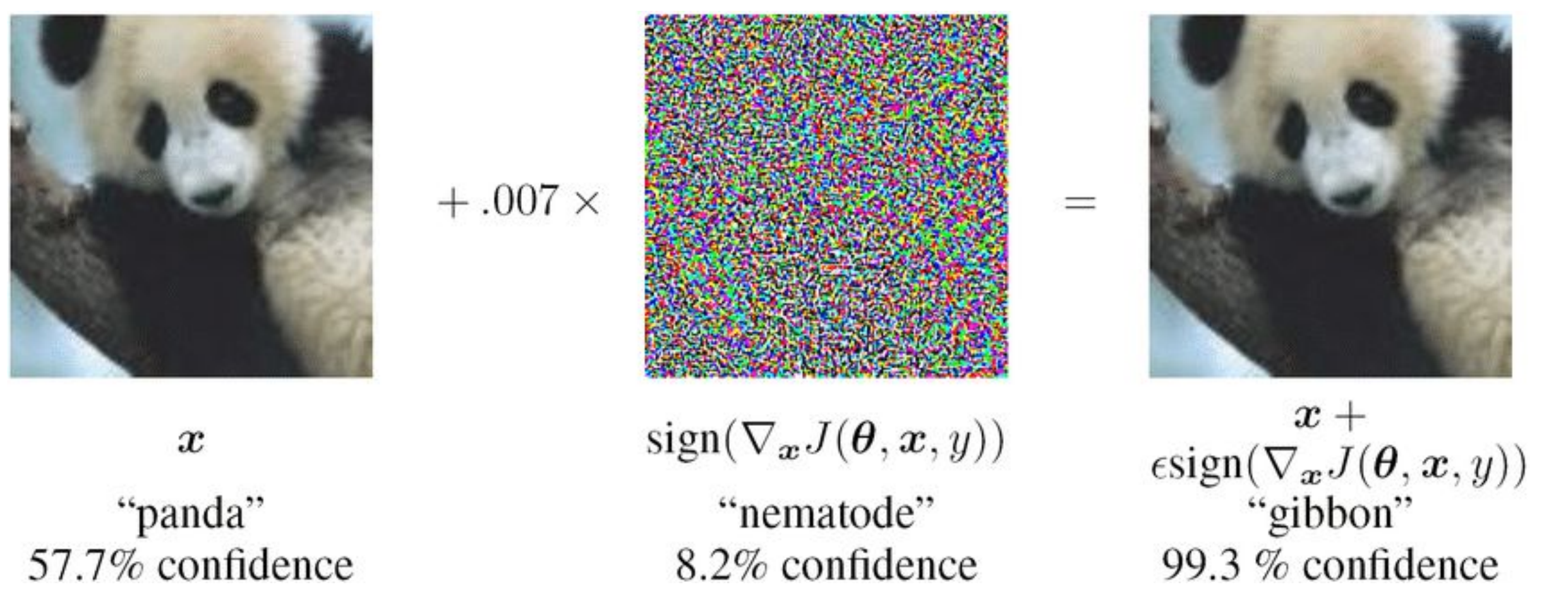
DeepManeuver: Adversarial Test Generation for Trajectory Manipulation of Autonomous Vehicles

Meriel von Stein   David Shriver   Sebastian Elbaum

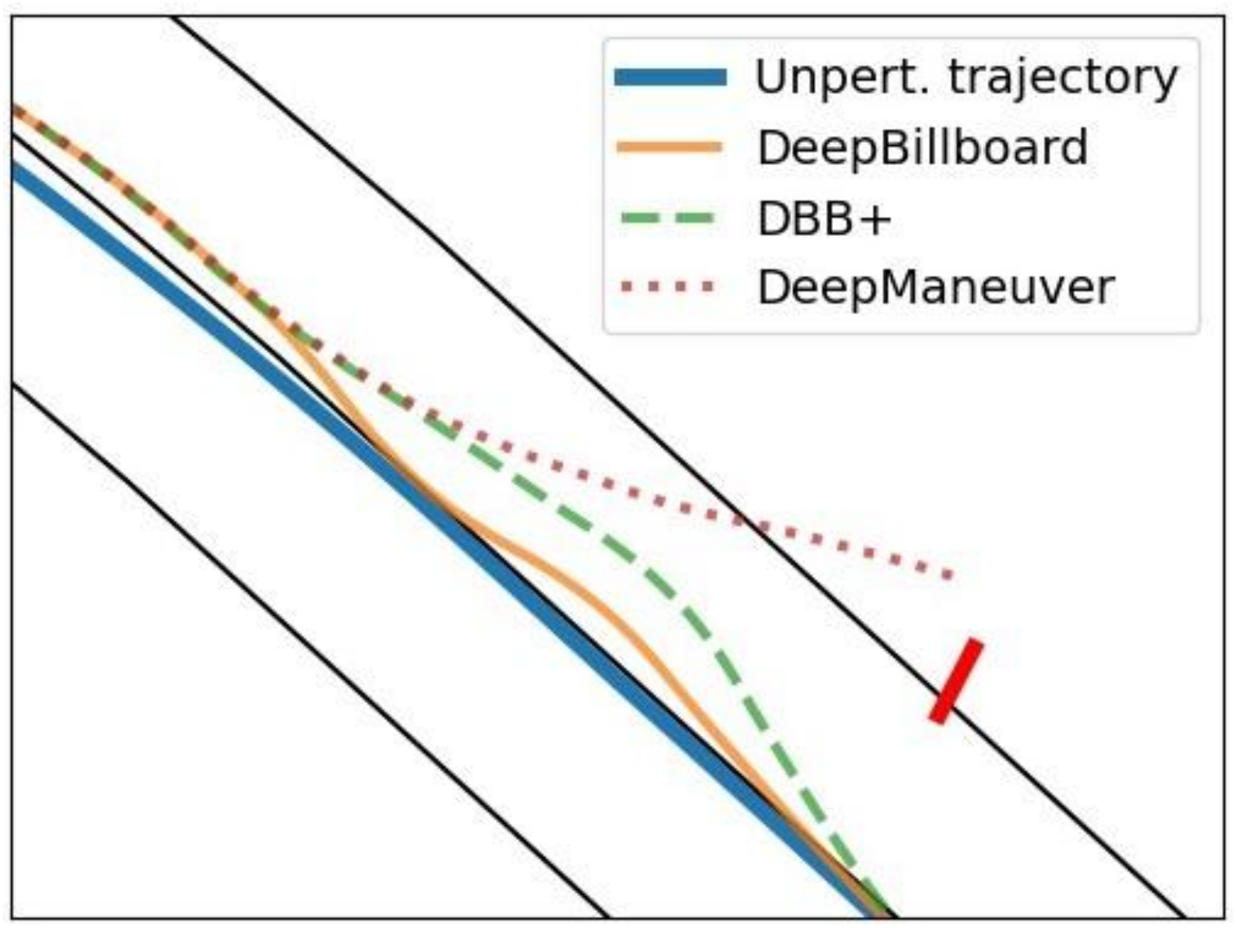
University of Virginia

Problem


+ .007 ×
=
x
"panda"
57.7% confidence
$\text{sign}(\nabla_x J(\theta, x, y))$
"nematode"
8.2% confidence
$x + \epsilon\text{sign}(\nabla_x J(\theta, x, y))$
"gibbon"
99.3 % confidence

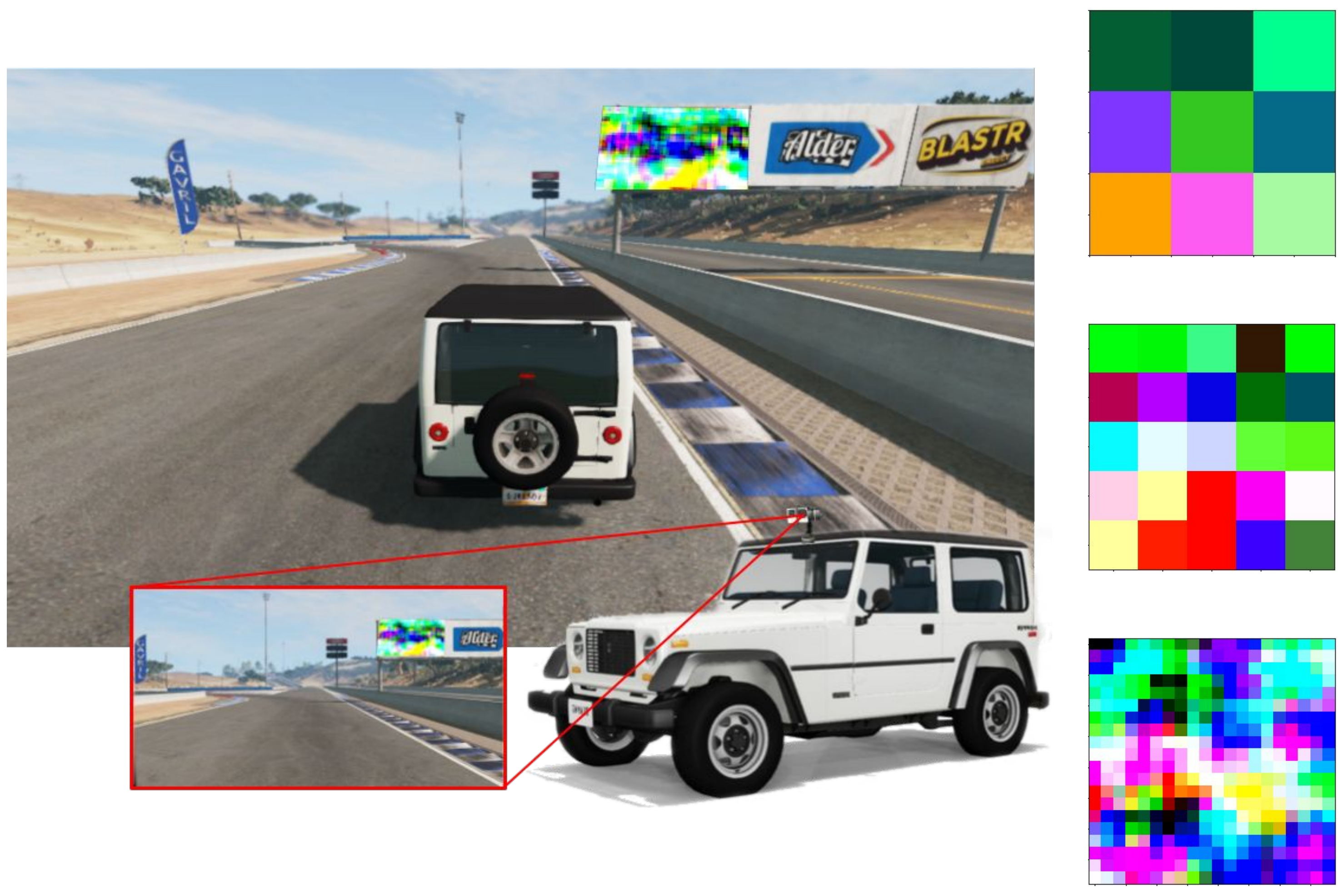
Adversarial attacks are often generated for individual images, or for an unordered collection of images meant to represent a test case.

In the real world, images in a test case are ordered, dependent, and tightly coupled with vehicle state. The evolution of state over time can affect perturbation strength. As the perturbation influences the vehicle trajectory, it may become less relevant to the inputs seen by the vehicle's DNN, reducing its perturbation strength.



Insight

Compounding effects of adversarial perturbations and spatiotemporally related changes in vehicle state must be taken into account during the perturbation generation process. We introduce DeepManeuver, the first state-adaptive approach to generate adversarial perturbations that cause complex maneuvers.



Solution

State-Adaptive Adversarial Testing

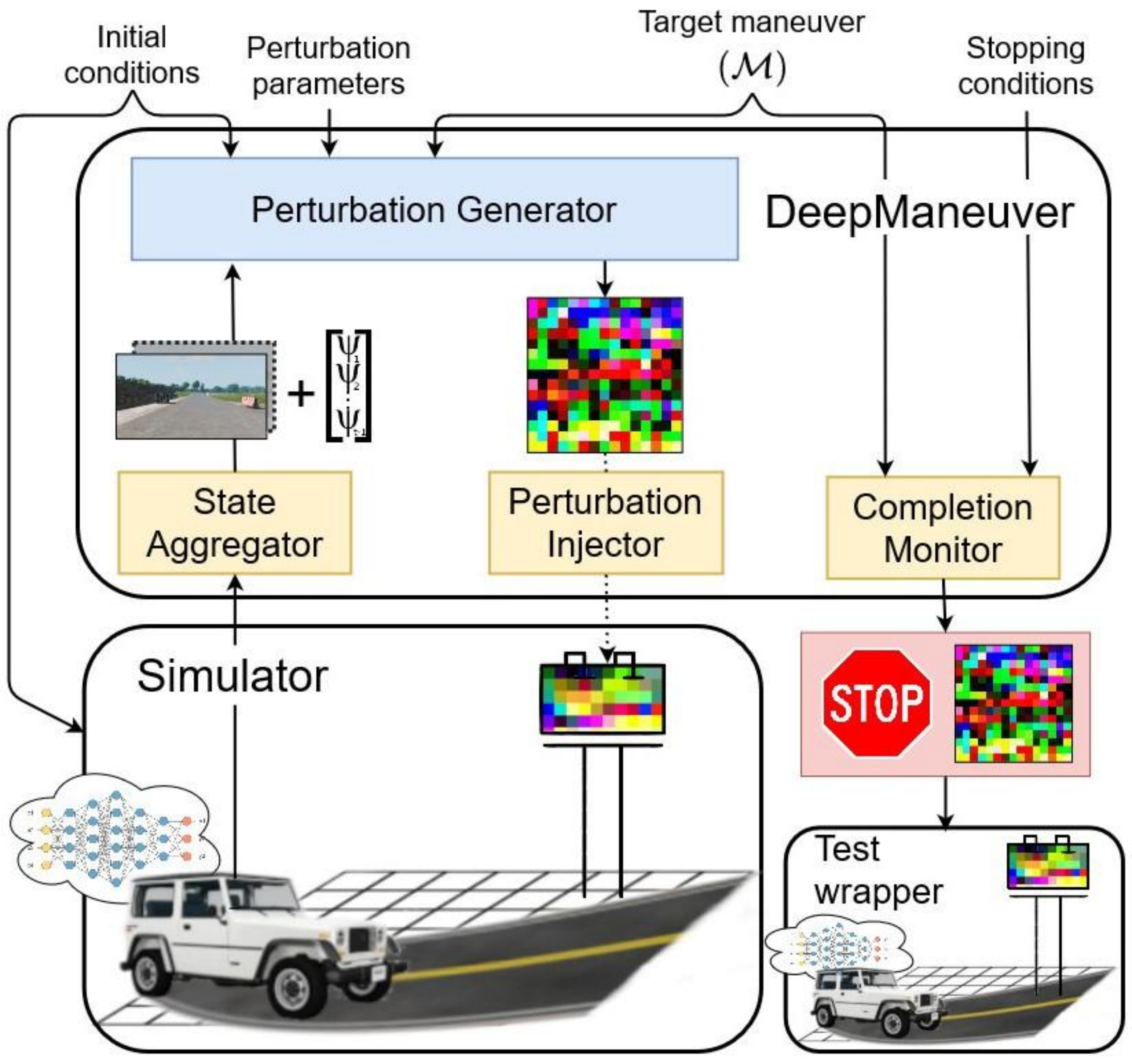
Given a DNN for vehicle actuation and a simulator, we build an automated, parameterized framework DeepManeuver that interleaves adversarial test generation with vehicle physics simulation. This creates a refined adversarial test case for the vehicle at the system level.

DeepManeuver

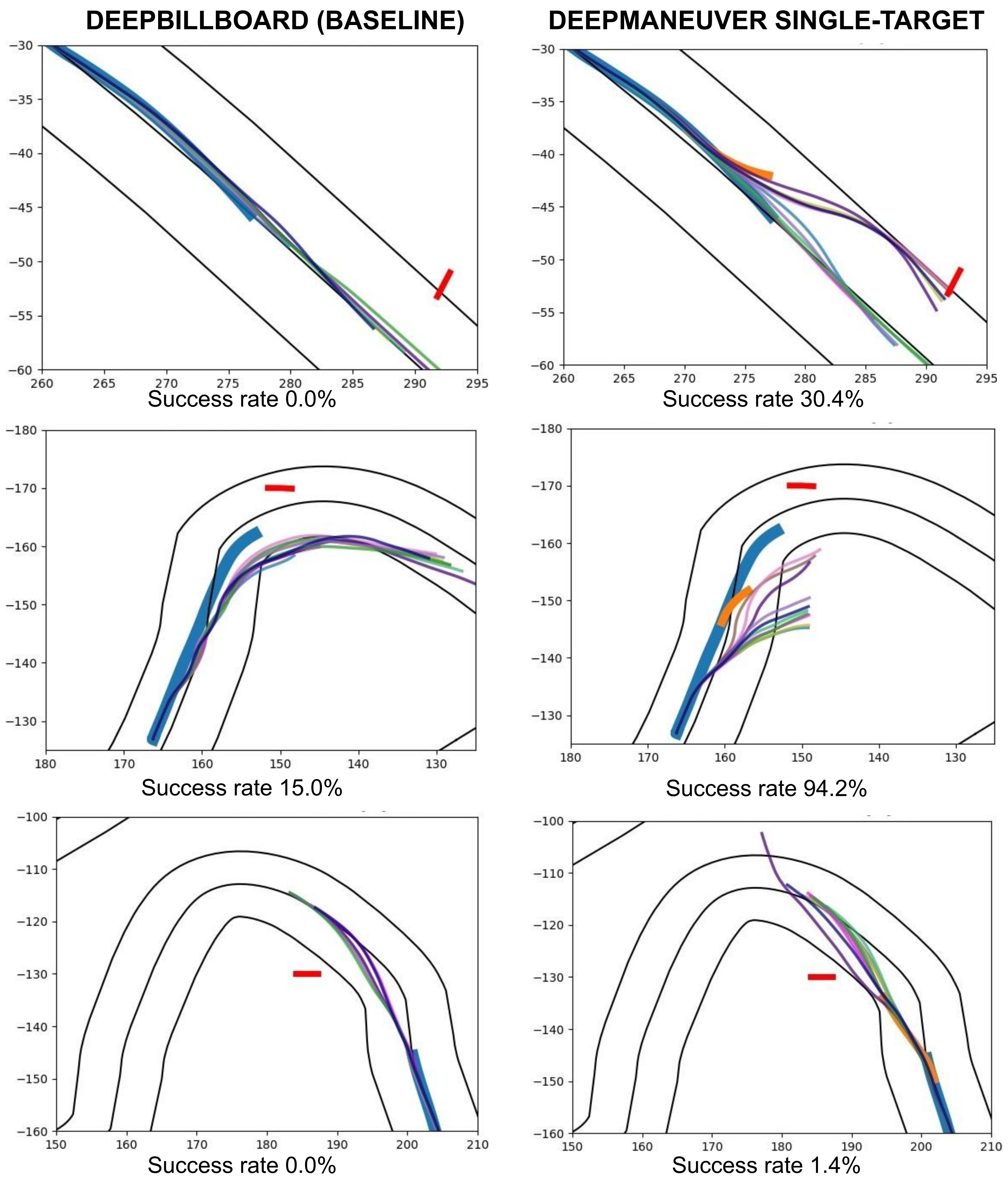
We modify the perturbation generation process to iteratively update the perturbation upon changes in vehicle state to preserve compounding effects of the perturbation in previous states. At each step, new DNN inputs and vehicle actuation values from the simulator are passed to the generator for refinement of the perturbation.

This is enabled through a two-part weighted loss function (below) for the generator that adjusts loss terms according to their importance to the trajectory.

$$\underset{\eta}{argmin}\left(w_n\mathcal{L}_1(\mathcal{N}(img_n+\eta), target_{\psi,n}) + \sum_{t=0}^{n-1} w_t\mathcal{L}_2(\mathcal{N}(img_t+\eta), \psi_t)\right)$$



Results



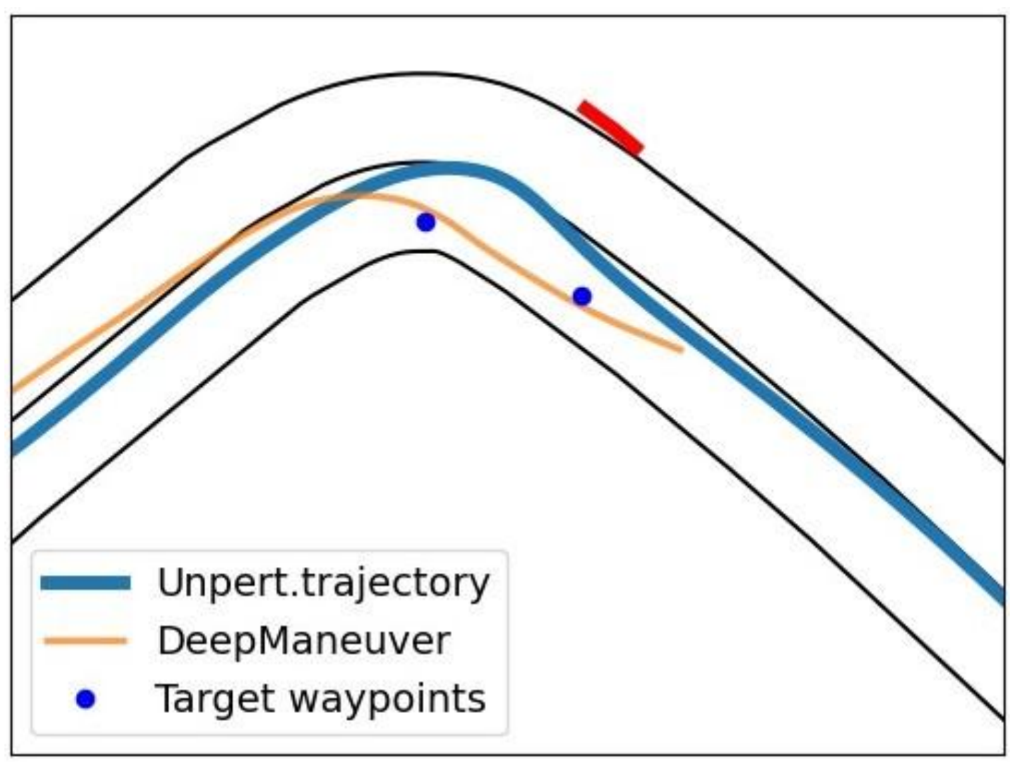
Left:
Performance comparison of state-of-the-art baseline technique DeepBillboard against DeepManeuver. Success rate denotes fulfillment of the maneuver. Across all parameterizations and environments, DeepManeuver shows a 20.7 percentage point increase in success rate over DeepBillboard.

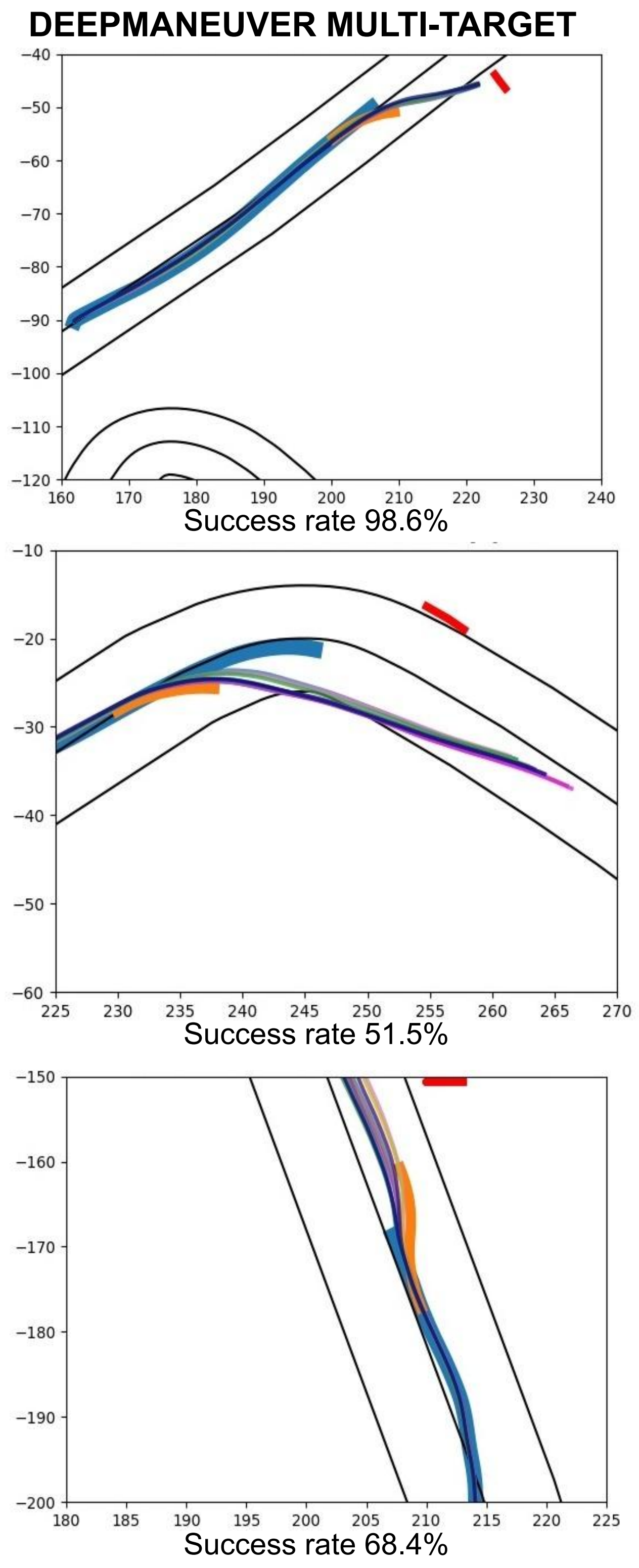
Right:
Performance for three multi-target maneuvers: hit a bullseye, cut a corner, and change lanes.

Bottom:
Simplified setup of a scenario to generate a perturbation for a multi-target maneuver.

RESULTS KEY

- 🟥 Billboard
- — Road
- Unpert. trajectory
- Generation trajectory
- Test trajectories





BASELINE TECHNIQUE: Husheng Zhou, Wei Li, Zelun Kong, Junfeng Guo, Yuqun Zhang, Bei Yu, Lingming Zhang, and Cong Liu. 2020. DeepBillboard: systematic physical-world testing of autonomous driving systems. In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE '20). Association for Computing Machinery, New York, NY, USA, 347–358. https://doi.org/10.1145/3377811.3380422

This work was supported in part by NSF Award #1924777 and AFOSR Award #FA9550-21-1-0164.