Adversarial Attacks on Autonomous Driving with Physically Realizable Patterns

Prasanth Buddareddygari, MS Computer Science
Mentors: Prof. Yezhou Yang[1] & Prof. Yi (Max) Ren[2]
[1] School of Computing, Informatics, and Decision Systems Engineering & [2] School for Engineering of Matter, Transport, and Energy

Introduction

- Deep Reinforcement Learning (RL) policies are vulnerable to adversarial attacks.
- Threat to Autonomous Driving (AD) system.

Motivation

- Effectiveness of targeted attacks [1].
- Practicality through physical object manipulation [2].

Contributions

- Targeted Physically Realizable Attack (TPRA) - a static perturbation on object to make AD reach target.
- Ablation studies to find best attack parameters.
- Robustness of object to translation.

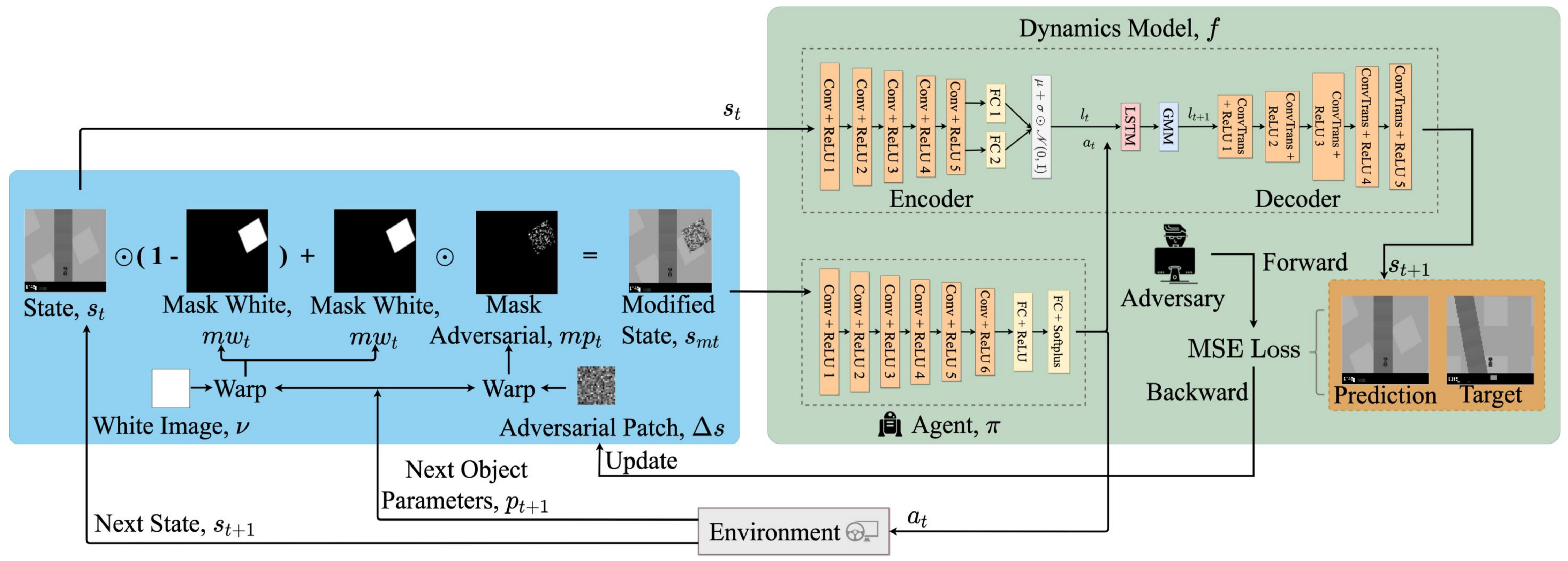


Fig 1: Illustration of Targeted Physical Adversarial Attack on CarRacing-V0 Environment.

Scenarios	Actions Error	Value Change (%)
Straight + Random	0.064	0
Left turn + Random	0.069	0
Right turn + Random	0.046	-10.72
Straight + Proposed	**0.126**	**-17.70**
Left turn + Proposed	**0.138**	**-32.26**
Right turn + Proposed	**0.062**	**-32.15**

Table 1: Quantitative Comparison with Baseline.

Robustness to Translation

- Moving towards track still facilitates attack.
- Moving away reduces attack effectiveness.

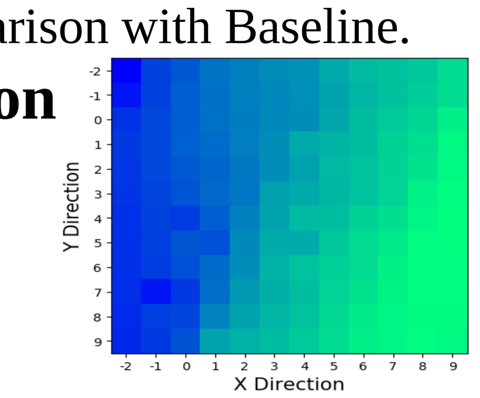


Fig 4: Attack Robustness to Object Position.

Targeted Physically Realizable Attack

- Agent dynamics are known.
- Environment dynamics are learned using a model.
- Target specified by attacker.

Attack Formulation

$$\min_{\|\Delta s\|_\infty \leq \epsilon} \sum_{t=1}^{T} \left\| s'_t - s'_{target} \right\|_2^2$$

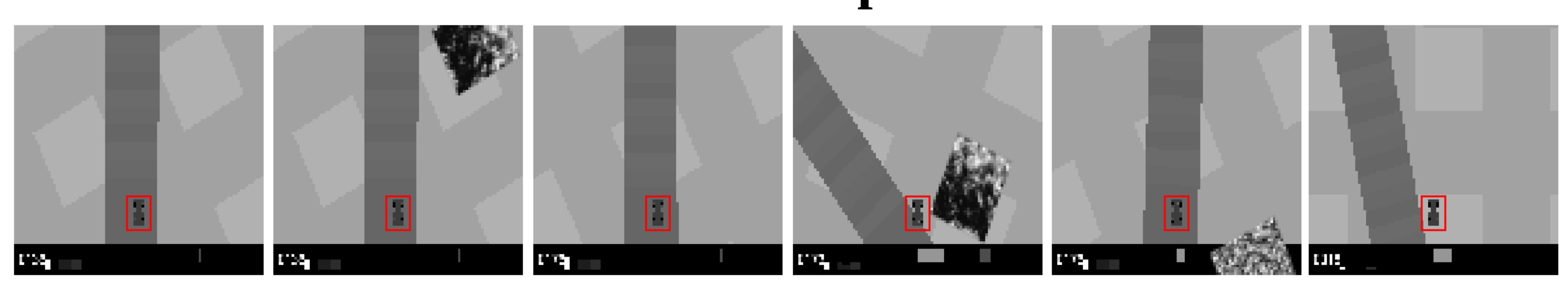
$$\begin{aligned}
s.t.\quad & a_t = \pi(s_{mt}), \\
& s'_{mt} = s'_t \odot (1 - mw_t) + mw_t \odot mp_t, \\
& mw_t = warp(\nu, p_t),\ mp_t = warp(\Delta s, p_t), \\
& p_t = \psi(\delta_t, \Phi), \\
& s'_{t+1} = f(s'_t, a_t), \\
& \delta'_{t+1} = g(\delta_t, a_t)
\end{aligned}$$

Where, π is the pretrained policy that outputs action a_t,
s_t is the image seen by agent without object,
$f(.,.)$ is the learned environment dynamics model,
$g(.,.)$ is the agent dynamics,
$\psi(\delta_t, \Phi)$ is the transition function,
ϵ is the strength of perturbation, and
Δs is the static perturbation that need to be found.

Experiments & Results

- OpenAI Gym's CarRacing-V0 environment
- Three driving scenarios
- Evaluation metrics – *Actions Error* and *Change in Value*

Baseline Comparison



No attack (t = 0) Ours (t = 0) No attack (t = T) Ours (t = T) Random (t = T) Target state

Fig 2: Targeted and Random Attacks on Straight Track Scenario.

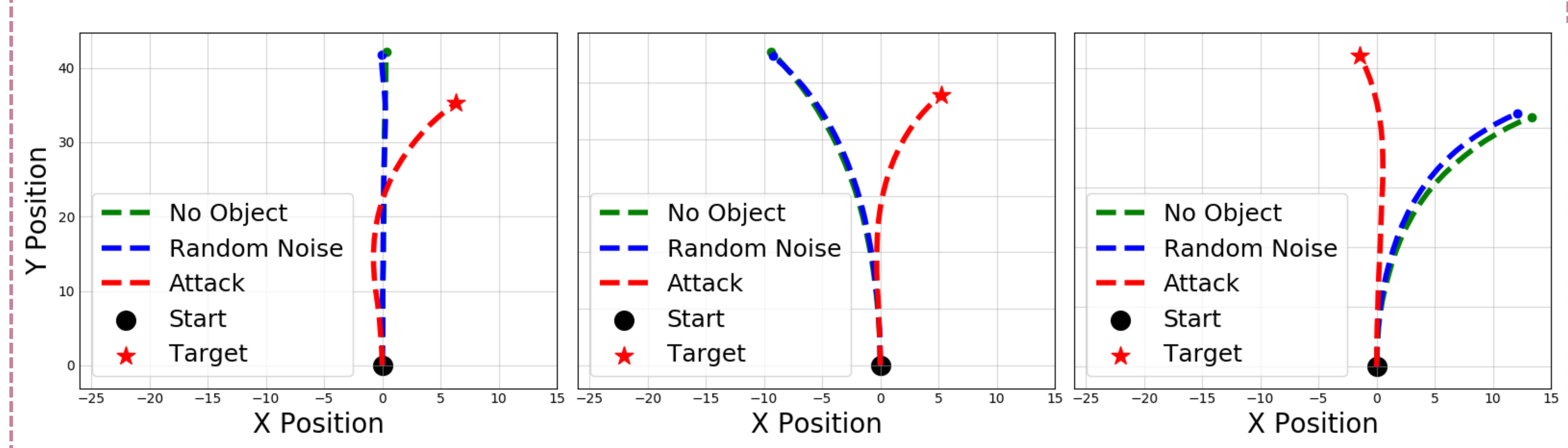


Fig 3: Trajectories in the Three Scenarios with No Attack, Random Attack, and Optimized Attacks.

Attack Strength vs Attack Length

Adversarial Bound ϵ	T = 15 Attack Loss	T = 15 Actions Error	T = 25 Attack Loss	T = 25 Actions Error	T = 30 Attack Loss	T = 30 Actions Error
0.1	0.091	0.064	0.090	0.064	0.088	0.063
0.3	0.088	0.078	0.087	0.069	0.085	0.066
0.5	0.086	0.113	0.077	0.107	0.083	0.070
0.9	0.081	0.125	**0.076**	**0.126**	0.078	0.093

Table 2: Ablation Studies on Attack Strength, ϵ vs Attack Length, T.

Conclusion & Future Work

- We presented TPRA by placing adversarial objects in the environment that can fool DNN policies to reach a target.
- Future work will study 3D and multi agent scenarios.

References

[1] Weng et al., "Toward evaluating robustness of deep reinforcement learning with continuous control," in ICLR, 2020.

[2] Z. Kong et al., "PhysGAN: Generating physical-world-resilient adversarial examples for autonomous driving," in CVPR, 2020.