Understanding Black-Box Attacks Against Object Detectors From a User's Perspective

Kim André Midtlid, Johannes Åsheim, and Jingyue Li[0000−−0002−7958−391X]

Norwegian University of Science and Technology, Trondheim, Norway
kamidtli@stud.ntnu.no, johannes.asheim@ntnu.no, jingyue.li@ntnu.no

Abstract. Due to recent developments in object detection systems, and the realistic threat of black-box adversarial attacks on object detector models, we argue the need for a contextual understanding of the attacks from the users' perspective. Existing literature reviews either do not provide complete and up-to-date summaries of such attacks or focus on the knowledge from the researchers' perspective. In this research, we conducted a systematic literature review to identify state-of-the-art black-box attacks and extract the information to help users evaluate and mitigate the risks. The literature review resulted in 29 black-box attack methods. We analyzed each attack from the following main aspects: attackers' knowledge needed to perform the attack, attack consequences, attack generalizability, and strategies to mitigate the attacks. Our results demonstrate an emerging increase in highly generalizable attacks, which now make up more than 50% of the landscape. We also reveal that more than 50% of recent attacks remain untested against mitigation strategies.

Keywords: artificial intelligence · object detection · image classification · adversarial attacks

1 Introduction

As Deep Neural Networks (DNNs) becomes more and more pertinent in image recognition and object detection tasks, their robustness also becomes more of a concern. Goodfellow et al. [14] have shown that the robustness of these models is susceptible to adversarial attacks. Such vulnerabilities have motivated researchers to develop adversarial attacks to exploit the object detection systems and contribute to improving their robustness. White-box attacks that assume knowledge about the target model continue to dominate the adversarial attack landscape, but there is an increase in black-box attacks. Black-box attacks assume no or very limited knowledge about the target model and are, therefore, more realistic approaches to adversarial attacks [34]. We argue that the increase in black-box attacks should be followed by a contextual understanding of the attacks from a user perspective. We define a user as a person who wants to know the risk and impact of adversarial attacks and how to defend against these attacks without knowing specific attack implementation details. Therefore, this paper omit the technical properties of the attacks for the traditional researcher

perspective. Existing surveys and reviews of adversarial attacks on image classification and object detection, e.g., [6, 20], focus mostly on the information needed by researchers and do not cover sufficient up-to-date black-box attacks. Our research motivation is to summarize the state-of-the-art black-box attacks targeting object detection models to help users evaluate and mitigate the risks. We focus on answering the following research questions.

- **RQ1:** What does the attacker need to know about the target model?
- **RQ2:** How generalizable is the attack?
- **RQ3:** What are the consequences of the attack?
- **RQ4:** Which mitigation strategies have been tested against the attack?

We performed a systematic literature review on articles published between 2017 and 2021 to collect state-of-the-art black-box attacks. Through the systematic literature review and snowballing, we uncovered 29 state-of-the-art attack methods, which we analyze and present in this paper. Our study benefits industrial practitioners and scientists. The contributions of the study are twofold.

- We provide comprehensive and up-to-date consolidated knowledge about black-box attacks targeting object detection models to help users to evaluate the risks and choose effective mitigation solutions.
- We identify the trends and weaknesses of existing studies in this field, which may inspire researchers' future work.

The rest of the paper is organized as follows: Section 2 introduces the background. Section 3 presents the related work. Section 4 explains our research methods, and Section 5 presents the results. We then discuss our results in Section 6. Conclusions and future work are in Section 7.

2 Background

Object detection is the field of Artificial Intelligence (AI) that uses deep learning to extract high-dimensional information from images and videos. An autonomous car with camera sensors uses image processing to navigate the road and detect obstacles.

2.1 Object Detection and Image Classification

Image classification is the task of classifying an input image by assigning it to a specific label [42], while object detection is the task of localizing and classifying distinct objects in an image or video. Current object detectors can be split into two main categories: two-stage and one-stage detectors. Two-stage detectors consist of two main parts. First, the detector uses a Region Proposal Network (RPN) to calculate proposed regions for objects. The RPN uses a set of predefined *anchor boxes* uniformly placed over the image to calculate proposed regions before outputting a predefined number of proposed bounding boxes with a corresponding objectiveness score. The objectiveness score indicates whether the proposed

region belongs to an object class or the background. These proposed regions significantly reduce the computational complexity needed to localize and classify an object. In the second stage, the proposed regions from the RPN are passed to a high-quality image classifier to recognize objects. One-stage detectors aim to improve the inference speed while still achieving acceptable accuracy. One-stage detectors achieve this goal by removing the region proposal stage required by the two-stage detectors. Instead, they run detection on a dense sampling of predefined default boxes. The ability to skip the region proposal step significantly decreases inference time and has led to the development of many one-stage detectors, e.g., [30, 38].

2.2 Threat Models

The threat model of an attack is based on what the adversary knows about the target model, thus we can categorize the attacks into three threat models. *White-box attacks*, e.g., FGSM [14], assume the adversary has complete knowledge of the target model , which include the model's internal structure, such as weights and parameters of the target model, and knowledge of the output given an input. In some cases, the adversary knows the training data distribution. This allows the adversary to construct attack methods specific to the given model. *Black-box attacks*, assume no internal information of the target model, but the ability to observe the output for a given input. Usually, black-box attack methods are constructed based on querying the target model [5, 8, 9]. Han Xu et al. [46] introduce *grey-box attacks* as a hybrid of white-box attacks and black-box attacks, where the attacker trains a generative model to create adversarial examples in white-box setting. Then the target model is attacked in the black-box setting with adversarial examples from the trained generative model.

3 Related Work

Bhambri et al. [6] performed a survey focusing on adversarial black-box attacks. The paper aims to conduct a comparative study of both adversarial attacks and defenses. Nineteen black-box attacks were compared on the number of queries, success rate, and perturbation norm. The survey categorizes the attacks based on gradient estimation, transferability, local search and combinatorics. Shilin Qiu et al. [37] presents a comprehensive study of the research of adversarial attack and defenses. The paper details white-box and black-box attack methods but mainly focuses on defense strategies. Kong et al. [25] reviewed adversarial attack literature in the different application fields of AI security. The fields include images, texts and malicious code. The paper presents attack algorithms for the different application domains and includes 13 attacks for the image domain, five of which are black-box attacks. The survey further elaborates on defense methods and how they affect the presented attacks. In order to help new researchers in the field, the paper introduces and discusses the different datasets and tools available. There are other surveys and articles, i.e., [1, 27, 46, 48], which discuss

adversarial attacks and defenses. The common limitation of these studies are the low number of included black-box attacks. In addition, the studies focus on consolidating information from the researchers' perspective.

4 Research Design and Implementation

We performed a Systematic Literature Review (SLR) and followed the SLR guidelines proposed by Kitchenham and Charters [24]. After analyzing the terms related to our research questions and their synonyms, we chose to use the search query: *Adversarial* AND *Attack* AND (*"Object detection"* OR *"Object detector"*).

We chose `oria.no`, a search engine that covers many scientific databases, including IEEE Xplore, Springer, ACM Digital library, and Scopus. To include only recent literature and to reduce the scope, we used the advanced search functionality in `oria.no`, and included only peer-reviewed and published scientific papers from the last 5 years back from 2021. The identified articles were filtered mainly based on their relevance to the research questions by reading their abstract, introduction, and, in some cases, methodology. After filtering, we identified 11 relevant primary studies. Then, we performed a snowballing search following the process proposed by [45], with the exception that forward and backward snowballing searches were limited to a single iteration each. The forward snowballing was performed using Google Scholar. The snowballing identified 16 more papers, resulting in 27 primary studies.

5 Research Results

In this section, we present our answers to each research question. Attack names preceded by asterisks (*) were not presented with a name in their corresponding paper. Therefore, a descriptive name is given based on the attack method.

5.1 RQ1—Attacker's Knowledge

How much information the attacker requires from the output labels varies across the identified papers but can be split into three categories: **Soft-labels** refer to the threat model where an attacker accesses the output probabilities $P(y|x)$ for y in the top k classes. Soft-labels also might include the label for each of the output probabilities. For object detectors, information about the bounding boxes indicates soft-labels. **Hard-labels** refer to a more restricted threat model where an attacker only has access to a list of $k \in \mathbb{Z}^+$ output labels. Different attacks make different assumptions about k. For $k = 1$, the attacker only has access to the single predicted class. In the case of $k > 1$, the list of classes is often ordered by decreasing probabilities but does not include the probabilities. For object detectors, the hard-label category signifies no information about the bounding boxes. Some attacks assume the target model outputs $k = 1$ or $k > 1$

Table 1: Attacks grouped by attacker knowledge

Attack Name	**Year**	**Knowledge**
NRDM [33]	2018	No-labels
DaST [51]	2020	Hard-labels and Soft-labels
HopSkipJumpAttack [9]	2020	Hard-labels
*Partial-retraining [36]	2020	Hard-labels
*Evolutionary Attack [13]	2019	Hard-labels
Label-Only Attack [20]	2018	Hard-labels
Opt-Attack [11]	2018	Hard-labels
Boundary Attack [8]	2017	Hard-labels
CMA-ES [19]	2021	Soft-labels
Simple Transparent Adversarial Examples [7]	2021	Soft-labels
*Discrete Cosine Transform Attack [26]	2021	Soft-labels
*Differential Evolution Attack [44]	2021	Soft-labels
BMI-FGSM [29]	2020	Soft-labels
*Transferable Universal Perturbation Attack [49]	2020	Soft-labels
Adv-watermark [23]	2020	Soft-labels
Evaporate Attack [43]	2020	Soft-labels
Daedalus [41]	2019	Soft-labels
One-Pixel-Attack [39]	2019	Soft-labels
Single Scratch attack [22]	2019	Soft-labels
GenAttack [2]	2019	Soft-labels
Universal perturbation attack [50]	2019	Soft-labels
Query-Limited Attack [20]	2018	Soft-labels
Partial-Info Attack [20]	2018	Soft-labels
Bandits [21]	2018	Soft-labels
Gradient Estimation Attacks [5]	2018	Soft-labels
R-AP [28]	2018	Soft-labels
ZOO [10]	2017	Soft-labels
LocSearchAdv [32]	2016	Soft-labels
*Substitute Attack [34]	2016	Soft-labels

labels. **No-labels** refer to the most restricted threat model, where an attacker requires no access to the output of the target model.

Table 1 presents the attacks grouped by the required attacker knowledge. We notice that more than 75% of the discussed attacks use the soft-labels approach. Table 1 also illustrates that about 25% of the discussed attacks use hard-labels as part of their method. We can also see that the number of hard-label attacks has tripled from 2017 to 2020, which might indicate that hard-label attacks are becoming more popular. The new trend might suggest that hard-label attacks have room for improvement in the coming years and should be investigated further. It is also worth noting **DaST** [51], which can be used in both a soft- and hard-label scenario because the attack is customizable. This might be an

indication of a new type of attack that can be modified based on the target model. **NRDM** [33] requires no labels at all. These two attacks illustrate a possibility in the landscape, as attacks can become more applicable to any target model and more independent of the attacker's knowledge.

5.2 RQ2—Attack Generalizability

The generalization of adversarial black-box attacks examines the number of different types of object detection models which are claimed to have been successfully attacked. We have defined four categories of generalization and present the results in Table 2. The categories are *None*, *Low*, *High* and *Very High*. The presented attack is tested on and successful against either one, two, three to five or six or more target models respectively. The term generalizability is only determined based on the number of attacked target models, and do not include datasets, model accuracy, attack hyperparameters and model hyperparameters. It is important to note that the generalizability is derived from the number of models claimed by the authors of the primary studies. Therefore an attack with *None* may be generalizable, but the authors only includes experiments against one target model.

Most of the attacks only target image classifiers, but the focus could be on one-stage models, two-stage models, or a combination of both for object detectors. An attack targeting both types of object detectors poses a significant threat, as it generalizes to most model architectures. This aspect is captured in the *target architecture* column in Table 2. Attacks targeting object detectors are labeled with one-stage, two-stage, or both, while attacks targeting image classifiers are labeled correspondingly.

From Table 2, we observe a balanced distribution between high and low generalizability. Both attack types show promising results, but the ones with high generalizability might be more interesting to be studied further, as they are successful across a broader range of object detectors. The number of highly generalizable attacks has increased from 2019, as shown in Figure 1. From Table 2, we also notice that **R-AP** [28] and **NRDM** [33] stand out. They are both classified as very high, meaning they have been tested and exhibited promising performance on six or more different models. Additionally, **NRDM** has been tested against both image classifiers and object detectors, demonstrating notable generalizability . It is also worth noting that [28] and [9] mention the possibility of combining **R-AP** and **HopSkipJumpAttack**, respectively, with other adversarial attacks as areas for future work. This combination demonstrates a potential to improve attacks through amalgamation, which is worth considering in future research. Many of the discussed attacks have also been tested on real-world APIs, which are listed in Table 3. From a user perspective, this illustrates a potential area of focus and risks to consider in the future.

Table 2: Attacks grouped by their level of generalizability

Attack Name	**Year**	**Generalization**	**Target Architecture**
NRDM [33]	2018	Very High	Image classifiers
R-AP [28]	2018	Very High	Two-stage
CMA-ES [19]	2021	High	One-stage and two-stage
*Differential Evolution Attack [44]	2021	High	Image classifiers
Adv-watermark [23]	2020	High	Image classifiers
Evaporate Attack [43]	2020	High	One-stage and two-stage
HopSkipJumpAttack [9]	2020	High	Image classifiers
*Partial-retraining [36]	2020	High	Image classifiers
*Transferable Universal Perturbation Attack [49]	2020	High	One-stage and two-stage
Daedalus [41]	2019	High	One-stage
One-Pixel-Attack [39]	2019	High	Image classifiers
Universal perturbation attack [50]	2019	High	Image classifiers
Single Scratch attack [22]	2019	High	Image classifiers
Bandits [21]	2018	High	Image classifiers
Gradient Estimation Attacks [5]	2018	High	Image classifiers
Boundary Attack [8]	2017	High	Image classifiers
*Substitute Attack [34]	2016	High	Image classifiers
*Discrete Cosine Transform Attack [26]	2021	Low	Image classifiers
BMI-FGSM [29]	2020	Low	Image classifiers
DaST [51]	2020	Low	Image classifiers
*Evolutionary Attack [13]	2019	Low	Image classifiers
GenAttack [2]	2019	Low	Image classifiers
Opt-Attack [11]	2018	Low	Image classifiers
Query-Limited Attack [20]	2018	Low	Image classifiers
Partial-Info Attack [20]	2018	Low	Image classifiers
Label-Only Attack [20]	2018	Low	Image classifiers
LocSearchAdv [32]	2016	Low	Image classifiers
Simple Transparent Adversarial Examples [7]	2021	None	Image classifiers
ZOO [10]	2017	None	Image classifiers

Table 3: Attacks against real-world APIs

Attack Name	**Year**	**Real-World API**
*Discrete Cosine Transform Attack [26]	2021	AWS Rekognition [4]
*Partial retraining [36]	2020	Google AutoML Vision [15]
Partial-Info Attack [20]	2018	Google Cloud Vision [16]
Gradient Estimation Attacks [5]	2018	Clarifai [12]
Boundary Attack [8]	2017	Clarifai [12]
*Substitute Attack [34]	2016	Amazon and Google Oracles [3, 16]

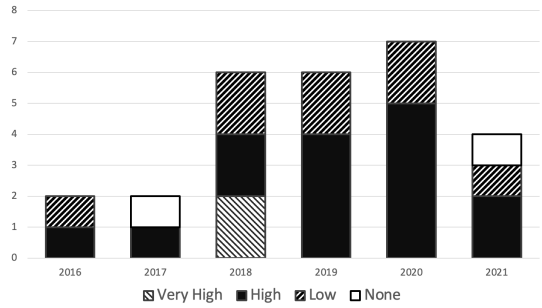

Fig. 1: The ratio of generalization levels for each year

5.3 RQ3—Attack Consequences

Classification attack is divided into targeted and untargeted attacks. Targeted attacks aim to misclassify a adversarial input image i' of class c', where the the target model would have classified input image i in to class c. In other words, the attacker wants to force the target model to predict a chosen class. Untargeted attacks aim to misclassify an adversarial input image i' in to any class c', where $c' \neq c$. Object detection attack can lead to object vanishing and object population. An object vanishing attack aims to suppress all object detection in a input image, while an object population attack aims to fabricate false objects in a predicted image.

Table 4 shows the consequences of each attack. Untargeted attacks are the most common, making up more than 75% of the discussed attacks. Even though these attacks make up the majority and pose a significant threat, targeted attacks might be more dangerous from a defender's perspective. Targeted attacks still make up about 65% of discussed attacks, and it is worth noting that most image classification attacks provide both targeted and untargeted versions. This trend suggests that attacks are not limited to a single purpose but can achieve multiple goals. In the realm of object detection attacks, we have looked at five attacks. Four of them exploit the object vanishing vulnerability, while only one focuses on object population. **CMA-ES** [19] stands out because it combines object detection and image classification attacks. **CMA-ES** is a very recently developed attack that could hint at a change of focus in the landscape. Additionally, **Daedalus** [41] is the only attack that can execute object population. Results in Table 4 also shows the emerging focus on attacks against object detectors from 2018.

5.4 RQ4—Mitigation Strategies

Table 5 contains a summary of all the mitigation strategies an attack is claimed to have been tested against. The *Vulnerable Mitigations* column lists all tested mitigation strategies where the attack is still able to reduce the overall accuracy of the system significantly. The definition of a significant drop in accuracy is claimed by each paper. The *Robust Mitigations* column lists all mitigation

Table 4: Attacks grouped by their consequences

Attack Name	**Year**	**Target Architecture**	**Consequenses**
CMA-ES [19]	2021	One-stage and two-stage	Vanishing, Targeted, and Untargeted
Evaporate Attack [43]	2020	One-stage and two-stage	Vanishing
*Transferable Universal Perturbation Attack [49]	2020	One-stage and two-stage	Vanishing
R-AP [28]	2018	Two-stage	Vanishing
Daedalus [41]	2019	One-stage	Population
*Differential Evolution Attack [44]	2021	Image classifiers	Targeted and Untargeted
BMI-FGSM [29]	2020	Image classifiers	Targeted and Untargeted
DaST [51]	2020	Image classifiers	Targeted and Untargeted
HopSkipJumpAttack [9]	2020	Image classifiers	Targeted and Untargeted
One-Pixel-Attack [39]	2019	Image classifiers	Targeted and Untargeted
Single Scratch attack [22]	2019	Image classifiers	Targeted and Untargeted
Gradient Estimation Attacks [5]	2018	Image classifiers	Targeted and Untargeted
Query-Limited Attack [20]	2018	Image classifiers	Targeted and Untargeted
Partial-Info Attack [20]	2018	Image classifiers	Targeted and Untargeted
Label-Only Attack [20]	2018	Image classifiers	Targeted and Untargeted
Bandits [21]	2018	Image classifiers	Targeted and Untargeted
Opt-Attack [11]	2018	Image classifiers	Targeted and Untargeted
Boundary Attack [8]	2017	Image classifiers	Targeted and Untargeted
ZOO [10]	2017	Image classifiers	Targeted and Untargeted
LocSearchAdv [32]	2016	Image classifiers	Targeted and Untargeted
*Discrete Cosine Transform Attack [26]	2021	Image classifiers	Targeted
*Partial-retraining [36]	2020	Image classifiers	Targeted
GenAttack [2]	2019	Image classifiers	Targeted
Simple Transparent Adversarial Examples [7]	2021	Image classifiers	Untargeted
Adv-watermark [23]	2020	Image classifiers	Untargeted
*Evolutionary Attack [13]	2019	Image classifiers	Untargeted
Universal perturbation attack [50]	2019	Image classifiers	Untargeted
NRDM [33]	2018	Image classifiers	Untargeted
*Substitute Attack [34]	2016	Image classifiers	Untargeted

strategies where the attack cannot reduce the overall accuracy of the system significantly. It is worth noting that *None tested* in the *Robust Mitigations* column only means that the attack has not been tested on any mitigation strategy. It does not mean that the attack is able to bypass all defense strategies. This also applies to the *Vulnerable Mitigations* column. A cell with "-" means that none of the tested mitigation strategies applies to that column. A list of defenses in the *Vulnerable Mitigations* column and "-" in the *Robust Mitigations* column means that none of the tested defenses successfully defended against the attack.

From Table 5, we notice that more than half of the discussed attacks have not been tested against any mitigation strategies. This illustrates that mitigation strategies have not been given enough attention. We also notice that Adversarial Training and Input Transformations repeat across different attacks in the *Vulnerable Mitigations* column. The repetition indicates that no single mitigation strategy works for all attacks, and that most modern mitigation strategies struggle to defend against the discussed attacks. It is worth noting that many of the mitigation strategies listed are umbrella terms, covering multiple defense implementations. For example, input transformations [18] cover multiple defense mechanisms such as JPEG-compression, clipping and median filtering. Although Figure 2 shows an increase in the number of mitigation strategies evaluated, we can also see a large emerging ratio of untested attacks from 2018.

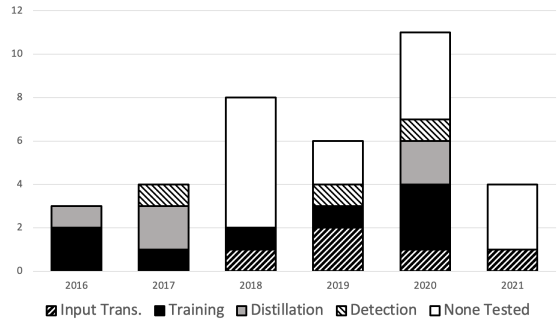

Fig. 2: The ratio of mitigation strategies each year

6 Discussion

The aim of our work is to summarize the state-of-the-art black-box attacks targeting object detectors to help users evaluate and mitigate the risks. No related work outlined in Section 3 takes the user's perspective but rather explains black-box attacks from a researcher's perspective and focuses on explaining the attack methods. For example, Kong et al. [25] and Bhambri et al. [6] provide categories of black-box attacks, but the categorization is based on the attack method. Understanding a black-box attack method requires a high level of competence in a user. Our study does not focus on the attack methods because they are not the most relevant information for a user. The main focuses from a user perspective are covered in our research questions. Results of RQ1 (Knowledge) can inform a user of the attacks which can and cannot be executed on a system. Results of RQ2 (Generalization) warns the user of which attacks have a large impact area and could affect the system. Results of RQ3 (Consequences) give the user insight into the attacks' results. Results of RQ4 (Mitigation strategies) are highly important to the user because they contain information that can help the user implement relevant defenses to the system.

Table 5: Attacks grouped by mitigation strategies they have been tested against

Attack	**Year**	**Vulnerable Mitigations**	**Robust Mitigations**
*Differential Evolution Attack [44]	2021	Feature squeezing [47] Input Transformations [18]	-
Adv-watermark [23]	2020	Adversarial Training [40] Input Transformations [18]	-
HopSkipJumpAttack [9]	2020	Adversarial Distillation [35], Region-based classification	Adversarial Training [40]
*Partial-retraining [36]	2020	Adversarial Detection [17] Adversarial Distillation [35] Adversarial Training [40] Feature squeezing [47]	-
GenAttack [2]	2019	Adversarial Training [40], Input Transformations [18]	-
One-Pixel-Attack [39]	2019	-	Adversarial Detection [17]
Daedalus [41]	2019	MagNet [31] Minimize bounding box size	-
Single Scratch attack [22]	2019	Input Transformations (JPEG-compression) [18] Input Transformations (Clipping) [18]	Input Transformations (Median Filtering) [18]
Gradient Estimation Attacks [5]	2018	Adversarial Training [40]	Rounded output probabilities
NRDM [33]	2018	Input Transformations [18]	-
Boundary Attack [8]	2017	Adversarial Distillation [35]	-
ZOO [10]	2017	Adversarial Detection [17] Adversarial Distillation [35]	Adversarial Training [40]
LocSearchAdv [32]	2016	Adversarial Training [40]	Query-access prevention
*Substitute Attack [34]	2016	Adversarial Distillation [35] Adversarial Training [40]	-
CMA-ES [19]	2021	None tested	None tested
*Discrete Cosine Transform Attack [26]	2021	None tested	None tested
Simple Transparent Adversarial Examples [7]	2021	None tested	None tested
DaST [51]	2020	None tested	None tested
Evaporate Attack [43]	2020	None tested	None tested
BMI-FGSM [29]	2020	None tested	None tested
*Transferable Universal Perturbation Attack [49]	2020	None tested	None tested
*Evolutionary Attack [13]	2019	None tested	None tested
Universal perturbation attack [50]	2019	None tested	None tested
Bandits [21]	2018	None tested	None tested
Label-Only Attack [20]	2018	None tested	None tested
Opt-Attack [11]	2018	None tested	None tested
R-AP [28]	2018	None tested	None tested
Query-Limited Attack [20]	2018	None tested	None tested
Partial-Info Attack [20]	2018	None tested	None tested

The results of the survey show that many modern adversarial attack studies have not focused on testing mitigation strategies, as shown in Table 5. Eighty percent of the discussed attacks against object detectors have not been tested against any mitigation strategies. Our study shows that the generalizability of recent attacks is increasing, which poses a more significant threat to the industry. No longer do the attacks focus on a single objective or target model, but rather, they combine all these goals into broader attacks. This means that modern attacks can bypass more defenses and achieve multiple attack objectives.

7 Conclusion and Future Work

We conducted a systematic literature review in order to summarize state-of-the-art black-box attacks targeting object detection models to help users evaluate and mitigate the risks. The literature review resulted in 29 unique black-box attack methods from 27 papers. Our analyses summarized the status and trends regarding attackers' knowledge needed to perform the attack, consequences, generalizability, and current mitigation strategies for each attack. We acknowledge that the SLR may have left out some papers due to missing search queries and limited database coverage. One finding from our study is that mitigation strategies should be comprehensively tested on the identified black-box attacks to find out which defenses are robust and which could be improved. We plan to focus on evaluating and improving different mitigation strategies as our future work.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access **6**, 14410–14430 (2018). https://doi.org/10.1109/ACCESS.2018.2807385
2. Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.J., Srivastava, M.: Genattack: Practical black-box attacks with gradient-free optimization (2018). https://doi.org/10.48550/ARXIV.1805.11090, `https://arxiv.org/abs/1805.11090`
3. Amazon: Aws machine learning (2021), `https://aws.amazon.com/machine-learning`
4. Amazon: Aws rekognition (2021), `https://aws.amazon.com/rekognition/`
5. Bhagoji, A.N., He, W., Li, B., Song, D.: Practical black-box attacks on deep neural networks using efficient query mechanisms. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 158–174. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_10
6. Bhambri, S., Muku, S., Tulasi, A., Buduru, A.B.: A survey of black-box adversarial attacks on computer vision models (2019). https://doi.org/10.48550/ARXIV.1912.01667, `https://arxiv.org/abs/1912.01667`
7. Borkar, J., Chen, P.Y.: Simple transparent adversarial examples (2021). https://doi.org/10.48550/ARXIV.2105.09685, `https://arxiv.org/abs/2105.09685`

8. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models (2017). https://doi.org/10.48550/ARXIV.1712.04248, https://arxiv.org/abs/1712.04248
9. Chen, J., Jordan, M.I., Wainwright, M.J.: Hopskipjumpattack: A query-efficient decision-based attack. In: 2020 IEEE Symposium on Security and Privacy (SP). pp. 1277–1294 (2020). https://doi.org/10.1109/SP40000.2020.00045
10. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: ZOO. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM (nov 2017). https://doi.org/10.1145/3128572.3140448, https://doi.org/10.1145%2F3128572.3140448
11. Cheng, M., Le, T., Chen, P.Y., Yi, J., Zhang, H., Hsieh, C.J.: Query-efficient hard-label black-box attack:an optimization-based approach (2018). https://doi.org/10.48550/ARXIV.1807.04457, https://arxiv.org/abs/1807.04457
12. Clarifai: The world's ai (2021), https://www.clarifai.com/
13. Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., Zhu, J.: Efficient decision-based black-box adversarial attacks on face recognition (2019). https://doi.org/10.48550/ARXIV.1904.04433, https://arxiv.org/abs/1904.04433
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014). https://doi.org/10.48550/ARXIV.1412.6572, https://arxiv.org/abs/1412.6572
15. Google: Automl (2021), https://cloud.google.com/automl
16. Google: Vision ai (2021), https://cloud.google.com/vision
17. Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.: On the (statistical) detection of adversarial examples (2017). https://doi.org/10.48550/ARXIV.1702.06280, https://arxiv.org/abs/1702.06280
18. Guo, C., Rana, M., Cisse, M., van der Maaten, L.: Countering adversarial images using input transformations (2017). https://doi.org/10.48550/ARXIV.1711.00117, https://arxiv.org/abs/1711.00117
19. Haoran, L., Yu'an, T., Yuan, X., Yajie, W., Jingfeng, X.: A cma-es-based adversarial attack against black-box object detectors. Chinese Journal of Electronics **30**(3), 406–412 (2021). https://doi.org/https://doi.org/10.1049/cje.2021.03.003, https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cje.2021.03.003
20. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information (2018). https://doi.org/10.48550/ARXIV.1804.08598, https://arxiv.org/abs/1804.08598
21. Ilyas, A., Engstrom, L., Madry, A.: Prior convictions: Black-box adversarial attacks with bandits and priors (2018). https://doi.org/10.48550/ARXIV.1807.07978, https://arxiv.org/abs/1807.07978
22. Jere, M., Rossi, L., Hitaj, B., Ciocarlie, G., Boracchi, G., Koushanfar, F.: Scratch that! an evolution-based adversarial attack against neural networks (2019). https://doi.org/10.48550/ARXIV.1912.02316, https://arxiv.org/abs/1912.02316
23. Jia, X., Wei, X., Cao, X., Han, X.: Adv-watermark: A novel watermark perturbation for adversarial examples (2020). https://doi.org/10.48550/ARXIV.2008.01919, https://arxiv.org/abs/2008.01919

24. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering **2** (2007)
25. Kong, Z., Xue, J., Wang, Y., Huang, L., Niu, Z., Li, F., Meng, W.: A survey on adversarial attack in the age of artificial intelligence. Wirel. Commun. Mob. Comput. **2021** (jan 2021). https://doi.org/10.1155/2021/4907754, https://doi.org/10.1155/2021/4907754
26. Kuang, X., Gao, X., Wang, L., Zhao, G., Ke, L., Zhang, Q.: A discrete cosine transform-based query efficient attack on black-box object detectors. Information Sciences **546**, 596–607 (2021). https://doi.org/https://doi.org/10.1016/j.ins.2020.05.089, https://www.sciencedirect.com/science/article/pii/S0020025520305077
27. Li, G., Zhu, P., Li, J., Yang, Z., Cao, N., Chen, Z.: Security matters: A survey on adversarial machine learning (2018). https://doi.org/10.48550/ARXIV.1810.07339, https://arxiv.org/abs/1810.07339
28. Li, Y., Tian, D., Chang, M.C., Bian, X., Lyu, S.: Robust adversarial perturbation on deep proposal-based models (2018). https://doi.org/10.48550/ARXIV.1809.05962, https://arxiv.org/abs/1809.05962
29. Lin, J., Xu, L., Liu, Y., Zhang, X.: Black-box adversarial sample generation based on differential evolution (2020). https://doi.org/10.48550/ARXIV.2007.15310, https://arxiv.org/abs/2007.15310
30. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. Lecture Notes in Computer Science p. 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2, http://dx.doi.org/10.1007/978-3-319-46448-0_2
31. Meng, D., Chen, H.: Magnet: a two-pronged defense against adversarial examples (2017). https://doi.org/10.48550/ARXIV.1705.09064, https://arxiv.org/abs/1705.09064
32. Narodytska, N., Kasiviswanathan, S.: Simple black-box adversarial attacks on deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1310–1318 (2017). https://doi.org/10.1109/CVPRW.2017.172
33. Naseer, M., Khan, S.H., Rahman, S., Porikli, F.: Task-generalizable adversarial attack based on perceptual metric (2018). https://doi.org/10.48550/ARXIV.1811.09020, https://arxiv.org/abs/1811.09020
34. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning (2016). https://doi.org/10.48550/ARXIV.1602.02697, https://arxiv.org/abs/1602.02697
35. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks (2015). https://doi.org/10.48550/ARXIV.1511.04508, https://arxiv.org/abs/1511.04508
36. Park, H., Ryu, G., Choi, D.: Partial retraining substitute model for query-limited black-box attacks. Applied sciences **10**(20), 1–19 (2020). https://doi.org/10.3390/app10207168
37. Qiu, S., Liu, Q., Zhou, S., Wu, C.: Review of artificial intelligence adversarial attack and defense technologies. Applied sciences **9**(5), 909 (2019). https://doi.org/10.3390/app9050909

38. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement (2018). https://doi.org/10.48550/ARXIV.1804.02767, https://arxiv.org/abs/1804.02767
39. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE transactions on evolutionary computation **23**(5), 828–841 (2019). https://doi.org/10.1109/TEVC.2019.2890858
40. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses (2017). https://doi.org/10.48550/ARXIV.1705.07204, https://arxiv.org/abs/1705.07204
41. Wang, D., Li, C., Wen, S., Han, Q.L., Nepal, S., Zhang, X., Xiang, Y.: Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples. IEEE Transactions on Cybernetics pp. 1–14 (2021). https://doi.org/10.1109/TCYB.2020.3041481
42. Wang, S., Su, Z.: Metamorphic testing for object detection systems (2019). https://doi.org/10.48550/ARXIV.1912.12162, https://arxiv.org/abs/1912.12162
43. Wang, Y., Tan, Y.a., Zhang, W., Zhao, Y., Kuang, X.: An adversarial attack on dnn-based black-box object detectors. Journal of network and computer applications **161**, 102634 (2020). https://doi.org/10.1016/j.jnca.2020.102634
44. Wei, X., Guo, Y., Li, B.: Black-box adversarial attacks by manipulating image attributes. Information sciences **550**, 285–296 (2021). https://doi.org/10.1016/j.ins.2020.10.028
45. Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14. ACM Press (2014). https://doi.org/10.1145/2601248.2601268
46. Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L., Jain, A.K.: Adversarial attacks and defenses in images, graphs and text: A review. International journal of automation and computing **17**(2), 151–178 (2020). https://doi.org/10.1007/s11633-019-1211-x
47. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. Proceedings 2018 Network and Distributed System Security Symposium (2018). https://doi.org/10.14722/ndss.2018.23198, http://dx.doi.org/10.14722/ndss.2018.23198
48. Zhang, J., Li, C.: Adversarial examples: Opportunities and challenges. IEEE Transactions on Neural Networks and Learning Systems **31**(7), 2578–2593 (2020). https://doi.org/10.1109/TNNLS.2019.2933524
49. Zhang, Q., Zhao, Y., Wang, Y., Baker, T., Zhang, J., Hu, J.: Towards cross-task universal perturbation against black-box object detectors in autonomous driving. Computer Networks **180**, 107388 (2020). https://doi.org/10.1016/j.comnet.2020.107388, https://www.sciencedirect.com/science/article/pii/S138912862030606X
50. Zhao, Y., Wang, K., Xue, Y., Zhang, Q., Zhang, X.: An universal perturbation generator for black-box attacks against object detectors. In: Qiu, M. (ed.) Smart Computing and Communication. pp. 63–72. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-34139-8_7
51. Zhou, M., Wu, J., Liu, Y., Liu, S., Zhu, C.: Dast: Data-free substitute training for adversarial attacks (2020). https://doi.org/10.48550/ARXIV.2003.12703, https://arxiv.org/abs/2003.12703