How stealthy is stealthy? Studying the Efficacy of Black-Box Adversarial Attacks in the Real World

Francesco Panebianco[0009−0007−1510−2594], Mario D'Onghia[0000−0001−9467−1523], Stefano Zanero[0000−0003−4710−5283], and Michele Carminati[0000−0001−8284−6074]

Politecnico di Milano, Milano Italy
Dipartimento di Elettronica, Informazione e Bioingegneria
{name.surname}@polimi.it

Abstract. Deep learning systems, critical in domains like autonomous vehicles, are vulnerable to adversarial examples (crafted inputs designed to mislead classifiers). This study investigates black-box adversarial attacks in computer vision. This is a realistic scenario, where attackers have query-only access to the target model. Three properties are introduced to evaluate attack feasibility: robustness to compression, stealthiness to automatic detection, and stealthiness to human inspection. State-of-the-Art methods tend to prioritize one criterion at the expense of others. We propose ECLIPSE, a novel attack method employing Gaussian blurring on sampled gradients and a local surrogate model. Comprehensive experiments on a public dataset highlight ECLIPSE's advantages, demonstrating its contribution to the trade-off between the three properties.

Keywords: Evasion · Adversarial Examples · Computer Vision · Machine Learning · Security · Deep Learning · Black-Box · Stealthiness

1 Introduction

Deep learning models for image classification and object detection are crucial in applications like self-driving vehicles [7], where they identify vehicles, pedestrians, traffic signs, and obstacles. Detection errors pose significant risks to passengers and others. Many proprietary classifiers and object detection networks are accessible via public Application Programming Interfaces (APIs), increasing their exposure. Examples of such APIs are api4ai [1] and Clarifai [2]. Evasion attacks exploit adversarial examples, malicious inputs crafted with noise patterns, to induce misclassification. Early works [25, 37, 10] demonstrated the feasibility of such attacks under an unrealistic white-box threat model, where attackers have full knowledge of the target system. Black-box *query-based* attacks [11, 4, 15, 18, 21, 6, 8, 24, 26, 38] operate under the more practical constraint of interacting with the model only through remote queries. White-box attacks benefit from gradient information, which guides the optimization process [37, 25, 10]. In contrast, black-box attacks address the absence of gradient access in two ways: by estimating gradients from queries when confidence scores are available [11], or using label-only strategies, which are generally less efficient [8, 24, 26, 38]. While label-only attacks are often the only viable option for many AI endpoints, computer vision endpoints frequently disclose top-k confidence scores [1, 2], enabling more effective

attack strategies. In this work, we explore the limitations and strengths of computer vision evasion attacks relative to their chances to succeed in a real-world scenario. We concentrate on some of the most powerful options in the black-box setting with confidence scores: SimBA [21], SimBA-DCT [21], and the Square Attack [6]. Prior work has provided definitions of "deployability" [18] of an attack in the real world. To the best of our knowledge, no existing work has presented a comprehensive formalization encompassing all characteristics of successful real-world attacks. We introduce a framework consisting of three *effectiveness properties*: *Robustness to Compression* (P1), *Stealthiness to Automatic Detection* (P2), and *Stealthiness to Human Inspection* (P3). The experimental evaluation shows that existing black-box attacks fulfill some effectiveness properties at the expense of others. While SimBA-DCT demonstrates strong performance on P3, its results on P1 and P2 are significantly weaker. Similarly, SimBA performs well on P2 but exhibits poor performance on P1 and P3. In contrast, the Square Attack shows a slight advantage on P1 but fails to achieve satisfactory results on P2 and P3. To address this, we propose *ECLIPSE* (**E**vasion of **C**lassifiers with **L**ocal **I**ncrease in **P**ixel **S**parse **E**nvironment). The attack resists JPEG compression (Joint Photographic Experts Group) while evading both automatic and human detection. ECLIPSE is a confidence-based black-box evasion attack based on Hill Climbing [30], a well-known metaheuristic for the optimization of functions lacking a closed form. The algorithm metaphorically "climbs the hill" of the objective function by iteratively improving the best solution found so far. Although ECLIPSE is not the first adversarial attack to rely on this method [16, 11], its notable effectiveness is attributed to two novel steps integrated within the optimization process. The first step involves applying Gaussian blurring [19] to the estimated gradients before updating the adversarial example. A comprehensive evaluation on a public dataset highlights the limitations of state-of-the-art attack methods while demonstrating the superior performance of ECLIPSE in resilience to image compression, detectability, and visual stealthiness.

We summarize the contributions of our research in what follows:

- We formalize the real-world feasibility of adversarial examples with three measurable effectiveness properties.
- We introduce ECLIPSE, a novel attack designed to achieve a balanced trade-off between all the effectiveness properties.
- We evaluate ECLIPSE in terms of effectiveness properties, comparing it with state-of-the-art baselines to identify their advantages and drawbacks.

2 Background and Motivation

Classifier *evasion* attacks induce misclassification in victim models by crafting *adversarial examples* — inputs perturbed with specialized noise. These attacks are categorized as *targeted*, where the target label is set, or *untargeted* otherwise. Attackers leverage prior knowledge (e.g., architecture, weights) in white-box attacks and empirical guidance (e.g., input-output behavior) in black-box attacks. Adversarial examples exhibit *transferability* [37, 20, 28, 33, 23], allowing them to fool multiple classifiers. Transfer-based attacks exploit this by crafting adversarial examples using surrogate models for black-box targets, albeit with reduced efficacy compared to native

black-box methods [23]. Initial efforts focused on white-box attacks (e.g., FGSM [20], I-FGSM [25], Carlini&Wagner [10]), but the impracticality of assuming full model access has redirected attention to black-box scenarios. Here, the model serves as a remote oracle providing either confidence scores or predicted labels. Confidence-based attacks, leveraging scores for efficient gradient estimation [11], are practical in domains like computer vision, where services often expose *top-k* confidence scores [1, 2].

Black-Box Attacks with Confidence Scores. Guo et al. [21] introduced *SimBA* (Simple Black-box Attack), which minimizes the search space using orthonormal basis perturbations of fixed step size. The same work introduces SimBA-DCT, which leverages the same strategy but translates it to the frequency domain. Andriushchenko et al. [6] formalized the *Square Attack*, employing random search with L_2 or L_∞ norm constraints. Giulivi et al. [18] proposed *Adversarial Scratches*, generating deployable adversarial examples via superimposed bezier curves. Among the recently proposed attack methods, SimBA, SimBA-DCT, and the Square Attack are particularly notable due to their widespread popularity and frequent use as baseline approaches in comparative studies.

Defenses. Adversarial example research exhibits a persistent cat-and-mouse dynamic typical of information security, with defenses repeatedly proposed and circumvented. While a comprehensive review is beyond this discussion's scope, Ahmed Aldahdooh et al. [3] provide a detailed overview of adversarial defenses. In real-world applications, preprocessing pipelines can disrupt adversarial perturbations. These pipelines may include defenses explicitly designed to counter such attacks. For example, Byun et al. [9] propose Small Noise Defense, which adds minor Gaussian noise to neutralize many black-box adversarial examples. However, even routine and often overlooked preprocessing steps, particularly image compression can significantly degrade the efficacy of adversarial examples.

2.1 Motivation

While prior work addresses the deployability of adversarial examples [18, 34], a systematic framework for defining attack realism remains absent. Three critical factors can impact attack success: (1) *Image Processing*: Compression methods like JPEG, which reduce file size while preserving perceptible details, can also erase subtle adversarial perturbations. (2) *Adversarial Detection*: Pre-inference detection algorithms can thwart attacks, necessitating minimal queries to avoid alerting the service provider. (3) *Visual Stealthiness*: Prominent adversarial noise increases the risk of detection and mitigation during deployment.

Threat Modeling. We consider an attacker aiming to mislead a computer vision classifier via remotely deployed adversarial examples. The attacker queries the model to obtain *confidence scores*, lacking knowledge of the architecture, weights, or training data. Confidence-based attacks enhance reliability by enabling gradient estimation through inference results [11]. Since confidence score access is common in computer vision systems, this attack scenario is realistic. We evaluate state-of-the-art confidence-based attacks, including *SimBA* [21], *SimBA-DCT* [21], and the *Square Attack* [6].

To ensure consistency, we focus on the L_∞-bound version of the Square Attack, aligning with the L_∞ bounds of the other methods.

3 Effectiveness properties of adversarial examples

Few works on classifier evasion address the real-world deployability of their proposed solutions. We formalize these considerations as measurable features of images and refer to them as *effectiveness properties*. These properties are organized into three orthogonal dimensions: *Robustness to Compression* (P1); *Stealthiness to Automatic Detection* (P2); *Stealthiness to Human Inspection* (P3). While some of these properties were individually evaluated in prior work [18], to the best of our knowledge we are the first to introduce such formalization of attack effectiveness in the real world.

Robustness to Compression (P1). Images shared on the internet often undergo various processing operations, such as resizing and compression, which are determined by the requirements of the hosting service. Compression, particularly the JPEG format (Joint Photographic Experts Group), is the most prevalent, serving to reduce data transfer and storage costs. As a form of processing, compression can impact the efficacy of adversarial examples to varying degrees. This characteristic was already partially formalized by prior work [18], although it was considered the only property to make attacks "deployable". This property is assessed by measuring the proportion of adversarial examples whose prediction confidence drops below a specified threshold following image compression.

Stealthiness to Automatic Detection (P2). Frequency-domain detection has demonstrated high efficacy against state-of-the-art black-box attacks. While ideal for perturbations crafted in the frequency domain (e.g., SimBA-DCT [21]), it can also detect some pixel-space attacks, as detailed in Section 5. Remote classification services are accessed via repeated queries, making attack patterns susceptible to stateful defenses [12, 17, 27, 13]. While recent methods like OARS [35] allow black-box attacks to evade such measures, firewalls may still flag anomalous query spikes, and frequent queries can inflate costs.

Stealthiness to Human Inspection (P3). Stealthiness to human inspection denotes the ability of adversarial examples to appear indistinguishable from benign inputs to human observers, achieved through minimal perturbations that blend with natural image features. Commonly, the perturbation norm [21, 6, 4] is minimized to improve stealthiness, though it inadequately captures human perception. A survey by Liu et al. [29] highlights the absence of a universal metric for this purpose, necessitating human evaluations for reliable assessments. Accordingly, this property is evaluated by requiring human auditors to assess the extent to which the image appears altered.

4 ECLIPSE

We propose *ECLIPSE* (**E**vasion of **C**lassifiers with **L**ocal **I**ncrease in **P**ixel **S**parse **E**nvironment), a targeted evasion attack incorporating two novel techniques aimed at satisfying all three effectiveness properties. Figure 1 illustrates how these techniques

Algorithm 1 ECLIPSE Algorithm, where f is the remote oracle returning the score of the target class, H is the GradCAM heatmap from the local model, M is the mask, τ_t is the mask threshold at iteration t

Require: the original image x, L_∞ perturbation budget β, the maximum number of iterations I, ϵ_0, the sample size s, the gaussian blur kernel size k, the gaussian distribution's standard deviation σ,$width$,$height$

Get GradCam heatmap from the local model $H = GradCAM_{local}(x)$
Initialize current best solution $C_0 = x$
Initialize gradient buffers $\nabla[i,j,c] = 0 \quad \forall i \in [1,height], j \in [1,width], c \in [1,3]$
Initialize mask $M_0[i,j] = 1 \quad \forall i \in [1,height], j \in [1,width]$
$fitness_0 = f(x)$
Initialize mask threshold $\tau_0 = 0.0$
for $t = 1$ to I **do**
    Sample batch B_s of s coordinates (i,j,c) in M_{t-1} without replacement
    $\nabla[i,j,c] = f(C_{t-1} + \mathbb{1}_{(i,j,c)}) - f(C_{t-1}) \quad \forall (i,j,c) \in B_s$
    Copy gradients to be processed $\delta = \nabla$
    $\delta = GaussianBlur(\delta, (k,k), \sigma)$
    $A_t = C_{t-1} + \epsilon_{t-1} \frac{\delta}{\text{maxabs}(\delta)}$
    Clip A_t such that $-\beta \leq A_t[i,j,c] - x[i,j,c] \leq \beta \quad \forall i \in [1,height], j \in [1,width], c \in [1,3]$
    Clip A_t such that $0 \leq A_t[i,j,c] \leq 1 \quad \forall i \in [1,height], j \in [1,width], c \in [1,3]$
    if $f(A_t) > fitness_{t-1}$ **then**
        $fitness_t = f(A_t)$
        $\epsilon_t = \max\{0.02, 0.95\epsilon_{t-1}\}$
    else
        $fitness_t = fitness_{t-1}$
    end if
    $\tau_t = \min\{0.5, \tau_{t-1} + 0.01\}$
    $M_t = ThresholdMask(H, \tau_t)$
    if $Area(M_t) < min\_area$ **or** we already sampled $> 0.75 Area(M_t)$ **then**
        $M_t[i,j] = 1 \quad \forall i \in [1,height], j \in [1,width]$
    end if
    if $fitness_t > fitness_{t-1}$ **then**
        $C_t = A_t$
        if $fitness_t > 0.5$ **then return** C_t
        end if
    end if
end for
Return failure if no solution is found within the max number of iterations

are embedded in the attack procedure. The first technique involves processing estimated gradients with Gaussian blurring [19]. The second technique masks the gradient sampling area using information from a local surrogate model. These two components are integrated into an optimization loop that performs *Hill Climbing* [30], a popular meta-heuristic that iteratively improves the candidate solution, effectively "climbing the hill" of the objective function. The quality of the solution is assessed based on the confidence of the prediction. Consequently, while this approach is effective when confidence scores are available, it is not applicable in label-only settings.

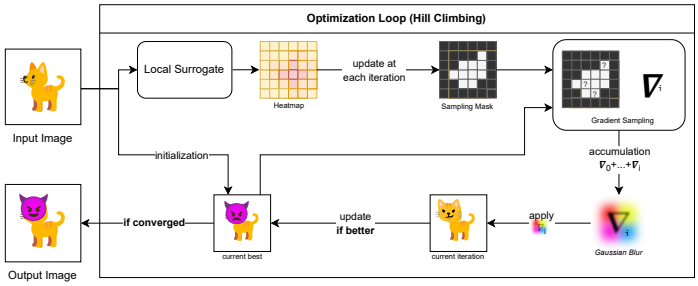


Fig. 1. Main steps of the ECLIPSE algorithm: The perturbation mask is computed initially, while other steps iterate until convergence.

Fig. 2. Example comparison of GradCAM heatmaps on a local surrogate and a remote model. Cat image is from the Animals-10 [14] dataset.

Local Surrogate. Adversarial optimization benefits from identifying relevant input features, which is challenging in query-only scenarios. Surrogate models trained on the same task can approximate remote behavior. We thus leverage a white-box explainability technique (GradCAM [36], Gradient-weighted Class Activation Mapping) to generate saliency maps on the local surrogate. These maps guide perturbations by creating masks that restrict gradient estimation to relevant areas, improving convergence. Figure 2 shows this concept in practice.

Gaussian Blurring. Gaussian blurring [19], a convolutional process with Gaussian filters, is used in ECLIPSE to smooth estimated gradients rather than image pixels. This approach reduces query count by interpolating sparse gradient values to neighbor coordinates, assuming gradient continuity. While Gaussian blur has been used in training optimizers [5] and defensive techniques [32], ECLIPSE uniquely integrates it into gradient-based adversarial attacks.

Attack Parameter Scheduling. ECLIPSE employs adaptive scheduling of attack parameters to facilitate convergence. The noise multiplier (*learning rate*) is exponentially decreased to stabilize convergence [22]. Similarly, the mask threshold is linearly reduced over iterations, focusing perturbations on increasingly relevant areas. These strategies ensure efficient and targeted attack progression.

5 Experimental Evaluation

All experiments have been performed on a ResNet152V2 model as the remote oracle. The local surrogate for ECLIPSE experiments is a DenseNet201. It was trained on a subset of ImageNet with 12 classes. Experiments have been run on the Animals-10 dataset [14], where images belonging to the ground truth of *cat* are used to generate targeted adversarial examples to be misclassified as *dog*.

Robustness to Processing (P1). To assess the robustness of adversarial examples against common image processing during upload, we evaluated JPEG compression, a prevalent internet image format. Table 2 enumerates attack parameters for the experiments. Three metrics were used to evaluate robustness: the median confidence score difference, the percentage of adversarial examples with low confidence loss (below 0.3), and the percentage of examples that "survived" compression (losing less than 0.05 confidence, remaining effective or improving). Results, summarized in Table 1, highlight the varying levels of robustness across the attacks. We can see from this evaluation that ECLIPSE excels in minimizing the loss for the vast majority of adversarial examples it generates. ECLIPSE and the Square Attack L_∞ have a similar survival rate, which is much higher than that of other attacks. In the median case, however, the Square Attack L_∞ performs much worse. Both SimBA and SimBA-DCT fail to produce compression-surviving perturbation, with a very low amount of adversarial examples still effective after processing. The remarkable performance of ECLIPSE, and to a lesser extent that of the Square Attack L_∞, is likely attributable to the coarser perturbation granularity. In fact, the former performs Gaussian blurring, while the latter overlays large squares on the original image.

Stealthiness to Automatic Detection (P2). We evaluated adversarial attack strategies on two aspects: detection avoidance and query efficiency. Detection was analyzed using a classifier trained to separate original images from adversarial examples, with features extracted from Discrete Cosine Transform (DCT) spectra. Experiments were conducted under consistent parameters for four attacks (Table 2), using 150 samples per attack. Adversarial attacks introduce distinct spectral changes observable in processed DCT spectra (Figure 3). SimBA-DCT perturbs primarily lower frequencies, creating a square of intensity, while Square Attack L_∞ introduces high-frequency artifacts due to its noise initialization process. We use t-SNE [31] (t-distributed Stochastic Neighbor Embedding) to perform dimensionality reduction. The technique identifies a lower-dimensional manifold that preserves the local structure of the high-dimensional input. This property is particularly useful for exploratory data analysis, as it eases the rapid recognition of data similarities. We observe distinct

Table 1. Comparison metrics for JPEG compression on different attacks.

Attack	Median Loss	Low-loss%	Surviving%
ECLIPSE	**0.15**	**89.33**	**18.67**
SimBA	0.5	4.00	2.00
SimBA-DCT	0.5	4.67	1.33
Square Attack L_∞	0.50	26.67	15.33

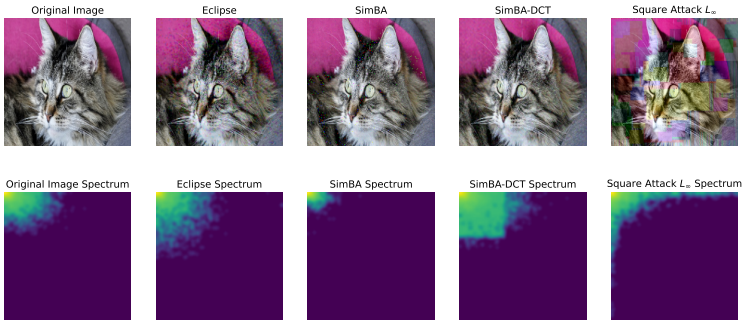


Fig. 3. Visual comparison of the processed DCT spectra of adversarial examples generated by each attack against the unaltered image (leftmost).

clusters for Square Attack L_∞ and, partially, for SimBA-DCT (Figure 4). Binary Support Vector Machines (SVMs) with polynomial kernels were trained to classify adversarial examples. ECLIPSE and SimBA were indistinguishable from benign images, while SimBA-DCT achieved actionable separability (precision = 1.0, recall = 0.47), and Square Attack L_∞ was highly detectable (Area Under the Curve, AUC = 0.96).

Query efficiency, measured as the number of calls to the victim model, showed that SimBA-DCT and Square Attack L_∞ converge faster than others, but this efficiency is undermined by their high detectability. The Square Attack required fewer queries but failed to converge for 13/150 samples. Despite superior efficiency, its high detectability diminishes practical utility. Our findings highlight the trade-offs between attack efficiency and susceptibility to detection, emphasizing the need for balanced evaluation metrics.

Stealthiness to Human Inspection (P3). To address the capacity of each attack to transparently blend and avoid human detection, we have administered a survey to 127

Table 2. Parameters chosen for comparisons on automatic detection and robustness to processing. k is the Gaussian blurring kernel size. p is the Square Attack's start area ratio. N/A indicates the parameter does not apply to the attack procedure.

Attack	Step Size	L_∞ budget	k	p	Max Iterations
ECLIPSE	0.1	0.1	3	N/A	1000
SimBA	0.1	0.1	N/A	N/A	100000
SimBA-DCT	0.1	0.1	N/A	N/A	100000
Square Attack L_∞	0.1	0.1	N/A	0.2	10000

Table 3. Cross-validation metrics of binary Support Vector Machine Classifiers to distinguish original images from adversarial examples using processed spectral features. The area under the ROC curve is highlighted in bold.

Comparison	Accuracy	Precision	Recall	F1-score	**ROC AUC**
Normal vs ECLIPSE	0.87 (± 0.01)	0.03 (± 0.20)	0.01 (± 0.08)	0.02 (± 0.11)	**0.50 (± 0.03)**
Normal vs SimBA	0.90 (± 0.03)	0.80 (± 0.33)	0.26 (± 0.23)	0.39 (± 0.29)	**0.63 (± 0.12)**
Normal vs SimBA-DCT	0.93 (± 0.03)	1.00 (± 0.00)	0.47 (± 0.25)	0.63 (± 0.24)	**0.73 (± 0.12)**
Normal vs Square Attack L_∞	0.99 (± 0.02)	0.99 (± 0.05)	0.91 (± 0.17)	0.95 (± 0.10)	**0.96 (± 0.09)**

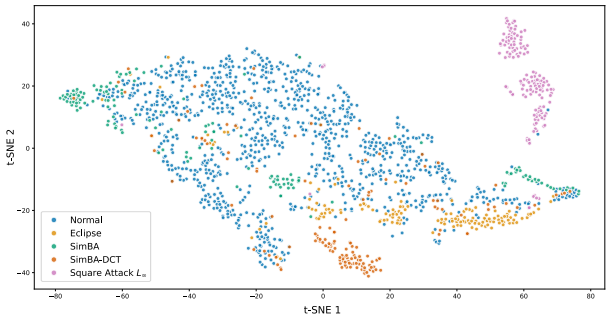


Fig. 4. Scatterplot of projected spectral features using t-SNE dimensionality reduction.

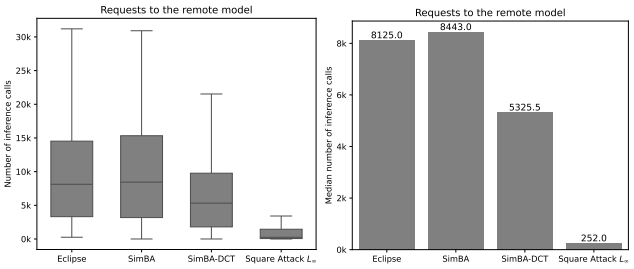


Fig. 5. Distribution of requests to the remote model as boxplot without outliers. On the right, the barplot shows the median request count for each attack.

people across different backgrounds. Figure 6 shows the distribution of the surveyed population in terms of scientific/non-scientific background. The Ethical statement at the end of this paper discusses the population distribution and other potential concerns. To mitigate bias in perception scores arising from question structure implying malicious edits in certain images, we first uniformly sampled some benign images and then incorporated all corresponding adversarial examples from considered attacks. Each question of the survey showed a picture (either a clean image or an adversarial example) and asked to rate how much they thought the image may have been altered. Valid scores for the answers are integer numbers from 0 to 3, where 0: Not Altered, 1: Slightly Visible, 2: Visible but could fool some people, and 3: Very much visible.

All images for the study were generated with a perturbation budget of $L_\infty = 0.05$. Before showing any picture, the survey asked the participant if they knew what an "adversarial example" is. Figure 6 shows the distribution of people who know the concept in the S.T.E.M. (Science, Technology, Engineering, and Mathematics) and Non-S.T.E.M. groups. Unsurprisingly, the latter seems to have a much smaller percentage of people that are aware of adversarial examples. Comparing the average score given for each attack and to unaltered images, it is apparent that the ranking of the visibility of attacks is the same in both population groups. However, the population that did not know the concept of adversarial examples was observed to be more "paranoid". The general scoring for all images is higher for this group, even for unaltered images. Figure

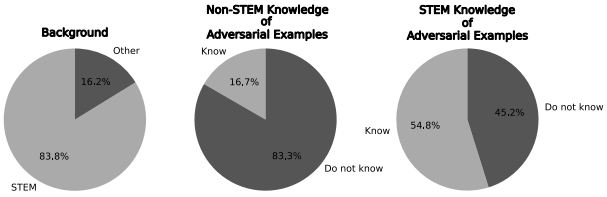


Fig. 6. Proportion of respondents from S.T.E.M. fields, awareness of adversarial examples within S.T.E.M. and Non-S.T.E.M. population.

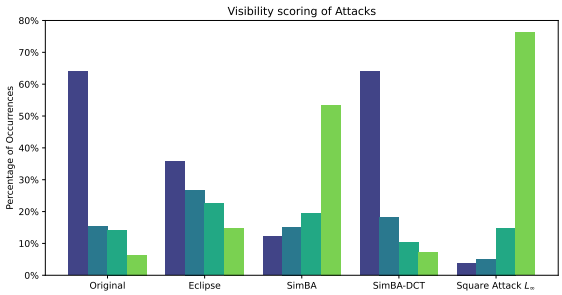
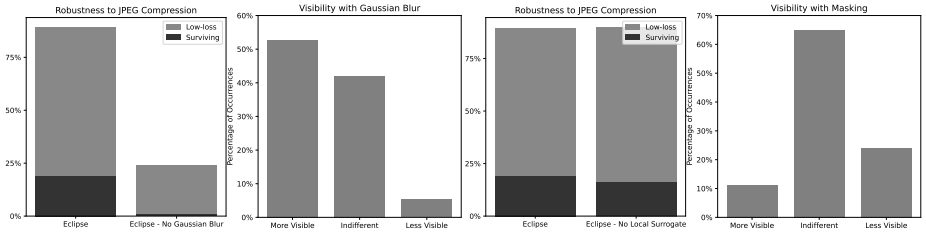


Fig. 7. Distribution of visibility scores for each attach on the general population. Scores are from 0 (not visible, blue) to 3 (very much visible, light green). The leftmost category corresponds to unaltered images.

7 shows a summary visualization of the scoring given by the general population for each attack and unaltered images. SimBA-DCT can produce very convincing adversarial examples, yielding a score distribution that is very similar to that of the original images. In second place we have ECLIPSE, which is slightly easier to recognize, but remains undetected by a large portion of survey subjects. If we consider both samples that are not recognized (score 0) and slightly visible ones (score 1) we observe a cumulative percentage of 62.55% for ECLIPSE and 82.34% for SimBA-DCT. SimBA and the Square Attack L_∞ follow with undeniably poor results: more than 50% of the survey participants consider adversarial examples from these attacks to be "very much visible". Their cumulative distribution for scores 0 and 1 is 27.23% and 8.94% respectively.

ECLIPSE Ablation Study. An ablation study was conducted to evaluate the impact of two novel components of the ECLIPSE attack: gradient Gaussian blurring and the sampling mask derived from the local surrogate. These components were assessed for their contribution to the three effectiveness properties. Robustness and detectability metrics mirror those used in attack comparisons, while visual stealthiness was evaluated via a survey. Participants were asked to compare ECLIPSE examples with ablated counterparts, choosing between three options: "The one on the left" (ECLIPSE), "The one on the right" (ablated version), or "I see no difference." Results illuminate the role of each component in the attack's overall effectiveness and stealth. Removing the Gaussian blur step significantly impacts all three effectiveness properties. Most notably, it improves robustness to processing, as evidenced by reduced loss and



(a) Gaussian blurring. (b) Local Surrogate.

Fig. 8. ECLIPSE ablation study results.

Table 4. Ablation study results of automatic detection of benign images and adversarial examples using spectral features. Each metric is a cross-validation score for Support Vector Machine classifiers. The are under the ROC curve is highlighted in bold.

Comparison	Accuracy	Precision	Recall	F1-score	**ROC AUC**
ECLIPSE	0.87 (± 0.01)	0.03 (± 0.20)	0.01 (± 0.08)	0.02 (± 0.11)	**0.50 (± 0.03)**
No Gaussian blur	0.90 (± 0.03)	0.69 (± 0.23)	0.39 (± 0.18)	0.50 (± 0.18)	**0.68 (± 0.09)**
No Local Surrogate	0.90 (± 0.04)	0.71 (± 0.21)	0.46 (± 0.28)	0.54 (± 0.22)	**0.71 (± 0.14)**

Table 5. Ablation study results. The table combines Robustness to Compression metrics and remote model query count.

	Compression Robustness			Remote Queries	
	Median Loss	Low-loss (%)	Surviving(%)	Median	IQR
ECLIPSE	0.15	89.33	18.67	8125	11212.50
No Gaussian blur	0.36	24.00	0.67	12870	14917.50
No Local Surrogate	0.17	90.00	16.00	9880	12382.50

higher survival rates (Figure 8a). Table 5 shows improved metrics, while stealthiness to automatic detection also benefits, with a 37% reduction in the median number of queries and a 25% decrease in the interquartile range (Table 5). Despite a drop in the Area Under the Curve of the Receiver Operating Characteristic (ROC AUC) from 0.68 to 0.5, indicating minimal automatic detection performance without the blur, the inclusion of Gaussian blur does not improve visual stealthiness, with over 50% of participants finding the adversarial examples more visible (Figure 8a). On the other hand, removing the local surrogate primarily benefits stealthiness, with a drop in ROC AUC from 0.71 to 0.5 (Table 5). It also reduces the number of queries by 18%, with a 10% reduction in the interquartile range (Table 5). However, it has minimal impact on robustness to processing or human visual stealth, with more than 60% of participants perceiving no difference (Figure 8b) and negligible effects on JPEG compression metrics, indicating that Gaussian blur is the key component affecting these properties.

Discussion. The results demonstrate that ECLIPSE achieves a superior balance across the three effectiveness properties. First, it generates the largest number of adversarial examples resilient to JPEG compression, substantially outperforming the second-best approach, the Square Attack. Second, ECLIPSE effectively bypasses

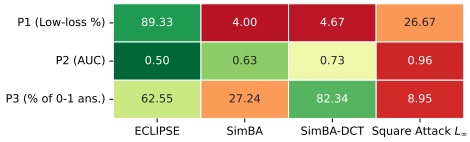


Fig. 9. Summary of evaluation of the main metrics of each effectiveness property.

defenses based on spectral features, as adversarial examples produced by the method are indistinguishable from benign images. Notably, a classifier trained for this distinction performs no better than random chance. While ECLIPSE does not set new benchmarks for query efficiency, its performance remains comparable to SimBA, one of the baseline methods. Furthermore, in terms of visual stealthiness, ECLIPSE achieves minimal degradation in human recognition rates relative to SimBA-DCT, the least detectable attack among those evaluated. A summary table of the main metrics that characterize the effectiveness properties for each attack is reported in Figure 9. To conclude, an ablation study of ECLIPSE's novel components confirms their critical contributions to its effectiveness across assessed properties. We shortly discuss the higher query count of ECLIPSE with respect to baselines. This is primarily due to the computational cost of gradient estimation via Hill Climbing, which exceeds that of heuristic baselines. However, as demonstrated in this section, this trade-off enhances real-world deployability. The increased query requirement imposes greater effort and cost on an attacker, necessitating multiple accounts and a more complex request distribution strategy. However, it does not drastically impact the real-world feasibility of the attack.

6 Conclusions

Adversarial examples present significant risks to machine learning systems, even in black-box scenarios where attackers rely on query-only access. While evasion attacks in computer vision are theoretically feasible, their practical deployment is constrained. We have formalized three *effectiveness properties* to measure the real-world feasibility of an adversarial attack: Robustness to Compression, Stealthiness to Automatic Detection, and Stealthiness to Human Inspection. We have further presented ECLIPSE, an attack that balances these properties through two novel components: Gaussian blurring of estimated gradients and gradient masking using heatmaps derived from surrogate models. ECLIPSE demonstrates superior robustness to JPEG compression, achieving adversarial success in 89% of cases compared to 27% for Square Attack L_∞. Against spectral-based detection, ECLIPSE achieves perfect stealthiness (AUC 0.5), significantly outperforming the Square Attack L_∞ (AUC 0.96). In terms of visibility, ECLIPSE ranks second, closely behind SimBA-DCT, with 63% of survey participants rating it negligible or invisible. Our research demonstrates the feasibility of adversarial attacks in real-world scenarios and highlights the necessity of developing defenses against more sophisticated threats. Through this evaluation, we demonstrated that existing State-of-the-Art attacks exhibit limited adherence to effectiveness properties, whereas ECLIPSE achieves a well-balanced trade-off.

Future Work and Limitations. While the work considers few baselines compared to the vast literature of adversarial examples, it considers the most meaningful State-of-the-Art attacks. Future work could extend robustness evaluations to include additional image transformations and defensive measures, such as physical deployment through printed adversarial examples [39]. Moreover, exploring perturbations beyond additive noise, such as changes to brightness, contrast, or color dynamics, could enhance stealthiness under diverse conditions. A broader study incorporating a wider range of attacks would further refine the analysis of effectiveness properties and their trade-offs.

Acknowledgments.. This work was partially supported by the Google.org Impact Challenge - Tech for Social Good Research Grant (Tides Foundation) and project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Disclosure of Interests.. The authors have no competing interests to declare that are relevant to the content of this article.

Ethical Statement. The survey on visual stealthiness did not collect any personal or identifiable information, including age, ensuring complete anonymity. Participation was voluntary and no incentive was given to participate. Even though no age-specific data was collected, we infer the survey demographic to be mostly under 30, given the channels where the survey was spread (student mailing list and social media).

References

1. api4ai Object Detection Endpoint (2024)
2. Clarifai Object Detection Endpoint (2024)
3. Ahmed Aldahdooh, W. Hamidouche, Sid Ahmed Fezza, O. Déforges: Adversarial example detection for DNN models: a review and experimental comparison. Artificial Intelligence Review (2021). https://doi.org/10.1007/s10462-021-10125-w
4. Al-Dujaili, A., O'Reilly, U.M.: Sign Bits Are All You Need for Black-Box Attacks. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SygW0TEFwH
5. Andrew Starnes, Clayton Webster: Gaussian smoothing stochastic gradient descent (GSmoothSGD) (2023), aRXIV_ID: 2311.00531
6. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. European Conference on Computer Vision (2020)
7. Balasubramaniam, A., Pasricha, S.: Object detection in autonomous vehicles: Status and open challenges. arXiv preprint arXiv:2201.07706 (2022)
8. Brendel, W., Rauber, J., Bethge, M.: Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. International Conference on Learning Representations (Feb 2018)
9. Byun, J., Go, H., Kim, C.: 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 (WACV). pp. 3051–3060 (Jan 2022)
10. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57. IEEE Computer Society, Los Alamitos, CA, USA (May 2017). https://doi.org/10.1109/SP.2017.49

11. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: 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. p. 15–26. AISec '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3128572.3140448
12. Chen, S.L., Chen, S.W., Carlini, N., Wagner, D.: Stateful Detection of Black-Box Adversarial Attacks. arXiv: Cryptography and Security (Jul 2019). https://doi.org/10.1145/3385003.3410925
13. Choi, S.H., Shin, J., Choi, Y.H.: PIHA: Detection Method Using Perceptual Image Hashing against Query-Based Adversarial Attacks. Future Gener. Comput. Syst. **145**(C), 563–577 (Aug 2023). https://doi.org/10.1016/j.future.2023.04.005, https://doi.org/10.1016/j.future.2023.04.005, place: NLD Publisher: Elsevier Science Publishers B. V.
14. Corrado, A.: Animals-10, https://www.kaggle.com/datasets/alessiocorrado99/animals10
15. Croce, F., Andriushchenko, M., Singh, N.D., Flammarion, N., Hein, M.: Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 6437–6445 (2022), issue: 6
16. Dang, H., Huang, Y., Chang, E.C.: Evading Classifiers by Morphing in the Dark. Conference on Computer and Communications Security (May 2017). https://doi.org/10.1145/3133956.3133978
17. Esmaeili, B., Azmoodeh, A., Dehghantanha, A., Karimipour, H., Zolfaghari, B., Hammoudeh, M.: IIoT Deep Malware Threat Hunting: From Adversarial Example Detection to Adversarial Scenario Detection. IEEE Transactions on Industrial Informatics **18**(12), 8477–8486 (2022). https://doi.org/10.1109/TII.2022.3167672
18. Giulivi, L., Jere, M., Rossi, L., Koushanfar, F., Ciocarlie, G., Hitaj, B., Boracchi, G.: Adversarial scratches: Deployable attacks to cnn classifiers. Pattern Recognition **133**, 108985 (2023)
19. Gonzalez, R.C., Woods, R.E.: Digital Image Processing (3rd Edition). Prentice-Hall, Inc., USA (2006)
20. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. International Conference of Learning Representations (Dec 2015), aRXIV_ID: 1412.6572
21. Guo, C., Gardner, J.R., You, Y., Wilson, A.G., Weinberger, K.Q.: Simple Black-box Adversarial Attacks. International Conference on Machine Learning pp. 2484–2493 (Jan 2019)
22. Hanneke, S.: Rates of convergence in active learning. Annals of Statistics **39**, 333–361 (2011). https://doi.org/10.1214/10-AOS843
23. Inkawhich, N., Liang, K.J., Binghui Wang, Binghui Wang, Wang, B., Inkawhich, M., Carin, L., Yiran Chen, Chen, Y., Chen, Y.: Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability. arXiv: Cryptography and Security (Apr 2020)
24. Jianbo Chen, Chen, J., Michael I. Jordan, Jordan, M.I., Michael I. Jordan, Martin J. Wainwright, Wainwright, M.J.: HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. IEEE Symposium on Security and Privacy pp. 1277–1294 (May 2020). https://doi.org/10.1109/sp40000.2020.00045
25. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv: Computer Vision and Pattern Recognition (Jul 2016). https://doi.org/10.1201/9781351251389-8
26. Li, H., Xu, X., Zhang, X., Yang, S., Li, B.: Qeba: Query-efficient boundary-based blackbox attack. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1218–1227. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00130

27. Li, H., Shan, S., Wenger, E., Zhang, J., Zheng, H., Zhao, B.Y.: Blacklight: Scalable Defense for Neural Networks against Query-Based Black-Box Attacks. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 2117–2134. USENIX Association, Boston, MA (Aug 2022), https://www.usenix.org/conference/usenixsecurity22/presentation/li-huiying
28. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into Transferable Adversarial Examples and Black-box Attacks. International Conference on Learning Representations (Nov 2016)
29. Liu, Z., Li, F., Lin, J., Li, Z., Luo, B.: Hide and Seek: on the Stealthiness of Attacks against Deep Learning Systems. In: European Symposium on Research in Computer Security. pp. 343–363. Springer (2022)
30. Luke, S.: Essentials of Metaheuristics. Lulu, second edn.
31. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
32. Muzammal Naseer, Naseer, M., Salman Khan, Khan, S., Salman Khan, Salman Khan, Salman Khan, Khan, S., Khan, S.A., Fatih Porikli, Porikli, F., Fatih Porikli: Local Gradients Smoothing: Defense Against Localized Adversarial Attacks. IEEE Workshop/Winter Conference on Applications of Computer Vision pp. 1300–1307 (Jan 2019). https://doi.org/10.1109/wacv.2019.00143
33. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, _eprint: 1605.07277
34. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical Black-Box Attacks against Machine Learning. ACM Asia Conference on Computer and Communications Security pp. 506–519 (Apr 2017). https://doi.org/10.1145/3052973.3053009
35. Ryan Feng, Ashish Hooda, Neal Mangaokar, Kassem Fawaz, S. Jha, Atul Prakash: Stateful Defenses for Machine Learning Models Are Not Yet Secure Against Black-box Attacks. Conference on Computer and Communications Security (2023). https://doi.org/10.1145/3576915.3623116
36. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626
37. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. International Conference on Learning Representations (Jan 2014)
38. Thibault Maho, Maho, T., Teddy Furon, Furon, T., Erwan Le Merrer, Le Merrer, E.: SurFree: a fast surrogate-free black-box attack. arXiv: Cryptography and Security (2020). https://doi.org/10.1109/cvpr46437.2021.01029
39. Wei, X., Guo, Y., Li, B.: Black-box adversarial attacks by manipulating image attributes. Information Sciences **550**, 285–296 (2021). https://doi.org/https://doi.org/10.1016/j.ins.2020.10.028