Understanding Object Detection Vulnerabilities in the Age of YOLO v11

Trinity Banks | College of Science and Technology, Florida Agricultural & Mechanical University, Tallahassee, FL, USA | triniy1.banks@famu.edu

Idongesit Mkpong-Ruffin | College of Science and Technology, Florida Agricultural & Mechanical University, Tallahassee, FL, USA | idongesit.ruffin@famu.edu

Chutima Boonthum-Denecke, PhD |Department of Computer Science, Hampton University, Hampton, VA | chutima.boonthum@hamptonu.edu

Deidre Evans | College of Science and Technology, Florida Agricultural & Mechanical University, Tallahassee, FL, USA | deidre.evans@famu.edu

Abstract

Object detection models are widely used in technologies such as surveillance, robotics, and autonomous driving. This literature review explores how models like YOLOv11 have shaped the field while also remaining vulnerable to adversarial attacks. These attacks, including adversarial patches, can cause models to misinterpret images in dangerous ways. By reviewing current research, this paper highlights how these vulnerabilities work, why they matter, and why further study is needed to improve the safety and reliability of modern object detection. Object detection models, particularly the "You Only Look Once" (YOLO) series, have seen rapid architectural evolution from YOLOv1 to the current YOLOv11. While deep learning models are traditionally considered vulnerable to adversarial attacks, this study identifies a significant shift in the effectiveness of these attacks within modern training frameworks.

The rapid integration of object detection models into these safety critical domains has made the security of these systems a paramount concern. While previous iterations of the “You Only Look Once” (YOLO) architecture have been extensively studied, the transition to the anchor-free model YOLOv11 presents a new landscape for adversarial robustness. This study performed a comprehensive analysis of established adversarial patch methodologies: the Dynamic Adversarial Patch (DAP), which utilized Creases Transformation (CT) blocks to account for movement, and the Remote Adversarial Patch (IPatch) which seeks to manipulate model semantics from a distance.

Based on the goal of securing autonomous vehicle environments, this study implemented a custom replication of the IPatch methodology, adapting it from images segmentation to object detection. The experimental framework utilized the BDD100K dataset and employed Expectation over Transformation (EoT) to ensure the patch remained effective across various angles and distances. Despite an optimization process spanning 1,000 epochs using the Adamax optimizer, the replication failed to achieve the intended adversarial suppression or false-positive results.

A significant outcome of this study arose during subsequent robustness testing, where I evaluated YOLOv11's response to various digital perturbations. These experiments revealed that the model consistently and correctly identified the adversarial patches themselves as distinct objects (such as a "teddy bear" or "person"), effectively neutralizing the attack's stealth. Building directly upon these findings, our future research will pivot to achieving adversarial stealth. Our upcoming work will explore the development of patches designed to be semantically hidden from the detection head by incorporating similarity objectives that blend the patch into the environmental background, testing these stealthier boundaries through physical experiments to improve the security of high-stakes technologies.

CS Concepts: Computing methodologies → Computer vision; Machine learning; Object detection; Neural networks; Adversarial machine learning

Keywords: YOLO; object detection; adversarial attacks; adversarial patches; computer vision