Adversarial Attacks on YOLOv9: White-box FGSM

November 24, 2025

Abstract

This report investigates the vulnerability of the YOLOv9e object detection model to adversarial perturbations generated using the Fast Gradient Sign Method (FGSM) under a white-box threat model. The goal is to understand how small, imperceptible perturbations can significantly degrade object detection performance. We implement a complete FGSM pipeline, generate adversarial versions of a validation dataset, and evaluate the performance drop in terms of mAP, precision, and recall. The results indicate that YOLOv9e remains highly susceptible to gradient-aligned perturbations due to its complex feature hierarchy and distribution-based regression head.

1 Introduction

Adversarial attacks exploit the sensitivity of deep learning models to carefully crafted perturbations that remain invisible to human observers. FGSM, one of the simplest and fastest adversarial techniques, generates perturbations by leveraging the gradient of the loss with respect to the input image.

Object detection models such as YOLOv9e rely on deep convolutional feature extractors and non-linear activation functions, which make them inherently vulnerable to such attacks. In this study, we apply a white-box FGSM attack to a YOLOv9e model trained on aerial imagery and quantify the resulting performance degradation on standard evaluation metrics.

The goals of this work are:

- Implement a white-box FGSM attack compatible with the YOLOv9 architecture.
- Generate adversarial images for an entire validation set.
- Evaluate the impact of the attack on detection accuracy.
- Analyse architectural reasons for YOLOv9e's vulnerability.

2 Algorithm and Implementation

2.1 FGSM Attack

The Fast Gradient Sign Method perturbs an input image x by adding noise in the direction of the gradient of the loss function J with respect to the input:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

Here:

- x is the clean input image,
- y is the ground-truth label,

- θ represents model parameters,
- ϵ is a small perturbation factor.

In YOLOv9e, the loss consists of classification, bounding box regression, and objectness components, all of which contribute gradients for generating perturbations.

2.2 Implementation Steps

The attack implementation proceeds as follows:

1. Load a pretrained YOLOv9e model.
2. Preprocess the input image and enable gradient tracking.
3. Perform a forward pass and compute the YOLO loss.
4. Backpropagate to compute $\nabla_x J$.
5. Generate adversarial perturbations using the FGSM formula.
6. Clamp the adversarial image to maintain valid pixel ranges.
7. Save the adversarial images and corresponding label files.

The final adversarial dataset mirrors the folder structure of the clean validation set to enable standard YOLO validation.

3 Results

3.1 Clean Set Performance

Metric	Value
mAP@0.50	0.6862
mAP@0.50:0.95	0.5514
Precision	0.8607
Recall	0.6235

3.2 Adversarial (FGSM) Set Performance

Metric	Value
mAP@0.50	0.4803
mAP@0.50:0.95	0.3297
Precision	0.7524
Recall	0.4104

3.3 Performance Drop

Metric	Drop
Δ mAP@0.50	0.2059
Δ mAP@0.50:0.95	0.2217
Δ Precision	0.1083
Δ Recall	0.2131

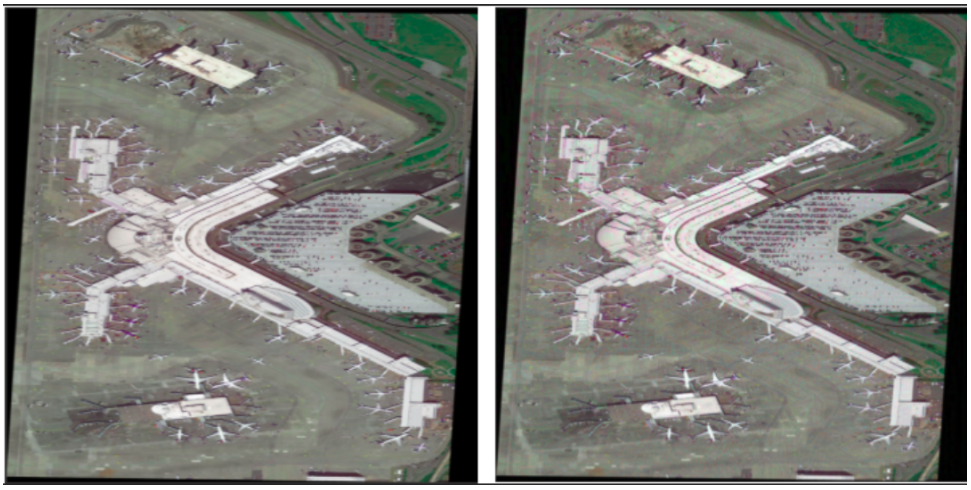

Figure 1: Original Image (L) and Adversarial Image (R)

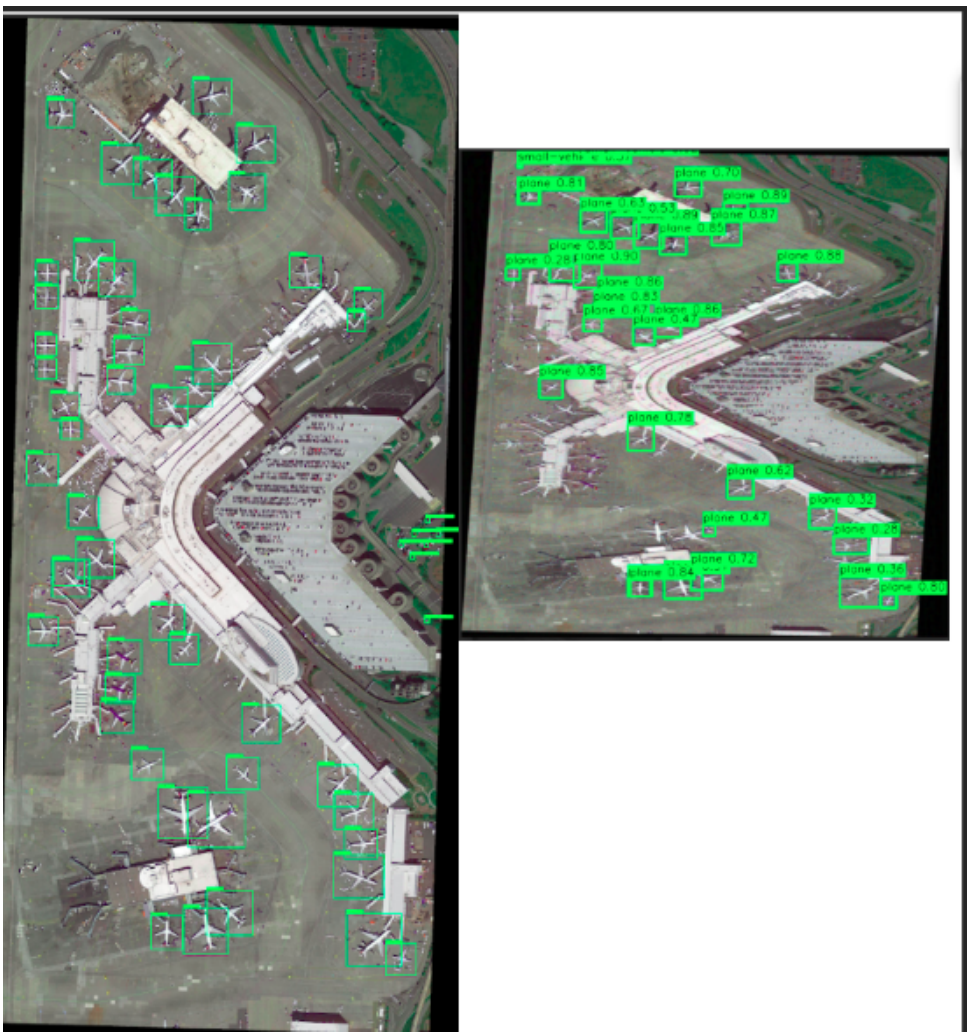

Figure 2: Original Output (L) and Adversarial Image output (R): Missed a few planes

3.4 Discussion

The attack causes:

- A significant reduction in recall, indicating many objects were not detected.
- A moderate drop in precision, suggesting increased uncertainty in predictions.
- Over **20% loss in mAP metrics**, showing clear degradation in bounding box localization.

Despite FGSM being a single-step attack, its impact is substantial, demonstrating the fragility of YOLOv9e under white-box adversarial conditions.

4 Why the Attack Works

FGSM is effective on YOLOv9e because of the following architectural characteristics:

4.1 Deep Non-linear Feature Hierarchy

YOLOv9e utilizes complex modules such as:

- C2f blocks,
- RepNCS convolutional units,
- Multi-scale feature pyramid structures.

These introduce non-linearities that amplify small gradient-aligned changes.

4.2 DFL-Based Bounding Box Regression

YOLOv9e predicts bounding boxes using Distribution Focal Loss (DFL), where each coordinate is represented as a discrete probability distribution. Small perturbations disrupt these distributions, causing:

- misaligned bounding boxes,
- reduced IoU,
- unstable confidence scores.

4.3 Sensitivity of Objectness Scores

Objectness logits pass through a sigmoid activation, making them extremely sensitive to tiny changes in the feature map.

4.4 White-Box Advantage

The attacker has full access to:

- gradients,
- losses,
- intermediary activations,

allowing precise targeting of model weaknesses.

5 Conclusion

This study demonstrates that YOLOv9e is highly vulnerable to white-box FGSM adversarial attacks. Even small perturbations cause significant degradation in key detection metrics such as mAP, precision, and recall. The vulnerability primarily arises from the model's deep non-linear architecture and distribution-based regression head.