Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL)
ISSN P: 2303-5528 | E-ISSN: 2303-6397
Vol. 13 (2) 2026

Computer Vision Adversarial Attacks and Defenses: A Comprehensive Review

Isabelle Marceau
Laurentian Institute for Innovation, Canada

Submission : 17/10/2025 Acceptance : 20/07/2026 Published : 06/08/2026

Abstract:
When malicious actors try to trick computer vision models into mislabeling images by creating small, nearly undetected perturbations, the reliability and security of these models are put at risk. gives a comprehensive review of adversarial attack methodologies and how they impact various computer vision applications, such as picture identification, object detection, and face recognition. In order to decrease adversarial vulnerabilities, we examine attack and defense techniques, including defensive distillation, adversarial training, input modification, and ensemble methods. We assess these safeguards against the development of attack tactics by analyzing the trade-offs between computing cost, model correctness, and robustness. critical challenges to secure computer vision models, such as issues with comprehension, adaptability to defenses, and transferability. Continuous research into constructing more robust computer vision systems that can resist hostile manipulation is necessary to increase the safety of AI applications in crucial domains like autonomous driving, healthcare, and security. Our results underscore this importance.
keywords Adversarial Attacks, Computer Vision Security, White-Box Attacks, Black-Box Attacks, Gray-Box Attacks

Introduction

Computer vision has become an essential component in many modern applications, including autonomous vehicles, medical imaging, surveillance, and face recognition. Adversarial assaults, which involve malicious inputs intended to fool models into producing erroneous predictions, are becoming increasingly common, despite the fact that AI-powered systems are extremely helpful and accurate. When model dependability is crucial, adversarial assaults, which involve inserting tiny, often undetectable alterations to images, can cause algorithms to misclassify objects, exposing serious security flaws. This is especially true in high-stakes situations. One common way to categorize attacks is by the level of model knowledge the adversary possesses. Assuming full access to the model's parameters, architecture, and training data, attackers conducting white-box assaults can manipulate a model's outputs thru meticulously constructed perturbations. In contrast, black-box attacks can generate harmful samples without knowing the model at all by utilizing query-based strategies or information gleaned from other models. Because attackers in the real world only have partial knowledge of the model, gray-box attacks are more realistic. These attack types constitute a threat to computer vision model security, hence robust defense methods are important. In response to these threats, various defensive strategies have been developed, each with its own pros and cons. As an example, adversarial training involves adding hostile cases to the training set in

order to make the model more resistant to known attacks. In order to make the model more resistant to input disturbances, defensive distillation modifies the softmax output while it trains. One input modification strategy that aims to reduce adversarial noise before the model processes data is picture preprocessing. Ensemble approaches, in contrast, use a large number of models to detect and mitigate malicious inputs. Defense solutions sometimes compromise processing cost, model accuracy, and adaptability to new forms of attacks, demonstrating how challenging it is to protect against adversarial attacks. offers a comprehensive review of adversarial attacks and responses in computer vision by analyzing the efficacy of various approaches and the challenges in constructing safe models. Based on their computational feasibility, durability, and ability to adapt to new attack methods, we assess security solutions and study the effects of adversarial attacks on different computer vision applications. This review aims to provide light on the current state of adversarial research in computer vision, highlighting key areas that need further research, with the ultimate goal of developing AI systems that can withstand manipulation from adversaries while still producing reliable results in the real world.

Different Forms of Attacks by Opponents

Adversarial attacks in computer vision exploit AI model flaws to mispredict or misclassify incoming images by adding tiny, maybe undetectable disruptions. This type of attack is categorized according on the attacker's expertise with the model's structure, parameters, and data. It is critical to comprehend the many forms of adversarial attacks in order to construct robust defenses for applications such as autonomous driving, image recognition, and surveillance. The following are examples of hostile attacks that are frequently encountered in computer vision.

1. White-Box Attacks

Assuming full familiarity with the model's structure, parameters, and training data, white-box assaults proceed as follows. With this kind of access, malicious actors can tweak inputs to make misclassification more likely. A common tactic of white-box attacks is to manipulate the model's loss function gradients in order to introduce perturbations that are undetectable to humans but can trick the model. A few examples of common white-box assault techniques are:

- **Fast Gradient Sign Method (FGSM)**: Produces counterexamples by introducing disturbances perpendicular to the model's loss gradient. The speed and accuracy with which FGSM generates adversarial inputs has earned it widespread renown.
- **Projected Gradient Descent (PGD)**: Refining the perturbations to maximize model misclassification, PGD utilizes numerous gradient steps, making iterative improvements over FGSM.

Carlini & Wagner (C&W) Attack: One of the most effective white-box attacks, it uses a targeted optimization strategy to generate perturbations with little visual distortion.

Defense solutions in high-security applications should take white-box attacks into account since they show how deep learning models are vulnerable when attackers have full sight.

2. Black-Box Attacks

The attacker in a black-box approach relies only on querying the model and watching its outputs; they do not have access to the model's parameters, internal structure, or training data.

Due to the lack of publically known model data, black-box assaults are more difficult to implement than white-box attacks; yet, they are more appropriate for many practical applications. Approaches to black-box attacks encompass:

- **Query-Based Attacks**: The attacker in such an approach would continuously query the model and, based on the observed outputs, incrementally refine the perturbations in order to build adversarial samples. By incrementally traversing the decision boundary in the absence of gradients, techniques such as the Boundary Attack generate adversarial samples.

Transferability Attacks: This approach takes advantage of the ubiquitous transferability of adversarial cases from one model to another, provided that the two models have comparable architectures or tasks. Attackers first train a dummy model that mimics the target model's architecture in order to supply it with adversarial instances.

Because adversarial examples can generalize beyond the specific model they were built for, black-box attacks show that adversarial examples are durable across models and highlight the significance of cross-model countermeasures.

3. Gray-Box Attacks

In gray-box attacks, the attacker only knows a little bit about the model, which is in between a white-box and a black-box attack. Even though they don't have the training data or precise parameters, attackers often know the model's architecture. Attackers with this degree of understanding can use targeted strategies to some extent, but they could still have to resort to more circumstantial means to succeed.

- **Limited Parameter Knowledge**: In this case, the attacker has knowledge of the model's architecture or some of its parameters, which allows them to use less accurate techniques such as FGSM or PGD to potentially misclassify the model.

Transfer-Based Approaches with Architecture Knowledge: An adversary could try to create transferable adversarial examples by creating a comparable model with different training data but the same architecture. In order to take advantage of architectural weaknesses, this method blends white-box and black-box tactics.

Defenses that depend only on parameter secrecy are inadequate in situations when some model features are known, like in academic or collaborative settings, where gray-box attacks become relevant.

4. Targeted vs. Untargeted Attacks

Additionally, the objectives of adversarial attacks can be categorized as either targeted or untargeted:

- **Targeted Attacks**: In this attack, the goal is to trick the model into making an inaccurate prediction for a certain class. One example is the intentional alteration of a "dog" photograph so that it is labeled as a "cat." Although targeted assaults are more challenging to implement, they are more effective in situations where the attacker stands to gain from a specific mistake.
- **Untargeted Attacks**: The goal of an untargeted attack is to generate a misclassification in general, rather than targeting a particular class. Due to the success metric of any departure from the correct class, these assaults are usually easier to generate.

5. Physical Adversarial Attacks

The goal of physical adversarial attacks is to trick computer vision models into thinking they are dealing with real objects instead than digital pictures by making alterations to the real environment. For uses like surveillance systems and autonomous cars, these kinds of attacks are a major worry.

- **Adversarial Patches**: Attackers can trick models into misclassifying real-world objects by printing out tiny patches with intricate patterns and hiding them in an object's perspective. In autonomous vehicles, for example, it has been demonstrated that adversarial patches can trick object detecting algorithms.

3D Adversarial Objects: Here, the adversary changes the items' physical characteristics or forms in order to influence the model's categorization. The effectiveness of facial recognition systems and security cameras has been shown through the use of such assaults.

Because of the serious dangers posed to applications that rely on security by physical adversarial attacks, it is essential to have strong defenses that can resist manipulations from both online and offline adversaries.

To build strong defense tactics in computer vision, it is vital to understand the many kinds of adversarial attacks. Each kind of assault, from those that operate with complete model knowledge (white-box attacks) to those that operate with partial knowledge (black-box and gray-box attacks), poses its own set of problems that draw attention to the weaknesses of AI models. Computer vision systems are useful in many different contexts, and by studying these assault tactics, experts in the field may better foresee such dangers and build thorough defenses to protect them.

Conclusion

The rapid growth of computer vision applications in domains such as healthcare, security, and autonomous systems has highlighted the critical need for robust defenses against malicious assaults. Adversarial attacks expose critical AI system weaknesses by subtly changing inputs to cause model misclassifications, which is especially problematic in high-stakes contexts where the models' accuracy and reliability are critical. This comprehensive examination has looked at a variety of adversarial attack strategies, including physical adversarial manipulations, white-box and black-box attacks, and more. There have also been evaluations of the effectiveness of defensive distillation, adversarial training, input transformation, and ensemble approaches. In spite of the fact that defense mechanisms have made models more resilient, we still found issues. Finding a happy medium between computational efficiency, interpretability, and model robustness is challenging because every defense method has its own set of trade-offs that could impact the model's accuracy and scalability. Given the dynamic nature of adversarial attack techniques, it is crucial to create hybrid and adaptable defenses capable of generalizing across various types of attacks. Finally, the need of ongoing innovation and interdisciplinary study is underscored by the fact that adversarial attacks and defenses in computer vision are constantly engaging in head-to-head competition. Developing trustworthy computer vision models that can withstand manipulation by adversaries will require cooperation between areas such as artificial intelligence (AI), cybersecurity (CSF), and

application-specific domains. Improving our understanding of attack tactics and response mechanisms is crucial for building a more safe and dependable AI ecosystem. After that, we will be able to build computer vision systems that are secure and resilient enough to resist real-world threats.

Bibliography

- Aravind Ayyagiri, Prof.(Dr.) Punit Goel, & A Renuka. (2024). Leveraging AI and Machine Learning for Performance Optimization in Web Applications. *Modern Dynamics: Mathematical Progressions*, *1*(2), 89–104. https://doi.org/10.36676/mdmp.v1.i2.13
- Lohia, A. (2024). Quantum Artificial Intelligence: Enhancing Machine Learning with Quantum Computing. *Journal of Quantum Science and Technology*, *1*(2), 6–11. https://doi.org/10.36676/jqst.v1.i2.9
- Alice Tan. (2024). Enhancing Cyber Defense Mechanisms: AI and Machine Learning-Based Threat Mitigation Strategies. *Journal of Quantum Science and Technology*, *1*(3), 90–93. https://doi.org/10.36676/jqst.v1.i3.30