Adversarial Detection: Attacking Object Detection in Real Time

*Han Wu, Syed Yunas, Sareh Rowlands, Wenjie Ruan, Johan Wahlstrom**
University of Exeter

IEEE IV 2023
IEEE INTELLIGENT VEHICLES SYMPOSIUM
Anchorage, Alaska, USA // June 4 – 7, 2023

Poster No. P13 - 02

Abstract

Intelligent robots rely on object detection models to perceive the environment. Following advances in deep learning security it has been revealed that object detection models are vulnerable to adversarial attacks. However, prior research primarily focuses on attacking static images or offline videos. Therefore, it is still unclear if such attacks could jeopardize real-world robotic applications in dynamic environments. This paper bridges this gap by presenting the first real-time online attack against object detection models. We devise three attacks that fabricate bounding boxes for non-existent objects at desired locations. The attacks achieve a success rate of about 90% within about 20 iterations. The demo video is available at: https://youtu.be/zJZ1aNlXsMU.

Introduction

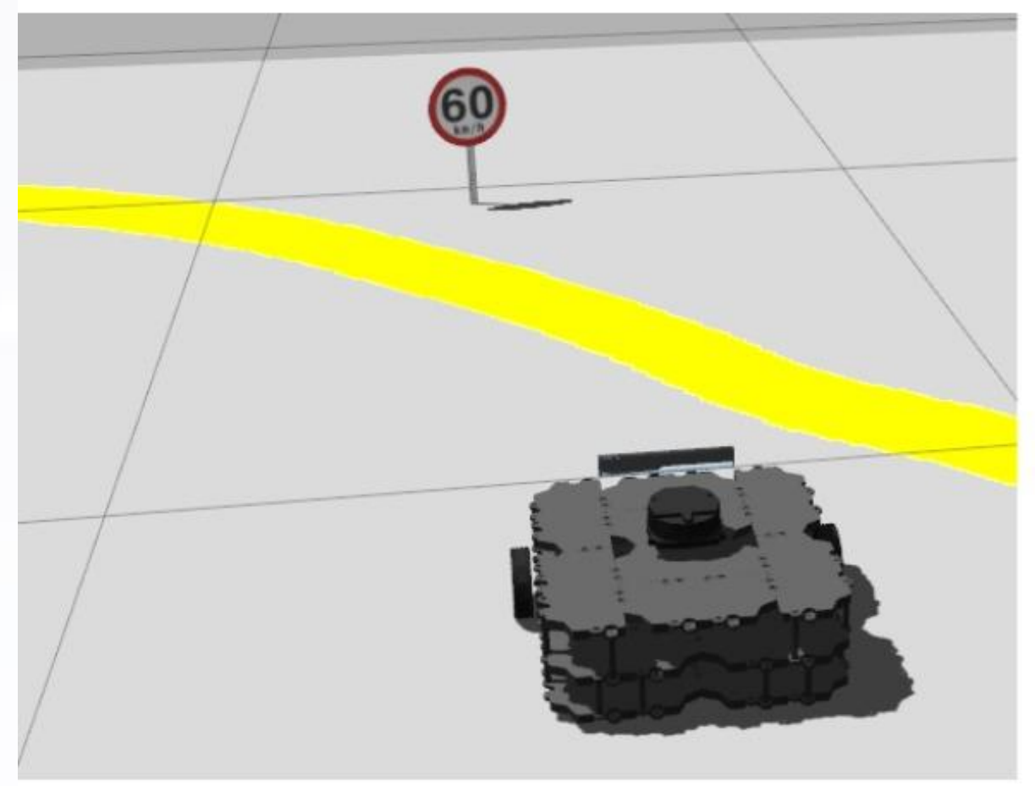
Reliable object detection is crucial for multiple safety-critical robotic applications. For example, an autonomous driving system relies on object detection models to perceive the environment and take action on it. While advances in deep neural networks have rapidly increased the availability of high-accuracy object detection models, this breakthrough has also revealed several potential vulnerabilities.

The first adversarial attacks fooled image classification models by adding imperceptible perturbations to the input image. Later research removed the restriction to imperceptible perturbations and instead designed an adversarial patch that can be printed in the physical world. Therefore, in 2018, Liu et al. introduced the DPatch, a digital patch that fools object detection models, and Lee et al. later extended the attack to physical patches.

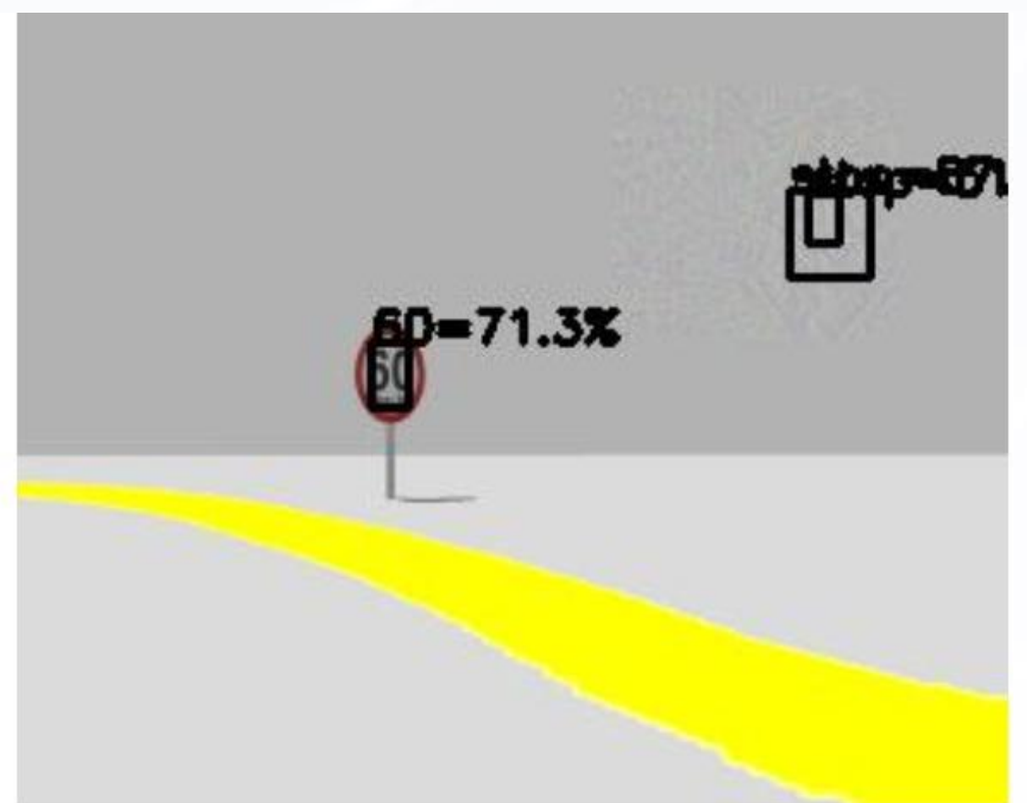
The generation of adversarial patches for attacking real-time robotic applications is still a rather challenging task. The generation process itself is often very computationally expensive. In addition, the efficiency of the physical patch is conditioned on strict requirements on the relative distance and orientation of the camera and the patch. These requirements are often difficult to satisfy for real-world robotic applications deployed in dynamic environments.

Contributions

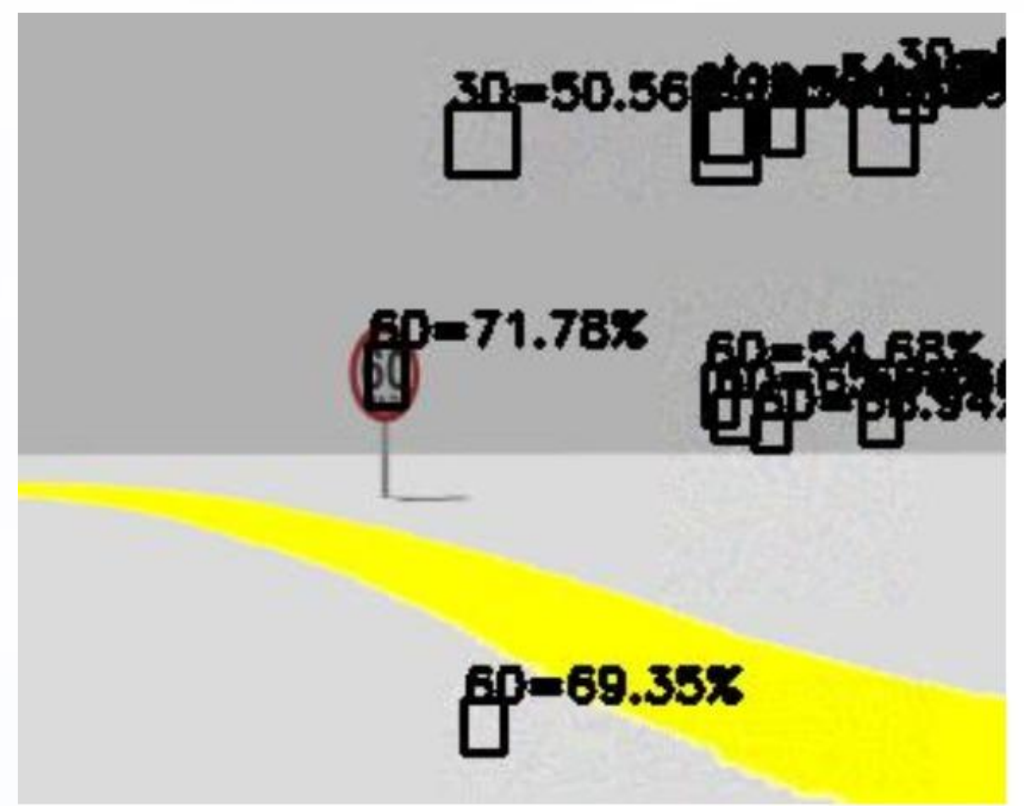
- We devise three adversarial attacks that can generate digital overlays of different shapes at specified locations in real time. This extends previous research that primarily has focused on square patches.
- Investigating the effect of various digital overlays it is found that the size of the overlay is critical for the success of the attack, whereas the attack performance is independent of the aspect ratio.
- We test our attacks in the Robot Operating System (ROS). The system is open-sourced to facilitate future extensions and comparisons.



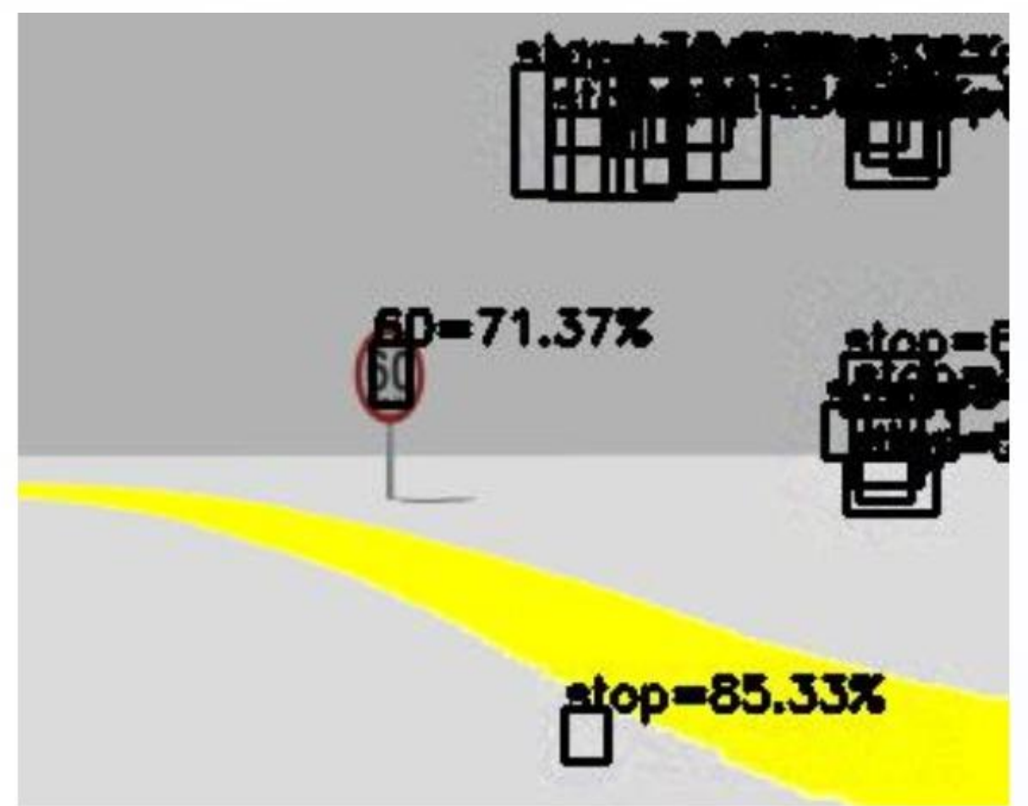
The original image



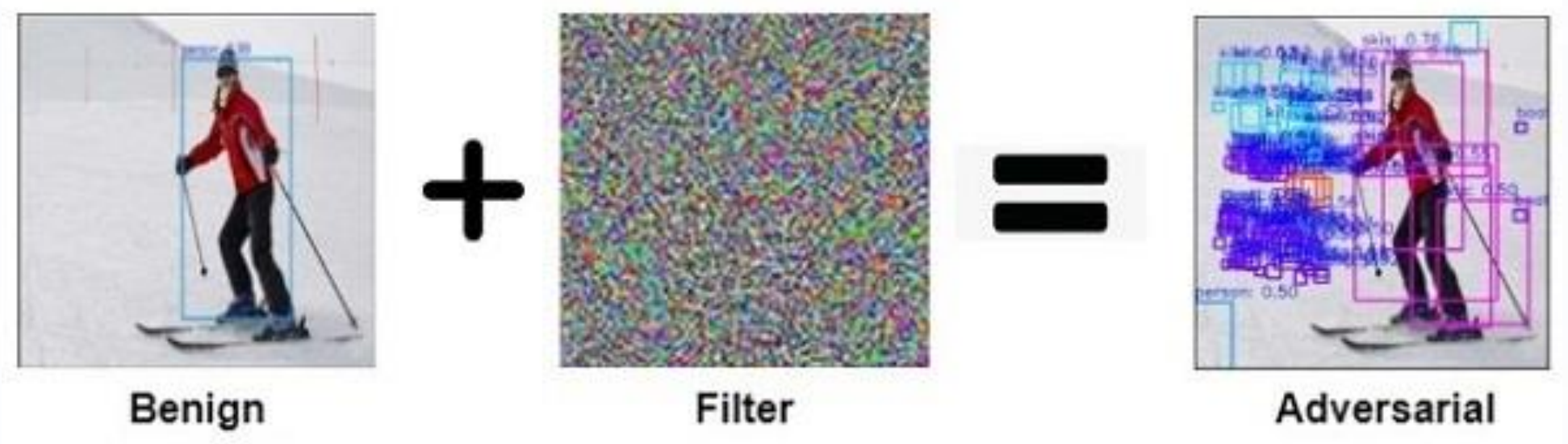
The One-Targeted Attack



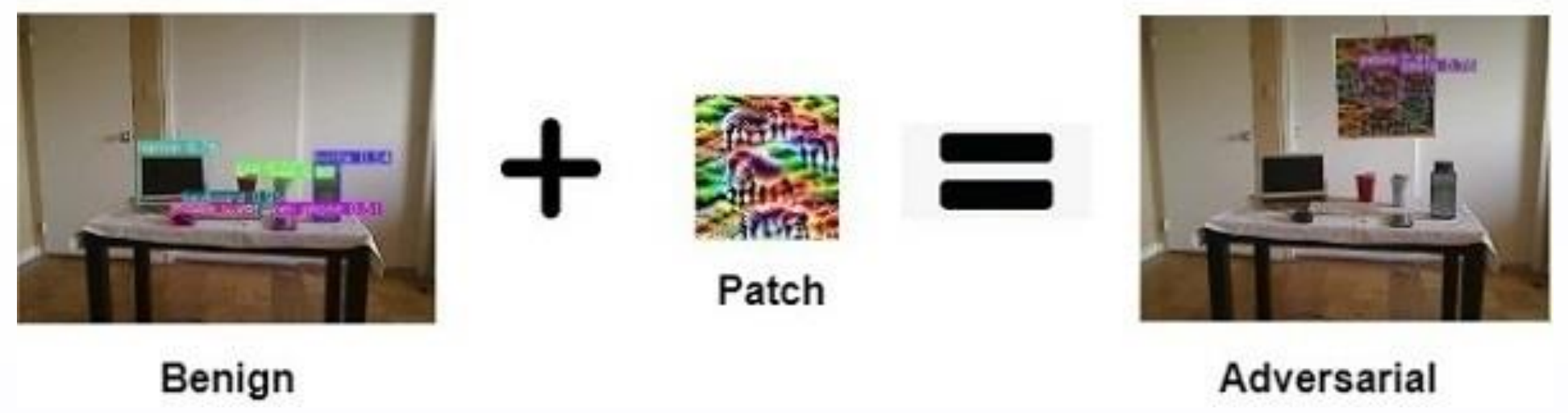
The Multi-Untargeted Attack



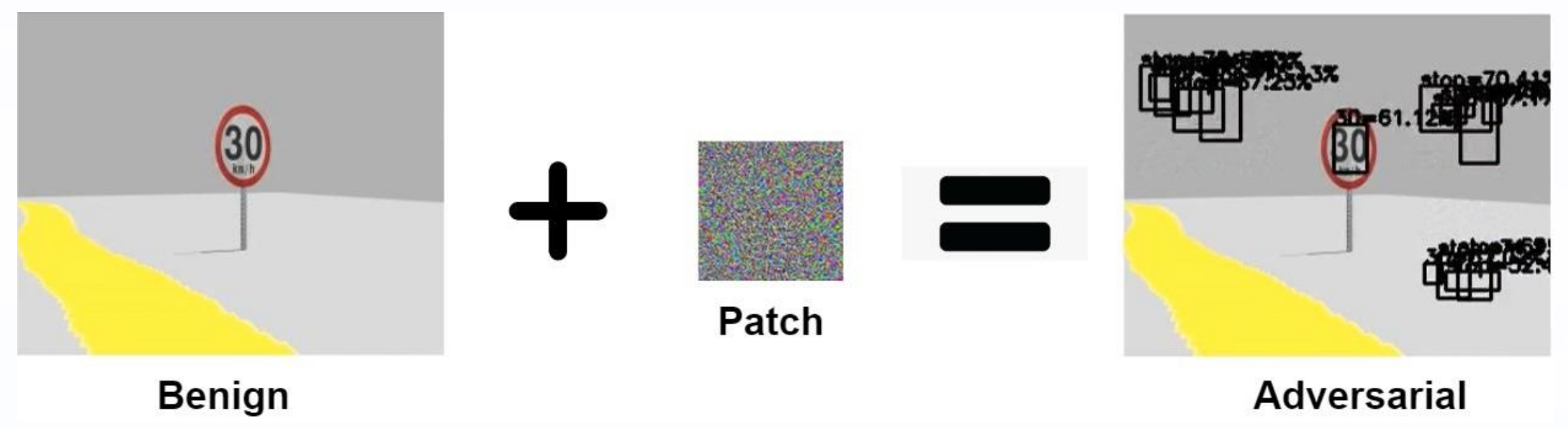
The Multi-targeted Attack



Adversarial Filter



Adversarial Patch



Adversarial Overlay (Ours)

One-Targeted Attack: $\max_{i \leq i \leq |\mathcal{O}|} \delta(c^i) * \delta(p_t^i)$

Multi-Untargeted Attack: $\sum_{i=1}^{|\mathcal{O}|} \sum_{j=1}^{K} [\delta(c^i) * \delta(p_j^i)]$

Multi-Targeted Attack: $\sum_{i=1}^{|\mathcal{O}|} [\delta(c^i) * \delta(p_t^i)]$

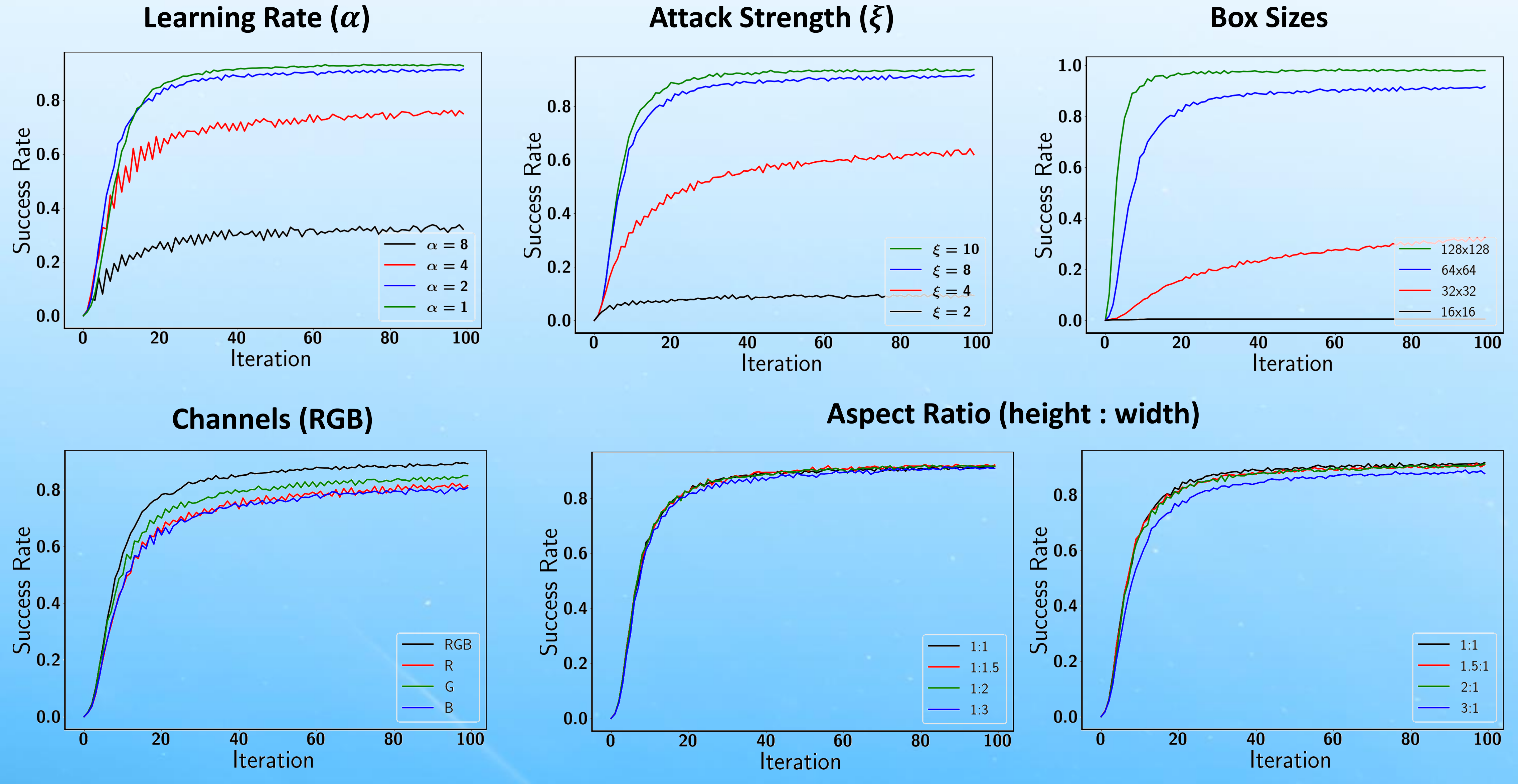
$$x'_{filter} = x + \delta$$

$$x'_{patch} = (1 - m) \odot x + m \odot \delta$$

$$\boldsymbol{x'_{overlay} = x + m \odot \delta}$$

Experimental Results

The effect of various hyper-parameters on the attack success rate


Learning Rate (α)

Attack Strength (ξ)

Box Sizes

Channels (RGB)

Aspect Ratio (height : width)

Real-time Performance

We measured the performance of the attack on an NVIDIA RTX 2080Ti GPU tested on the VOC 2012 validation set, which includes 5823 images in total.

The attack achieved 24 FPS (1 iteration costs 41 ms). It should be noted that the performance of the attack also depends on model size. A larger model requires more computations to compute the gradient, while the computation time does not grow as the box size increases.

For an online attack, the attack can be even more efficient. We can save computations by reusing the overlay generated in the previous timestep since there is a high correlation between consecutive video frames and iterations.

Box Sizes	1 iteration	10 iterations	20 iterations
64x64	41 ms	410 ms	780 ms
128x128	41 ms	410 ms	781 ms

TABLE I: The time cost of the attack with different numbers of iterations ($\alpha = 2,\ \xi = 8$).

Box Sizes	N=1	N=3	N=5
64x64	7.47 it (306 ms)	12.03 it (493 ms)	13.62 it (558 ms))
128x128	4.40 it (180 ms)	7.64 it (313 ms)	10.20 it (418 ms)

TABLE II: The average number of iterations and time cost of generating N bounding boxes ($\alpha = 2,\ \xi = 8$).

Conclusion:

This paper has demonstrated that it is possible to attack an object detection system in real time. We generate human unperceivable adversarial overlays of arbitrary shapes to fabricate bounding boxes at desired locations. This attack could be a threat to the areas of traffic sign recognition and the autonomous driving field.

Future Research:

In the future, we plan to investigate the effect of the attack on modular autonomous driving systems that rely on object detection models to perceive the environment. In addition, we will explore how to detect adversarial attacks so that we can embrace deep learning models in safety-critical robotic applications in a safe way.

Talk Slides

Our Research