are different from the ones generated by standard AT, and we are trying to solve the following optimization problem:

$$\min_{\theta} \mathbb{E}_{(\mathbf{X}, Y)} [\mathcal{L}(f_{\theta}(\mathbf{X} + \delta_Z), Y)], \quad (3)$$

where δ_Z refers to the perturbation obtained using ZQBA. The backward pass for Guided Backpropagation is defined as:

$$\frac{\partial f}{\partial x_i^{(l)}} = \frac{\partial f}{\partial x_i^{(l+1)}} \cdot \mathbb{1}[x_i^{(l)} > 0] \cdot \mathbb{1}\left[\frac{\partial f}{\partial x_i^{(l+1)}} > 0\right], \quad (4)$$

where ∂ is the partial derivative, f is the output of the model, x is the input to layer l , and $\mathbb{1}$ is the indicator function. The proposed strategy considers Eq. (3) as a tractable surrogate of Eq. (2). Furthermore, because the perturbation is derived from a class-discriminative, sign-masked gradient (Eq. (4)), rather than the mixed-sign loss gradient

Table 1

Classification accuracy of various architectures on clean examples across multiple datasets. Par (M) denotes total parameters in millions.

Model	Par (M)	Clean accuracy		
		CIFAR-10	CIFAR-100	TinyImageNet
MobileNetv2	3.3	85.08	57.27	45.79
EfficientNetB2	8.7	84.99	58.70	58.90
ResNet18	11.1	94.43	69.64	55.85
ResNet50	24.4	96.65	78.33	63.83
WideRN28-10	36.4	89.52	56.43	48.11
ResNet101	42.5	96.25	78.64	63.99
VGG19	137.0	91.63	63.17	55.38

used by PGD, it is explicitly steered toward the model decision hyper-plane rather than some local direction that rapidly increases loss. This makes minimizing Eq. (3) conceptually closer to margin maximization, and consequently less sensitive to the specific attack algorithm than standard AT, which is consistent with the results presented in the experiments.

Standard AT methods also require the generation of adversarial examples in each epoch, constantly adapting to the gradients of the model, which creates a significant computational overhead. ZQBA perturbations are generated from a surrogate model in an independent phase from training, requiring no additional queries to the model being trained. This makes ZQBA AT more practical in settings with limited resources and with less accuracy degradation since the SSIM constraint (≥ 0.95) ensures that the perturbations are well-balanced.

Methodology. Our AT strategy extends ZQBA [14] by taking advantage of its perturbations as data augmentation to improve the robustness of models. This process starts with the creation of feature maps (Setup Phase) from different architectures that are trained on the same dataset as the one we intend to achieve robustness, creating a group of adversarial perturbations. We apply Guided Backpropagation to each training and testing image, for each model, to extract the corresponding feature map, and the adversarial samples are created using Eq. (1). As such, we have a pool of adversarial examples D_{adv} , created before the training phase, without requiring interaction with the model to be trained.

During the ZQBA AT phase, the trained model f_{θ} is trained on a combination of clean and adversarial samples, using the standard Cross-Entropy (CE) loss on both of them, defined as:

$$\min_{\theta} \mathbb{E}_{(\mathbf{X}, Y)} [\mathcal{L}(f_{\theta}(\mathbf{X}), Y) + \mathcal{L}(f_{\theta}(\mathbf{X} + \delta_Z), Y)]. \quad (5)$$

In this case, the feature map used to perturb each image is randomly selected, with the constraint that it should be from a class distinct than the one assigned to the original image $\mathbf{X}$. Contrary to the standard AT, the inner maximization is replaced with a pre-computed perturbation obtained during the Setup Phase. With this step, we have a model trained to be resilient to feature-aware manipulations, which seemingly generalizes better across the different attack types considered in this work. The weights of this model can be loaded as a pretrain to any AT approach, associated with improved generalization capacity, making the proposed approach flexible.

4. Experiments

This section elaborates on the experiments that were conducted to assess the effectiveness of the proposed approach. The experimental setup is outlined first, followed by the performance evaluation, and finally the presentation of the ablation study results. For reference, Table 1 summarizes the clean accuracy and number of trainable parameters for all evaluated datasets and architectures. The performance of these networks serves as a comparison for all experiments.

Table 2

Performance of ZQBA in cross-architecture settings across multiple datasets. MNv2, E2, RN18, RN50, RN101, and WRN denote MobileNetV2, EfficientNetB2, ResNet18, ResNet50, ResNet101, and WideResNet28-10, respectively. F and T denote the source architecture used for feature map extraction and the target model, respectively. Δ represents the difference between clean accuracy and the strongest attack accuracy (highlighted in bold).

Dataset	F \ T	MNv2	E2	RN18	RN50	WRN	RN101	VGG19	Δ
CIFAR-10	MNv2	50.39	<u>54.22</u>	67.27	69.18	65.35	69.91	62.43	-34.69
	E2	53.69	<u>54.39</u>	67.28	68.98	65.50	69.41	62.88	-31.30
	RN18	73.07	74.73	78.63	81.17	<u>74.18</u>	81.22	75.24	-21.36
	RN50	76.22	<u>77.19</u>	81.33	82.00	<u>80.26</u>	83.50	78.61	-20.43
	WRN	65.57	<u>67.05</u>	71.83	74.98	68.71	74.79	69.52	-23.95
	RN101	75.33	<u>77.28</u>	79.05	79.73	78.66	82.53	<u>76.04</u>	-20.92
	VGG19	66.51	<u>67.42</u>	74.00	75.55	71.71	76.43	67.50	-25.12
CIFAR-100	MNv2	16.36	<u>22.02</u>	30.07	34.72	33.41	33.45	23.17	-40.91
	E2	25.82	<u>25.87</u>	32.98	35.34	35.43	35.07	28.53	-32.88
	RN18	30.18	34.20	42.16	45.29	46.86	42.31	<u>32.16</u>	-39.46
	RN50	38.98	41.27	46.84	51.20	54.15	53.02	<u>40.91</u>	-39.35
	WRN	36.31	36.24	34.88	36.47	29.80	37.38	<u>33.98</u>	-26.63
	RN101	<u>39.27</u>	42.42	47.13	53.19	54.95	51.51	38.23	-40.41
	VGG19	<u>30.88</u>	31.08	37.44	41.13	38.46	39.87	27.19	-35.98
Tiny ImageNet	MNv2	16.23	<u>22.46</u>	25.39	26.56	26.21	25.09	25.51	-29.56
	E2	37.92	35.09	37.13	38.69	<u>37.12</u>	37.36	39.57	-23.81
	RN18	36.38	36.03	29.43	34.35	34.14	<u>32.63</u>	35.90	-26.42
	RN50	44.49	43.31	38.59	<u>37.72</u>	41.60	37.09	43.53	-26.74
	WRN	30.58	30.87	<u>29.52</u>	31.92	27.86	32.32	31.90	-20.25
	RN101	42.73	41.29	<u>38.06</u>	39.46	41.62	33.06	42.41	-30.93
	VGG19	34.09	34.83	33.52	36.01	35.41	35.54	<u>33.54</u>	-21.86

4.1. Experimental setup

Datasets. Experiments were performed on three datasets used in adversarial robustness research: CIFAR-10 and CIFAR-100 [34], both containing 50,000 training and 10,000 test images with 10 and 100 classes, respectively; and Tiny ImageNet [35], a 200-class subset of ImageNet [36] with 100,000 and 10,000 training and validation images, respectively.

Models. Each architecture was trained in the CIFAR-10, CIFAR-100, and Tiny ImageNet datasets, and models with the best *clean accuracy* were selected. These models are used to obtain the feature maps for each dataset, and the models susceptible to perturbations (target models) differ from those used to obtain these feature maps. Similar to the literature, the first Convolutional layer of ResNet18, 50, and 101 was modified to be more suitable for the image size of CIFAR. The remaining networks follow the implementation provided by PyTorch [37]. The models were trained for 100 epochs for CIFAR-10 and 200 epochs for CIFAR-100 and Tiny ImageNet, with a batch size of 128, and random crop and random horizontal flip as data augmentation strategies.

Adversarial Attacks. To thoroughly evaluate the proposed AT methodology, we consider *white-box*, *black-box*, and transfer attacks (*gray-box*). We selected PGD-CE, PGD with Difference of Logits Ratio (DLR), FGSM, AA standard and AA+ as *white-box* attacks, Square and RayS as *black-box* attacks, and adversarial samples generated for surrogate models, using FGSM, PGD-CE, and AA standard, as *gray-box* attacks.

Evaluation. We evaluate model robustness using accuracy on natural test samples, referred to as *Clean Accuracy*, and on adversarial samples generated under *white-box*, *black-box*, and *gray-box* threat models, each identified by the name of the corresponding attack. For results with mean and standard deviation, we have trained the models 5 times with random seeds. Additionally, the perceptual quality of the ZQBA adversarial images is assessed using the PSNR, L_{inf} , L_2 , and SSIM metric [30], which correlates with human visual perception. Using ResNet18, surrogate model training took 1 h and 20 min, followed by 17 min and 42 s for feature map extraction (average of three datasets). The generated feature maps occupied 30 MB for CIFAR-10 and CIFAR-100 and 118 MB for Tiny ImageNet. When averaged across the evaluated datasets, the attack consumed 406 MiB of memory and

required 0.018 s per image during execution, which represents 83% less than Square Attack and 98% less than PGD with 40 steps.

Implementation Details. For PGD training, we use the PGD-CE with 10 steps (as stated below), CE loss, and Stochastic Gradient Descent as optimizer with a momentum of 0.9 for all datasets. A learning rate of 0.01 and weight decay of 0.0005 for CIFAR-10, a learning rate of 0.05 and weight decay of 0.0005 for CIFAR-100, and a learning rate of 0.1 and weight decay of 0.0001 for Tiny ImageNet. We use the StepLR scheduler with a step size of 30 and γ of 0.1 for CIFAR-10 and CosineAnnealingLR for CIFAR-100 and Tiny ImageNet. Regarding the *white-box* attacks, for all datasets, we consider an $\epsilon = 8/255$. The PGD-CE with random starts has $\alpha = 2/255$ and 10 and 40 steps versions; PGD-DLR has 100 steps and 5 restarts; AA standard and plus versions use L_{∞} norm. Square attack uses only one restart and a varying number of queries, explicitly stated in the results. The hardware consists of 1 NVIDIA A40, with 46 GB of memory, Intel Xeon @ 2.10 GHz, with 12 cores, and 128 GB of RAM. All the code was developed using PyTorch and the standard test sets associated with each dataset were used. The attacks use the Torchattacks library implementation, except for RayS, which uses the official GitHub implementation provided by the authors.

4.2. ZQBA performance evaluation

Architecture Influence on Attack Performance. The first experiment focuses on evaluating the performance of ZQBA in different architectures and datasets to determine whether the feature maps generalize across various architectures or are dependent on prior knowledge. Table 2 displays the accuracy under ZQBA attack for cross-architecture in CIFAR-10, CIFAR-100, and Tiny ImageNet. The obtained results show that: (1) for CIFAR-10, the smallest networks generate the best perturbations for all architectures; (2) for CIFAR-100, the smallest network achieves the best overall results, except for the larger networks, where the same architecture generates better perturbations; and (3) for Tiny ImageNet, overall the same network type is the best at generating the perturbations. In sum, the feature maps used to generate adversarial images are not dependent on the architecture of the target model (i.e., the performance of the models is always below the clean accuracy presented in Table 1, as highlighted by the Δ), meaning that the attacker does not need previous details of the target model architecture, although this knowledge may increase the strength of the attack.

Table 3

Transferability of ZQBA across multiple datasets. The first and third rows of the table header denote the dataset used to extract the feature maps and the dataset used for evaluation, respectively. Values in parentheses indicate the relative decrease with respect to clean accuracy.

Model	CIFAR-10 ↓ Tiny	CIFAR-100 ↓ Tiny	Tiny ↓ CIFAR-10	Tiny ↓ CIFAR-100
MNv2	27.60 (40%)	27.91 (39%)	78.00 (8%)	42.54 (26%)
E2	40.34 (32%)	40.55 (31%)	77.12 (9%)	44.90 (24%)
RN18	38.78 (43%)	38.99 (43%)	89.39 (5%)	55.10 (21%)
RN50	47.74 (39%)	49.21 (37%)	93.04 (4%)	66.79 (15%)
WRN	31.61 (34%)	29.30 (39%)	81.80 (9%)	36.96 (35%)
RN101	49.51 (35%)	47.41 (38%)	92.43 (4%)	66.83 (15%)
VGG19	38.37 (31%)	37.46 (32%)	86.74 (5%)	54.18 (14%)

Table 4

Classification accuracy of state-of-the-art black-box attacks vs. ZQBA using a ResNet architecture across multiple datasets.

Attack	#Queries	CIFAR-10	CIFAR-100	Tiny ImageNet
Square	1	88.47	51.55	44.03
ZOO	1	94.43	69.63	67.78
ZQBA	0	78.63	42.16	29.43

Dataset Transferability. We evaluate the transferability of ZQBA to datasets distinct from the one used to generate the feature maps, in Table 3, highlighting that models used to obtain these maps do not need to be trained on the same data as the target models. Since CIFAR-10 and CIFAR-100 share the same data, we do not consider this combination in this experiment. The transferability strength is evaluated using the model accuracy and its relative decrease to the clean accuracy, represented by the values between parentheses, as a percentage. The results show that: (1) the transferability from CIFAR-10/100 to Tiny ImageNet has a significant impact on the model accuracy, justified by the perturbations having a smaller size with a more scattered representation of the objects in the image; and (2) the perturbations obtained from Tiny ImageNet have less impact due to its size being adjusted to match the image size of the target dataset, resulting in a loss of detail in the used perturbation; and (3) the difference between Tiny→CIFAR10 and Tiny→CIFAR100, which relates to the inherent complexity of manipulating the classifications for a dataset with fewer classes (10 vs. 100), *i.e.*, perturbations are more likely to impair the classification of a model with 100 classes than one with 10 classes, given that the latter has more precise decision boundaries for classification. These results indicate that transferability of ZQBA is improved when perturbations are obtained from smaller images and applied to larger images, and that datasets with more classes are more susceptible to changes in final classification.

State-of-the-art Comparison. Current *black-box* attacks use multiple queries to obtain perturbations for the target models, while our approach does not require any previous interaction with these models. Therefore, for a fair comparison with ZQBA, we consider state-of-the-art *black-box* adversarial attacks with a single query. Table 4 presents the results for a ResNet-18, as the reference model, on various datasets. Regarding the ZOO attack, the accuracy remains the same as the clean accuracy since the attack cannot, in one query, reliably find a perturbation. Furthermore, Square attack shows a slight decrease in model performance, but its efficiency relies on a greater number of queries due to its approach of refining the noise added to the image. ZQBA has a greater performance degradation relative to the clean accuracy, showing the impact of this attack without queries for all the considered datasets.

4.3. Ablation studies

Selecting Noise Perturbation. The motivation behind ZQBA is to use information obtained from DNNs to create effective perturbations

Table 5

Performance ZQBA using a ResNet model when susceptible to multiple variations of noise, in different datasets. SSIM is used to assess image quality.

Variations	CIFAR-10		CIFAR-100		Tiny ImageNet	
	Acc.	SSIM (↑)	Acc.	SSIM (↑)	Acc.	SSIM (↑)
Feature Map	78.63	0.95	42.16	0.95	29.43	0.95
EigenGradCAM	90.40	0.83	58.50	0.88	43.38	0.89
GradCAM	90.58	0.83	58.61	0.88	43.39	0.89
FullGrad	91.77	0.86	60.89	0.89	46.43	0.94

Table 6

Performance of ZQBA using a ResNet model when selecting the feature maps based on different image similarity metrics, in different datasets. SSIM is used to assess image quality.

Approach	Metric	CIFAR-10		CIFAR-100		Tiny ImageNet	
		Acc.	SSIM(↑)	Acc.	SSIM(↑)	Acc.	SSIM(↑)
Least	RMSE	78.46	0.95	41.97	0.95	31.28	0.95
	SSIM	83.80	0.95	48.32	0.95	30.33	0.95
	Pearson	78.68	0.95	42.30	0.95	31.24	0.95
Most	RMSE	83.09	0.95	46.93	0.95	33.17	0.95
	SSIM	78.66	0.95	41.80	0.95	30.49	0.95
	Pearson	79.19	0.95	42.65	0.95	30.69	0.95
Random	–	78.63	0.95	42.16	0.95	29.43	0.95

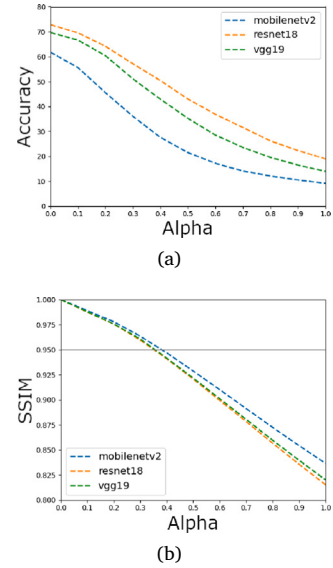


Fig. 4. Accuracy (%) and SSIM (%) of ZQBA with different weights for the feature maps in multiple architectures. The lines refer to an average of each model performance across the three considered datasets. The horizontal line represents the threshold of human perceptibility in detecting that an image has been attacked.

that deceive model classification. Thus, we explore various methods for extracting the representative information of the original images, using Feature Maps, EigenGradCAM [38], GradCAM [39], and FullGrad [40] in Table 5. The results show that extracting the Feature Maps, using Guided Backpropagation, leads to greater performance degradation and achieves better image quality, as depicted by the SSIM values, throughout the datasets. Furthermore, the FullGrad approach has less effective results (given by greater accuracy), despite achieving better image quality than other GradCAM approaches due to more concise attention maps. In sum, the Feature Maps provide the best relation between image quality and performance decrease given its representation of the entire object in the image, as opposed to the GradCAM approaches, which emphasize specific parts of the image (*i.e.*, model attention).

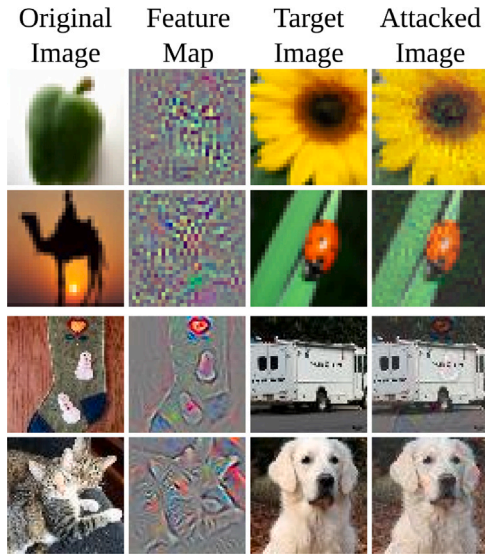


Fig. 5. Original Image, Feature Map, Target Image, and Attacked Image for CIFAR, in the first two rows, and Tiny ImageNet, in the last two. The Original Image refers to the image used to obtain the Feature Map, the Target Image is the one whose classification the attacker wants to compromise, and the Attacked Image is the combination of the Feature Map and the Target Image.

Table 7

Comparison of perturbation magnitude across datasets using PSNR, L_{inf} , and L_2 . Higher PSNR values and lower norm magnitudes indicate that the perturbations remain visually imperceptible.

Datasets	PSNR(↑)	L_{inf} (↓)	L_2 (↓)
CIFAR-10	24.96	0.23	4.67
CIFAR-100	34.82	0.14	2.83
Tiny ImageNet	39.11	0.11	3.79

Table 8

Performance of ZQBA in cross-architecture settings in different datasets, using Adversarial Training. We considered the feature maps that most diminish the performance of the model. Δ displays the improvement of including Adversarial Training and Tiny refers to the Tiny ImageNet dataset.

Dataset	Models	Strongest		
		Clean	ZQBA	Δ
CIFAR-10	MobileNetv2	73.62	64.81	+14.42
	EfficientNetB2	77.49	70.08	+16.39
	ResNet18	86.54	81.52	+8.44
	ResNet50	93.20	77.23	+1.01
	WideResNet28-10	99.17	84.35	+18.78
	ResNet101	93.47	75.49	+0.16
	VGG19	95.93	79.46	+12.95
CIFAR-100	MobileNetv2	35.49	25.23	+8.87
	EfficientNetB2	40.93	32.08	+6.26
	ResNet18	70.28	64.04	+33.86
	ResNet50	69.80	63.62	+24.64
	WideResNet28-10	76.90	71.53	+41.73
	ResNet101	71.16	62.05	+23.82
	VGG19	41.66	33.18	+5.99
Tiny	MobileNetv2	39.99	34.39	+18.16
	EfficientNetB2	53.07	48.12	+13.03
	ResNet18	40.84	34.27	+4.84
	ResNet50	45.64	38.28	+1.19
	WideResNet28-10	52.16	45.09	+17.23
	ResNet101	42.99	36.47	+3.41
	VGG19	42.45	35.77	+2.25

Image Similarity Effect. We further assess whether the attack, when using feature maps, has a greater impact depending on the

similarity between the images used to obtain the feature maps and the target image, as shown in Table 6. We do not consider feature maps from the same class as the target image, since it would lead the model to correctly classify the attacked image either way. The results show that selecting the feature map based on similarity does not affect the quality of the attacked image, since all approaches achieved adequate image quality (≥ 0.95). Regarding the accuracy, randomly selecting feature maps is a reliable strategy for all the datasets and has the inherent advantage of not requiring a preliminary evaluation of the similarity between images, with increasing overhead with dataset size. Therefore, the considered strategy was the Random selection of feature maps to perturb the image, which further diminishes the knowledge required by the attacker, widening the applicability of ZQBA. To enhance the evaluation of image quality, we extend our evaluation beyond SSIM and incorporate widely adopted perturbation-magnitude metrics, such as L_{inf} , L_2 , and Peak Signal-to-Noise Ratio (PSNR) in Table 7. The results show that the attack produces low-magnitude perturbations and high imperceptibility, as evidenced by the L-values and PSNR values, respectively, while also highlighting that the method generalizes well to more complex datasets, in our setting.

Effect of Feature Map Weight. When adding the perturbations to the target images, ZQBA considers an α value to weight the impact of the perturbation relative to the original image that decreases models performance, without compromising image quality, i.e. SSIM ≥ 0.95 [33]. We evaluate different α values, in Fig. 4, relative to accuracy and SSIM, considering different architectures, with the values averaged through the three datasets. We choose MobileNetv2, ResNet18, and VGG19 to depict diverse model families and sizes. The results show that increasing the value of α reduces model accuracy, regardless of the architecture, as shown in Fig. 4(a). However, solely considering accuracy to assess the best value for α is not sufficient since ZQBA needs to produce perturbations imperceptible to the human eye. As such, Fig. 4(b) displays the relation between α and SSIM, with the threshold for image quality (0.95), and the results show a decrease in image quality with the increase of α , emphasizing that the best value above the threshold is 0.4, throughout the architectures and datasets.

Qualitative Analysis of ZQBA. We provide representative samples for the considered datasets in Fig. 5 to ensure that the proposed attack does not result in significant visual disruption. The examples show that the attacked image does not undergo significant changes relative to its original state, and it is still possible for humans to correctly classify the objects existing in these images. Furthermore, the feature maps obtained for Tiny ImageNet are more concise and clearly represent an object, while CIFAR tends to scatter more around the region of the object, mainly due to the size of the original images. This aspect is also consistent with the results for dataset transferability, in Table 3. These images demonstrate the stealthiness and the applicability of using feature maps to produce adversarial attacks, without needing to query the target model, favoring ZQBA for imperceptible adversarial attacks.

4.4. ZQBA adversarial training

To investigate whether ZQBA and AT provide complementary robustness, we evaluate the effect of combining ZQBA pretraining with PGD AT (+ZQBA) and assess whether the sequential application of ZQBA perturbations as a pretraining strategy followed by PGD AT yields cumulative improvements over PGD AT alone (PGD-AT).

Adversarial Training on ZQBA Performance. The first step of evaluation for ZQBA-AT consists in assessing the effect of the ZQBA perturbations on different models, for the considered datasets, whose accuracy is displayed in Table 8. The results show that the usage of ZQBA-AT can overall improve the robustness against ZQBA attack, at the cost of reducing the accuracy of the smaller architectures on clean settings. Despite the inclusion of ZQBA samples during the training phase, the models are still slightly affected by the ZQBA attack (i.e.,

Table 9

Performance on state-of-the-art white-box attacks of PGD Adversarial Training, and respective addition of ZQBA pretrain, considering a ResNet model in different datasets.

Dataset	Models	White-box accuracy					
		FGSM	PGD-CE-10	PGD-CE-40	PGD-DLR	AA	AA+
CIFAR-10	PGD-AT	48.03 $\pm$ 0.51	42.80 $\pm$ 0.50	41.54 $\pm$ 0.52	41.07 $\pm$ 0.37	38.91 $\pm$ 0.29	38.84 $\pm$ 0.29
	+ZQBA	51.59 $\pm$ 0.14	46.61 $\pm$ 0.11	45.41 $\pm$ 0.14	44.14 $\pm$ 0.08	42.15 $\pm$ 0.14	42.08 $\pm$ 0.12
CIFAR-100	PGD-AT	21.83 $\pm$ 0.23	19.83 $\pm$ 0.40	19.34 $\pm$ 0.49	17.71 $\pm$ 0.14	15.98 $\pm$ 0.29	15.97 $\pm$ 0.27
	+ZQBA	23.25 $\pm$ 0.30	21.10 $\pm$ 0.29	20.67 $\pm$ 0.19	18.84 $\pm$ 0.18	16.86 $\pm$ 0.21	16.86 $\pm$ 0.20
Tiny	PGD-AT	13.21 $\pm$ 0.17	11.42 $\pm$ 0.20	11.05 $\pm$ 0.08	9.08 $\pm$ 0.15	7.61 $\pm$ 0.05	7.59 $\pm$ 0.06
	+ZQBA	17.40 $\pm$ 0.39	15.62 $\pm$ 0.41	15.23 $\pm$ 0.36	12.92 $\pm$ 0.22	11.50 $\pm$ 0.15	11.49 $\pm$ 0.15

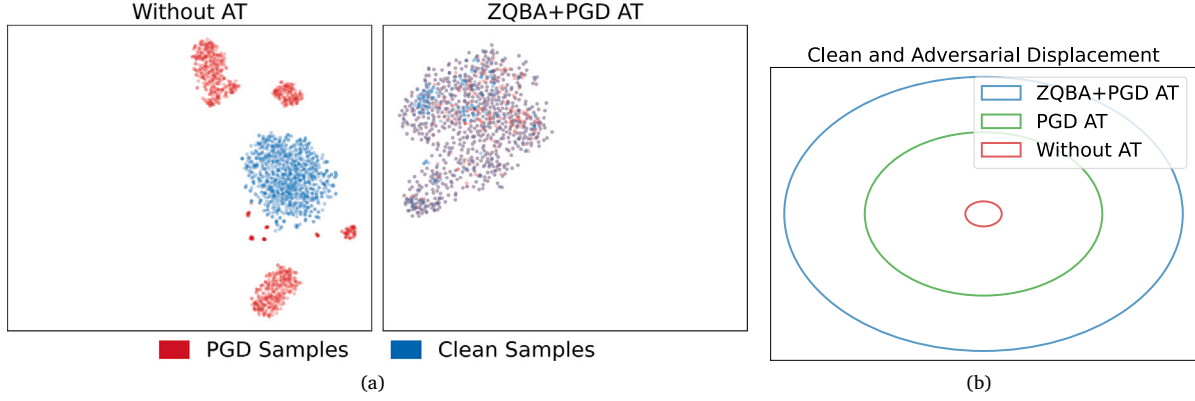


Fig. 6. Visualization of t-SNE distribution (on the left) for adversarial and clean samples of one class in the CIFAR-10 dataset, and displacement between the Clean and Adversarial Samples (on the right) using t-SNE values, for models without AT, with PGD AT, and with the combination of ZQBA-AT pretrain and PGD AT, in CIFAR-10 dataset.

Table 10

Performance on state-of-the-art black-box attacks with increasing number of queries of PGD Adversarial Training, and respective addition of ZQBA pretrain, considering a ResNet model in different datasets. C-10, C-100, and Tiny refer to the CIFAR-10, CIFAR-100, and Tiny ImageNet datasets.

Dataset	Models	Square			RayS		
		50	1000	10 000	50	1000	10 000
C-10	PGD-AT	66.23	48.72	41.90	72.60	73.32	73.49
		± 0.59	± 0.49	± 0.31	± 0.72	± 0.33	± 0.28
	+ZQBA	73.31	55.34	47.27	77.57	78.76	78.95
		± 0.27	± 0.09	± 0.14	± 0.83	± 0.15	± 0.14
C-100	PGD-AT	34.86	20.88	17.03	41.96	42.36	42.58
		± 0.57	± 0.54	± 0.26	± 0.81	± 0.57	± 0.74
	+ZQBA	39.75	24.80	20.37	46.06	46.75	46.77
		± 0.36	± 0.41	± 0.33	± 0.55	± 0.34	± 0.49
Tiny	PGD-AT	25.90	15.70	12.15	31.31	31.53	31.64
		± 0.37	± 0.25	± 0.28	± 0.58	± 0.61	± 0.59
	+ZQBA	30.98	20.12	16.27	35.57	36.18	36.34
		± 0.49	± 0.42	± 0.46	± 0.53	± 0.55	± 0.57

ZQBA accuracy is less than Clean), which can be explained by the wide variety of examples that can be created. Additionally, when considering the results for CIFAR-100 dataset, we can observe that, globally, the models obtain a greater improvement, since this dataset contains a small number of examples for each class and highly benefits from the increase of data available during training. The results highlighted in 4 show that WideResNet28-10 consistently benefits the most from ZQBA-AT to improve the accuracy for ZQBA and, when compared with Table 1, there are also increases in Clean accuracy for WideResNet28-10, across the three datasets. This suggests that wider networks have a greater capacity of hyperplane refinement and adequate clustering.

ZQBA Pretrain on Various Attack Settings. Since ZQBA proposes the creation of feature-aware perturbations, which differs from the approaches commonly used in *white-box* attacks, we wanted to assess

Table 11

Performance on gray-box attacks of PGD Adversarial Training, and respective addition of ZQBA pretrain, considering a ResNet model in different datasets. These attacks are obtained white-box attacks to surrogate models. Tiny refers to the Tiny ImageNet dataset.

Dataset	Models	Gray-box accuracy		
		FGSM	PGD	AA
CIFAR-10	PGD-AT	72.53 $\pm$ 0.30	72.81 $\pm$ 0.17	72.96 $\pm$ 0.28
	+ZQBA	77.67 $\pm$ 0.08	78.07 $\pm$ 0.11	78.29 $\pm$ 0.14
CIFAR-100	PGD-AT	42.42 $\pm$ 0.91	42.32 $\pm$ 0.84	42.64 $\pm$ 0.86
	+ZQBA	46.19 $\pm$ 0.45	46.07 $\pm$ 0.42	46.43 $\pm$ 0.51
Tiny	PGD-AT	31.37 $\pm$ 0.55	31.48 $\pm$ 0.55	31.60 $\pm$ 0.53
	+ZQBA	36.35 $\pm$ 0.43	36.48 $\pm$ 0.40	36.63 $\pm$ 0.51

if the inclusion of ZQBA-AT could improve the standard AT (*i.e.*, using only PGD). The model used throughout the following experiments is ResNet18, with the ZQBA perturbations obtained from a MobileNetV2, for CIFAR-10 and CIFAR-100, and from a ResNet18 for Tiny ImageNet, which showed higher attack success rate in Table 2. The reported results presume that the “+ZQBA” addition is loading the weights and performing standard AT. Tables 9, 10 and 11 display the performance on *white-box*, *black-box*, and *gray-box* settings, respectively. The evaluation on *white-box* attacks show a slight improvement of accuracy from the inclusion of ZQBA-AT pretraining, across the three datasets, which can be justified by the different approach used by ZQBA (*i.e.*, feature-aware perturbations). To further evaluate whether the observed robustness resulted from enhanced generalization capacity, we assess the effect of a *black-box* attacks considering varying iterations. In this scenario, regardless of the impact of the number of queries to the attack success, ZQBA-AT pretraining yields better results than PGD-AT alone (7%–4%), which is consistently retained across the three datasets. To conclude this part of the analysis, we extend our evaluation to the *gray-box* settings, using adversarial samples previously generated for a

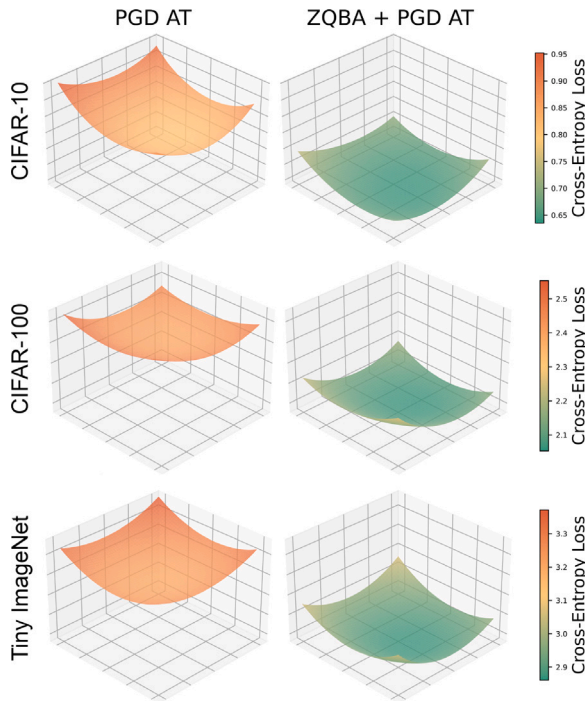


Fig. 7. The loss surfaces of PGD AT, on the left, and ZQBA-AT pretrain with PGD AT, on the right, in different datasets. The surface reflects the variation of the loss function (z-axis) when susceptible to random perturbation vectors (represented by x- and y-axes). Smoother surface represents better generalization.

ResNet18 using *white-box* attacks. The results confirm that the accuracy increases, in a similar proportion to the one observed in the *black-box* context, across the datasets and attacks. These results were achieved by only loading the weights of a previously trained model and can be applied to multiple AT approaches. In our setting, the inclusion of ZQBA-AT pretraining can transversely increase the robustness on the three attack settings, suggesting that further exploring feature-aware perturbations could be a future strategy to increase generalization capacity.

High-dimensional Data Visualization. To clarify the overall improvement provided by the inclusion of ZQBA-AT pretrain, we use the t-SNE [41] algorithm to conceptualize the distributions of the clean and adversarial samples for CIFAR-10. This dataset was chosen because it contains the largest number of images per class. For the sake of the explanation, we selected a random class to produce the t-SNE visualization, due to the clutter caused by using 10 classes and a similar effect is observed throughout all classes. Fig. 6(a) displays the cluster distribution for Clean and Adversarial examples, for the “Frog” class, considering a model without AT and one using ZQBA-AT pretrain and PGD-AT. It suggests that a model without any AT is not able to create cluster containing both Clean and Adversarial images, confirming the decrease in accuracy usually seen with adversarial attacks, and that the inclusion of AT adapts the hyperplanes to include both of these. To have a global picture, we obtained the t-SNE values for all clean and adversarial images, calculated the Euclidean distance between them, and obtained the mean value for this distance. Fig. 6(b) represents this distance, showing that the combination of ZQBA-AT pretrain and PGD-AT has a bigger distance than the remaining models, which is associated with the proposed AT creating broader clusters, as illustrated in Fig. 3.

Loss Landscape Analysis. We finish the analysis of whether including feature-aware perturbations helps increase the model generalization capacity by visualizing its loss landscape [42]. Fig. 7 demonstrates loss surfaces for PGD AT and combination of ZQBA-AT pretrain and

PGD AT, across the considered datasets. Overall, we observe that the loss value for the proposed AT is lower than PGD AT one, explaining the improvement seen across different scenarios and suggesting that feature-aware perturbations act as a loss regularizer. The flat surfaces seen in both AT approaches suggest that these models are less susceptible to small perturbations, achieving greater generalization. These qualitative tests, together with the quantitative results, suggest that incorporating feature-aware perturbations into AT can improve model robustness, and this was achieved solely by loading the weights of a previously trained model.

5. Conclusions

This paper investigates the relevance of incorporating feature-aware perturbations, generated by ZQBA, into training as a strategy to enhance model robustness across multiple attack scenarios. This work follows as a corollary of the ZQBA proposal presented in [14], which comprises an attack that leverages feature maps to create perturbations that can fool target models without querying them. By loading the weights of ZQBA AT into an AT approach, the models achieve improved generalization under white-box, black-box, and gray-box attacks. These results motivate the development of novel robustness approaches that integrate feature-aware perturbations, given the observed regularizing effect on AT in our experiments. Building on the insights from this paper, future work could evaluate this methodology in other computer vision tasks and other data types (e.g., audio or text), and investigate the effect of feature-aware perturbations in Transformer architectures.

CRedit authorship contribution statement

Joana Cabral Costa: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Tiago Roxo:** Writing – review & editing, Validation, Methodology, Conceptualization. **Hugo Proença:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Pedro R.M. Inácio:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Joana Cabral Costa reports financial support was provided by Foundation for Science and Technology. Tiago Roxo reports financial support was provided by Foundation for Science and Technology. Joana Cabral Costa reports financial support was provided by Institute for Telecommunications. Tiago Roxo reports financial support was provided by Institute for Telecommunications. Hugo Proença reports financial support was provided by Institute for Telecommunications. Pedro R. M. Inacio reports financial support was provided by Institute for Telecommunications. Given their role as Associate Editor, Hugo Proença had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to another journal editor. If there are other authors, they 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

Research Link Provided

ZQBA: Zero Query Black-box Adversarial Attack (Original data) (GitHub)

References

- [1] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023, URL: <https://arxiv.org/abs/2307.09288>, arXiv:2307.09288.
- [2] J.C. Costa, T. Roxo, J.B. Sequeiros, H. Proença, P.R.M. Inácio, Predicting CVSS metric via description interpretation, *IEEE Access* 10 (2022) 59125–59134.
- [3] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, *Nature Med.* 29 (2023) 1930–1940.
- [4] C. Patrício, J.C. Neves, L.F. Teixeira, Coherent concept-based explanations in medical image and its application to skin lesion diagnosis, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3799–3808.
- [5] T. Roxo, J.C. Costa, P.R. Inácio, H. Proença, On exploring audio anomaly in speech, in: *2023 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE*, 2023, pp. 1–6.
- [6] T. Roxo, J.C. Costa, P.R. Inácio, H. Proença, Bias: A body-based interpretable active speaker approach, *IEEE Trans. Biom. Behav. Identity Sci.* (2024).
- [7] T. Roxo, J.C. Costa, P.R. Inácio, H. Proença, Asdnb: Merging face with body cues for robust active speaker detection, in: *2025 IEEE International Joint Conference on Biometrics, IJCB, IEEE*, 2025, pp. 1–10.
- [8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: *2nd International Conference on Learning Representations, ICLR 2014*, 2014.
- [9] A.K. Aggarwal, Explainable and uncertainty-aware deep learning for image classification under noisy and augmented conditions, *Int. J. Signal Process.* 10 (2025) 39–42.
- [10] A.K. Aggarwal, Comparative analysis of non-local means and hybrid median bilateral filtering for mobile image denoising, *Smart Wearable Technol.* (2024).
- [11] P.Y. Chen, H. Zhang, Y. Sharma, J. Yi, C.J. Hsieh, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.
- [12] J. Liu, J. Zhou, J. Zeng, J. Tian, Difattack: Query-efficient black-box adversarial attack via disentangled feature space, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 3666–3674.
- [13] J. Chen, H. Chen, K. Chen, Y. Zhang, Z. Zou, Z. Shi, Diffusion models for imperceptible and transferable adversarial attack, *IEEE Trans. Pattern Anal. Mach. Intell.* (2024).
- [14] J.C. Costa, T. Roxo, H. Proença, P.R.M. Inácio, ZQBA: Zero query black-box adversarial attack, in: *Proceedings of the 18th International Conference on Agents and Artificial Intelligence - Volume 5: ICAART, INSTICC, SciTePress*, 2026b, pp. 3982–3989, <http://dx.doi.org/10.5220/0014237300004052>.
- [15] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2014, arXiv preprint arXiv:1412.6572.
- [16] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE, 2016, pp. 372–387.
- [17] S.M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.
- [18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: *ICLR*, 2018.
- [19] A. Dabouei, S. Soleymani, F. Taherkhani, J. Dawson, N. Nasrabi, Smoothfool: An efficient framework for computing smooth adversarial perturbations, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2665–2674.
- [20] F. Croce, M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, in: *ICML, PMLR*, 2020, pp. 2206–2216.
- [21] M. Andriushchenko, F. Croce, N. Flammarion, M. Hein, Square attack: a query-efficient black-box adversarial attack via random search, in: *ECCV*, Springer, 2020, pp. 484–501.
- [22] J. Chen, Q. Gu, Rays: A ray searching method for hard-label adversarial attack, in: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1739–1747.
- [23] J. Park, P. Miller, N. McLaughlin, Hard-label based small query black-box adversarial attack, in: *Proceedings of the IEEE/CVF WACV*, 2024, pp. 3986–3995.
- [24] S. Roy, A. Jain, M. Vatsa, R. Singh, Taigen: Training-free adversarial image generation via diffusion models, 2025, arXiv preprint arXiv:2508.15020.
- [25] C. Huang, H. Tang, Scoreadv: Score-based targeted generation of natural adversarial examples via diffusion models, 2025, arXiv preprint arXiv:2507.06078.
- [26] D. Wu, S.T. Xia, Y. Wang, Adversarial weight perturbation helps robust generalization, *Adv. Neural Inf. Process. Syst.* 33 (2020) 2958–2969.
- [27] H. Kuang, H. Liu, Y. Wu, R. Ji, Semantically consistent visual representation for adversarial robustness, *IEEE Trans. Inf. Forensics Secur.* 18 (2023) 5608–5622.
- [28] F. Li, K. Li, H. Wu, J. Tian, J. Zhou, Towards robust learning via core feature-aware adversarial training, *IEEE Trans. Inf. Forensics Secur.* (2025).
- [29] Y. Zhang, Y. Chen, Z. Chen, W. Ruan, X. Huang, S. Khastgir, X. Zhao, Adversarial training for probabilistic robustness, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 1675–1685.
- [30] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.* 13 (2004) 600–612.
- [31] J.C. Costa, T. Roxo, H. Proença, P.R. Inácio, LISArD: learning image similarity to defend against gray-box adversarial attacks, *PeerJ Comput. Sci.* 12 (2026a) e3735.
- [32] S. Mostafa, D. Mondal, M.A. Beck, C.P. Bidnost, C.J. Henry, I. Stavness, Leveraging guided backpropagation to select convolutional neural networks for plant classification, *Front. Artif. Intell.* 5 (2022) 871162.
- [33] J.R. Flynn, S. Ward, J. Abich IV, D. Poole, Image quality assessment using the ssim and the just noticeable difference paradigm, in: *International Conference on Engineering Psychology and Cognitive Ergonomics*, Springer, 2013, pp. 23–30.
- [34] A. Krizhevsky, G. Hinton, et al., Learning Multiple Layers of Features from Tiny Images, (Master's Thesis), University of Tront, 2009.
- [35] Y. Le, X. Yang, Tiny imagenet visual recognition challenge, *CS 231N* 7 (2015) 3.
- [36] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 248–255.
- [37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [38] M.B. Muhammad, M. Yeasin, Eigen-cam: Class activation map using principal components, in: *2020 International Joint Conference on Neural Networks, IJCNN, IEEE*, 2020, pp. 1–7.
- [39] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [40] S. Srinivas, F. Fleuret, Full-gradient representation for neural network visualization, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [41] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, *J. Mach. Learn. Res.* 9 (2008).
- [42] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2018.