Curiosity-Driven and Victim-Aware Adversarial Policies

Annual Computer Security Applications Conference (ACSAC) 2022

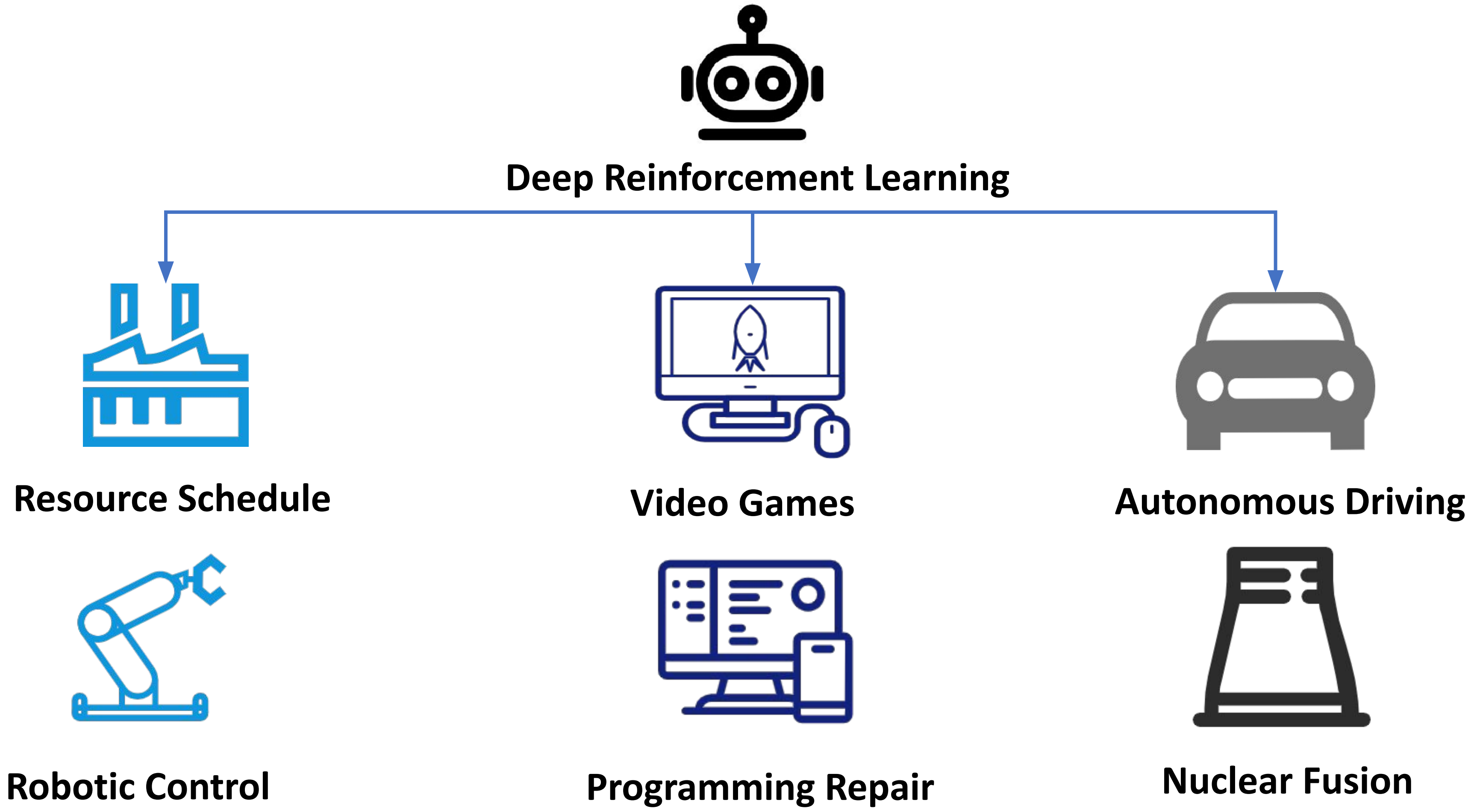
Chen Gong[1, 2], Zhou Yang[2], Yunpeng Bai[1], Jieke Shi[2], Arunesh Sinha[3], Bowen Xu[2], David Lo[2], Xinwen Hou[1], Guoliang Fan[1]

[1]Institute of Automation, Chinese Academy of Sciences

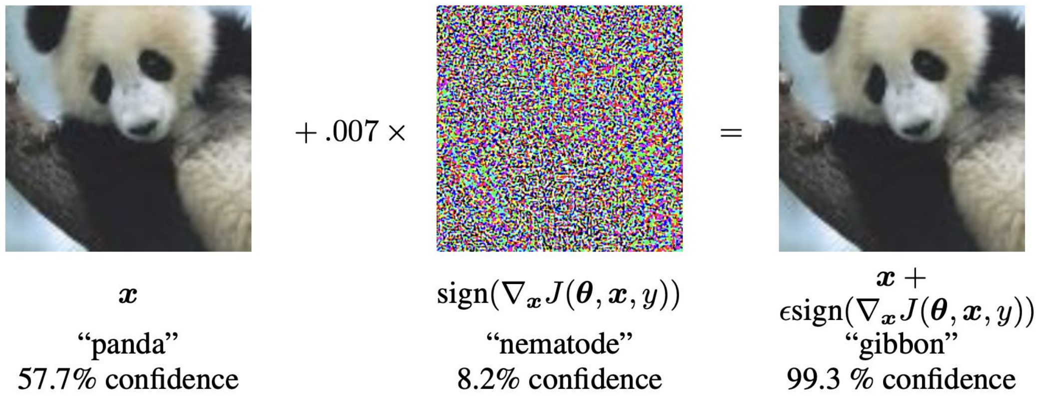
[2]School of Computing and Information Systems, Singapore Management University

[3]Management Science & Information Systems, Rutgers Business School, Rutgers University

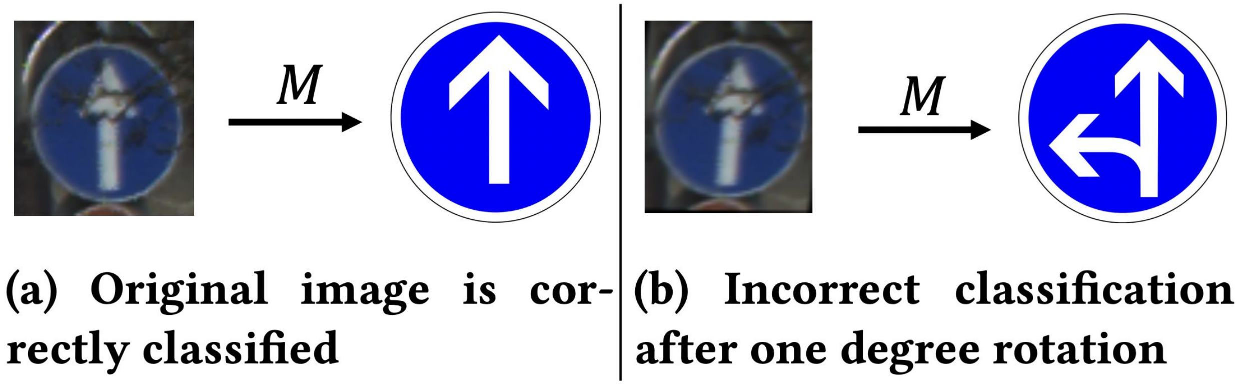
Success of DRL



DNNs are Vulnerable to Attacks



Adding imperceptible noise [1]

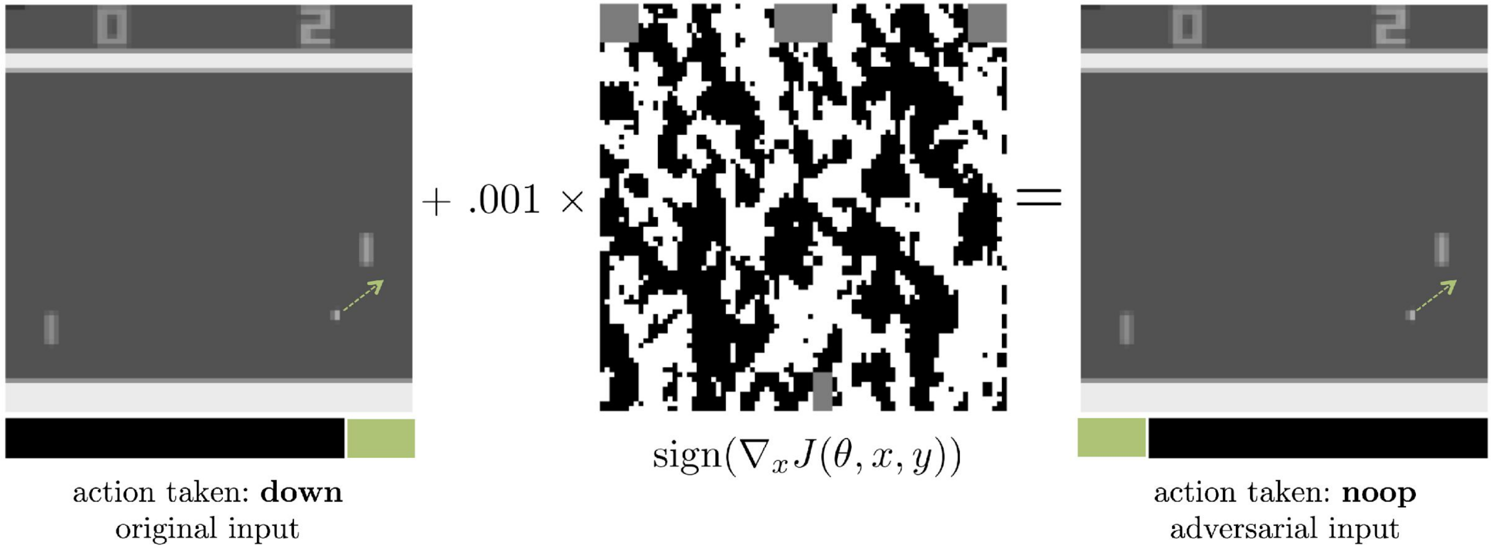


Rotation with small degrees [2]

Applying small perturbations or transformations on inputs can change DNNs' outputs.

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *ICLR 2015*.
[2] Gao, X., Saha, R. K., Prasad, M. R., & Roychoudhury, A. (2020). Fuzz testing based data augmentation to improve robustness of deep neural networks. *ICSE 2020.*

DRL agents are Vulnerable to Attacks, as well

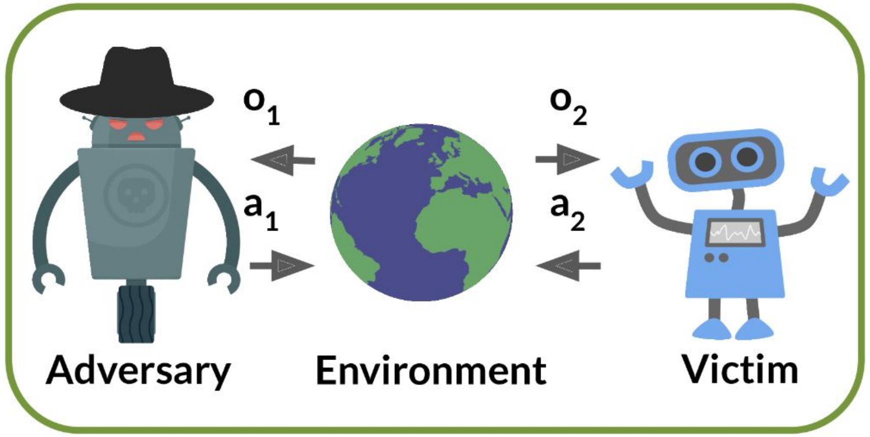


Adding invisible noise to the background image in games can fool DRL agents [1]

Unrealistic in practice as it requires an attacker to hack into the system.

[1] Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. *arXiv preprint*.

Realistic Threat Model of Attacking DRL



Realistic multi-agent threat model [1, 2]

Assumptions:

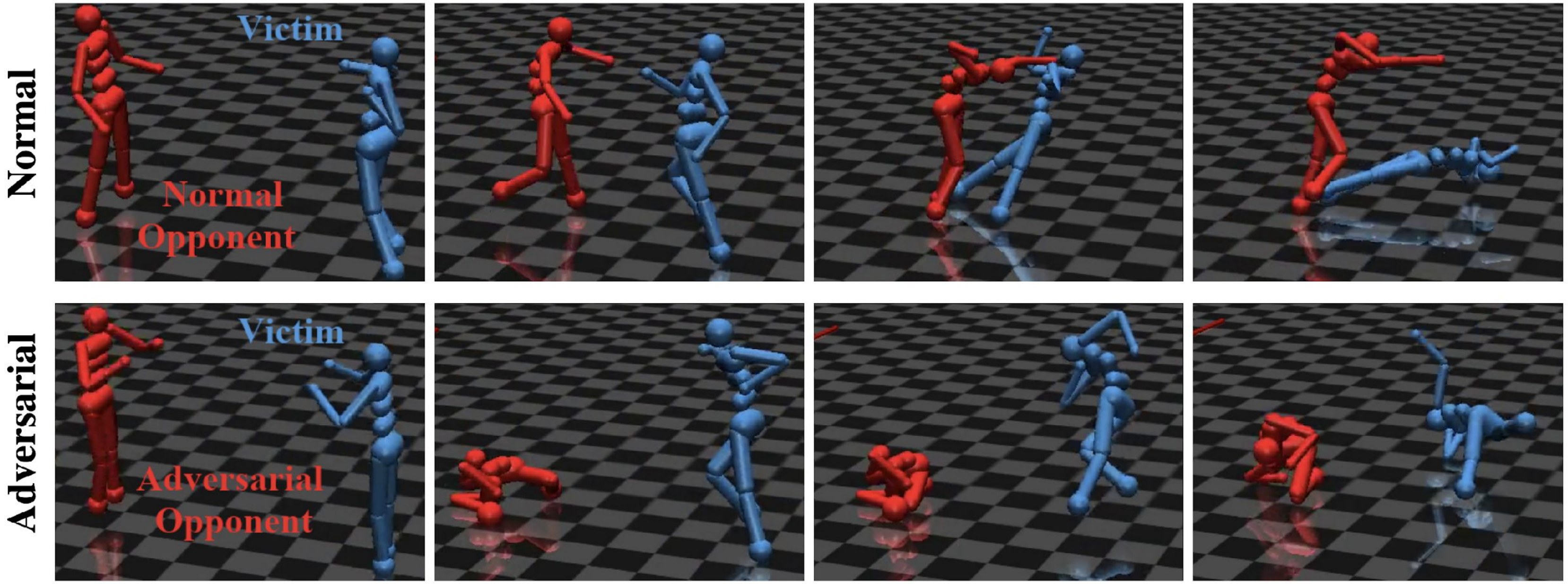
- ❑ Victim agent's parameters are unavailable
- ❑ The victim agent plays a fixed policy
- ❑ An attacker cannot make changes to game environments
- ❑ The attacker can control one agent

[1] Gleave, A., Dennis, M., Kant, N., Wild, C., Levine, S., & Russsell, S. (2020). Adversarial Policies: Attacking Deep Reinforcement Learning. *ICLR 2020*.

[2] Guo, W., Wu, X., Huang, S., & Xing, X. (2021). Adversarial policy learning in two-player competitive games. *ICML 2021*.

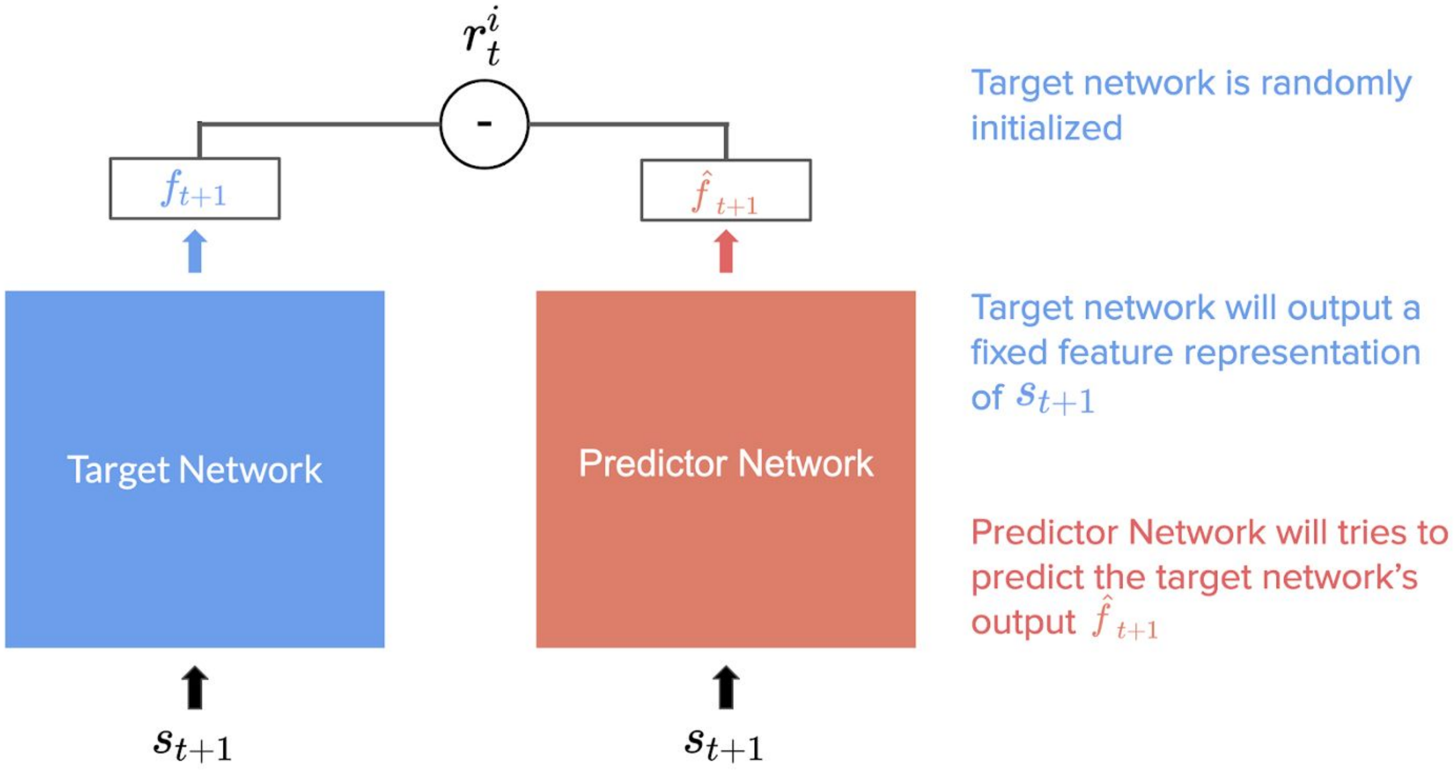
Insights from Prior Studies

- The attacker can fool the victim by taking uncommon actions to lead the game into unfamiliar states
- The victim can exhibit undesired sub-optimal behaviors in unfamiliar states [1, 2]



[1] Gleave, A., Dennis, M., Kant, N., Wild, C., Levine, S., & Russsell, S. (2020). Adversarial Policies: Attacking Deep Reinforcement Learning. *ICLR 2020*.
[2] Guo, W., Wu, X., Huang, S., & Xing, X. (2021). Adversarial policy learning in two-player competitive games. *ICML 2021*.

Curiosity Mechanism in DRL



Random Network Distillation (RND) [1]
image from [2]

$$r_t^i = \Big\| \underbrace{\hat{f}(s_{t+1})} - \underbrace{f(s_{t+1})} \Big\|_{\boxed{2}}^{\boxed{2}}$$

Prediction of the Feature Representation of next state

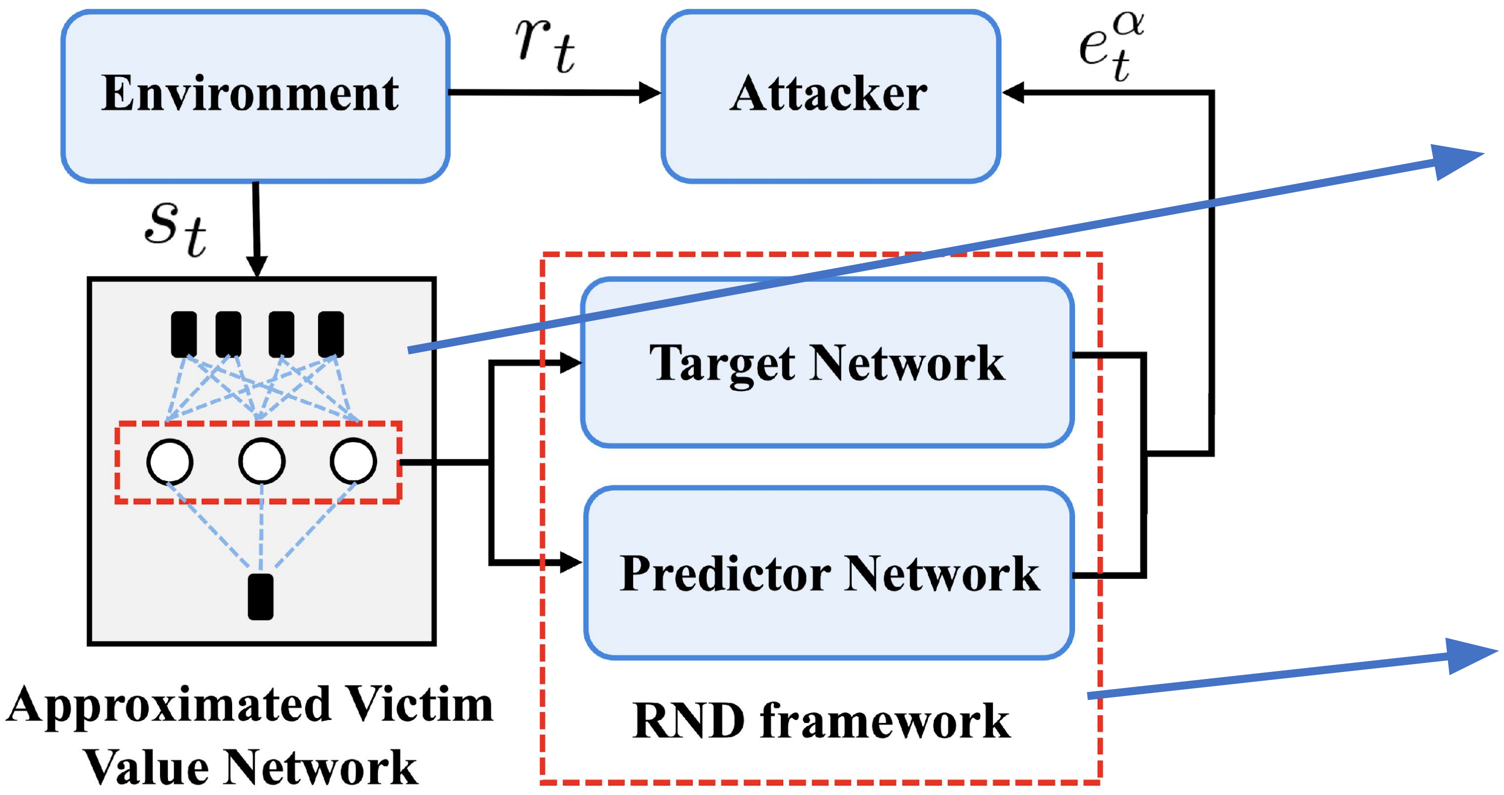
Target Feature Representation of next state

L2 norm (to output a scalar)

- An attacker's capable of exploring such unfamiliar states may increase the effectiveness of the attack
- Incorporating curiosity mechanisms in attacking can be promising.

[1] Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). Exploration by random network distillation. *ICLR 2018*.
[2] https://medium.com/data-from-the-trenches/curiosity-driven-learning-through-random-network-distillation-488ffd8e5938

Our Method of Training Adversarial Policies



Architecture of curiosity-driven and victim-aware adversarial policies

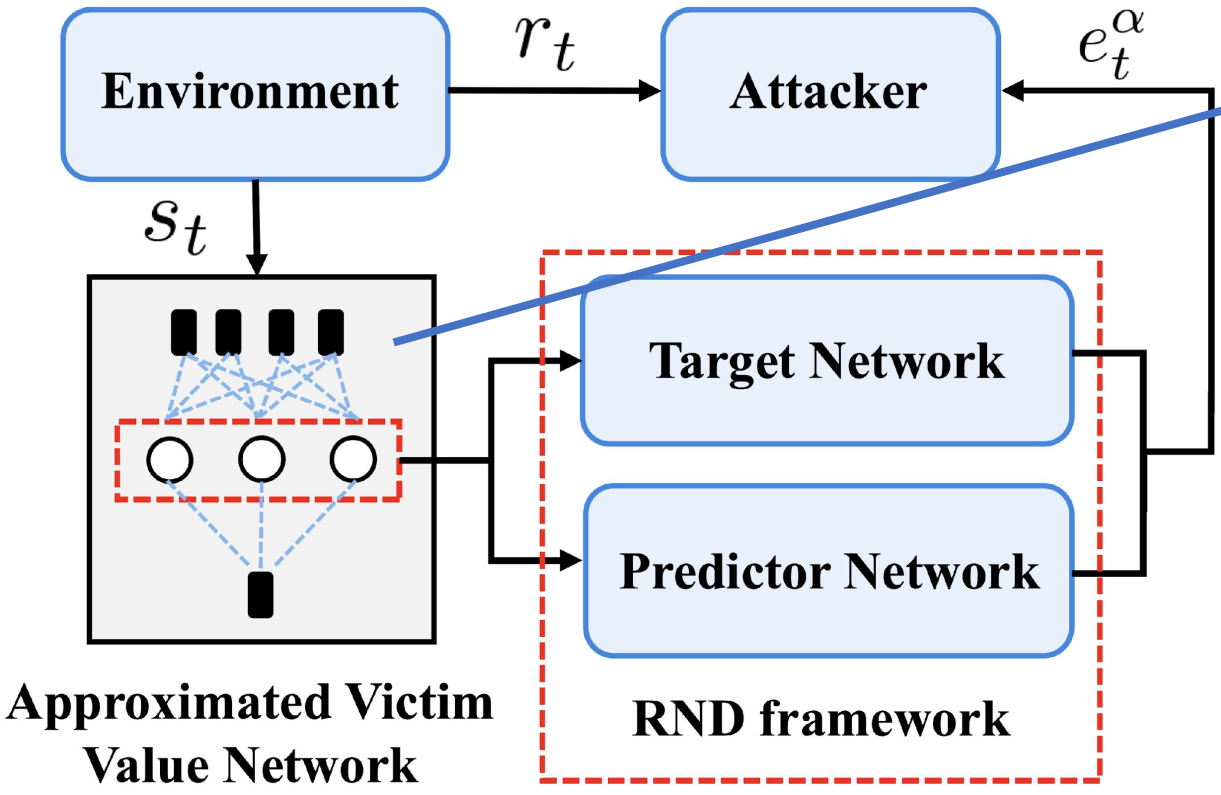
Victim-aware module

- Approximate victim state-value function, which allows the attacker to leverage the victim information

Curiosity-driven module

- RND framework to encourage the attacker to explore various states to uncover the vulnerabilities of the victim

Approximating Victim Information



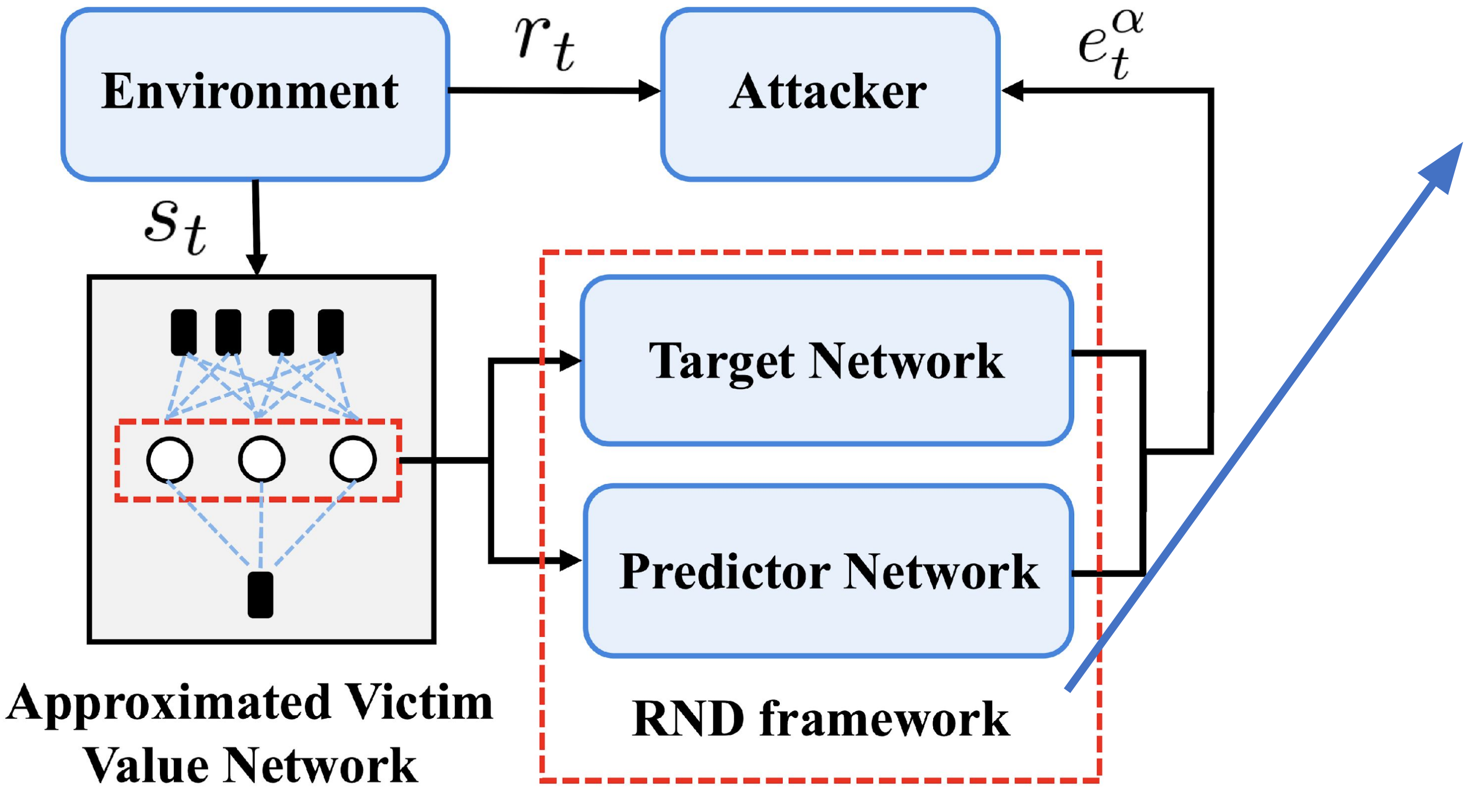
Victim-aware module

The outputs of the victim state-value function only depends on the state and the attacker policy, which allows to build a surrogate network to approximate it.

$$\underset{\theta^\nu}{\operatorname{argmin}} \left\| V_{\pi^\alpha}^{\nu}(s_t) - \left(R^\nu(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}\sim P}[V_{\pi^\alpha}^{\nu}(s_{t+1})]\right)\right\|^2$$

training objective when approximating victim state-value function

Curiosity-Driven Exploration



Curiosity-driven module

- We feed the output of a hidden layer of victim state-value function into RND framework
- The output expected mean square error of RND is utilized as the intrinsic reward

$$e^{\alpha} = \left\| \hat{g}_{\theta_{\hat{g}}}(\phi(s)) - g_{\theta_g}(\phi(s)) \right\|^2$$

the intrinsic reward to drive the attacker's exploration

Adversarial Policies Training

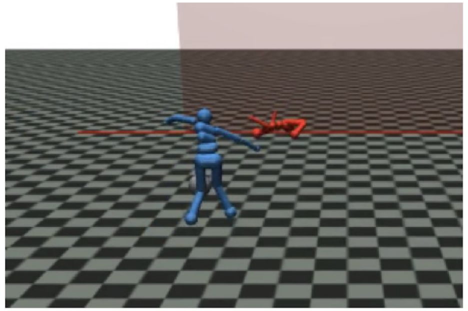
$$\arg\max_{\theta} \mathbb{E}_{(a_t^\alpha, s_t) \sim \pi_{\text{old}}^\alpha} \Big[\min\left(\text{clip}(\rho_t, 1-\epsilon, 1+\epsilon) A_t^\alpha, \rho_t A_t^\alpha\right)$$
$$- \min\left(\text{clip}(\rho_t, 1-\epsilon, 1+\epsilon) A_t^\nu, \rho_t A_t^\nu\right)\Big],$$
$$\rho_t = \frac{\pi_\theta^\alpha(a_t^\alpha | s_t)}{\pi_{\text{old}}^\alpha\left(a_t^\alpha | s_t\right)}, A_t^\nu = A_{\pi_{\text{old}}^\alpha}^\nu(a_t^\alpha, s_t),$$
$$A_t^\alpha = A_{\pi_{\text{old}}^\alpha}^\alpha(a_t^\alpha, s_t) + \lambda A_{\pi_{\text{old}}^\alpha}^{\alpha,\text{ins}}(a_t^\alpha, s_t)$$

Training objective of our adversarial policy

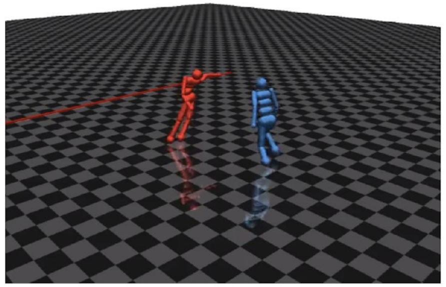
Follows the technique in [1] to train adversarial policy using the PPO algorithm [2]

[1] Guo, W., Wu, X., Huang, S., & Xing, X. (2021). Adversarial policy learning in two-player competitive games. *ICML 2021*.
[2] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

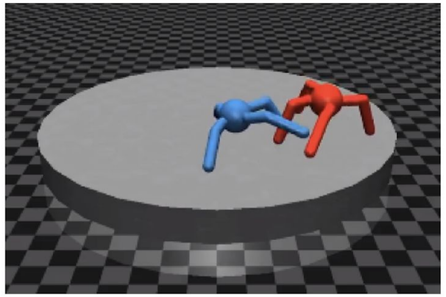
Experiment Setup



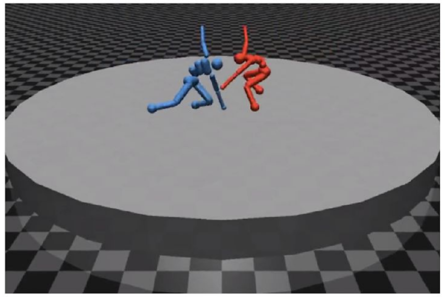
(a) Kick-And-Defend



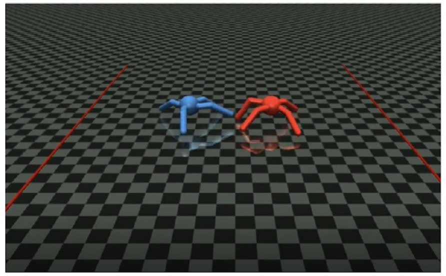
(b) You-Shall-Not-Pass



(c) Sumo-Ants



(d) Sumo-Humans



(e) Run-To-Goal-Ants

(f) StarCraft II

Experiment games

Winning rate:

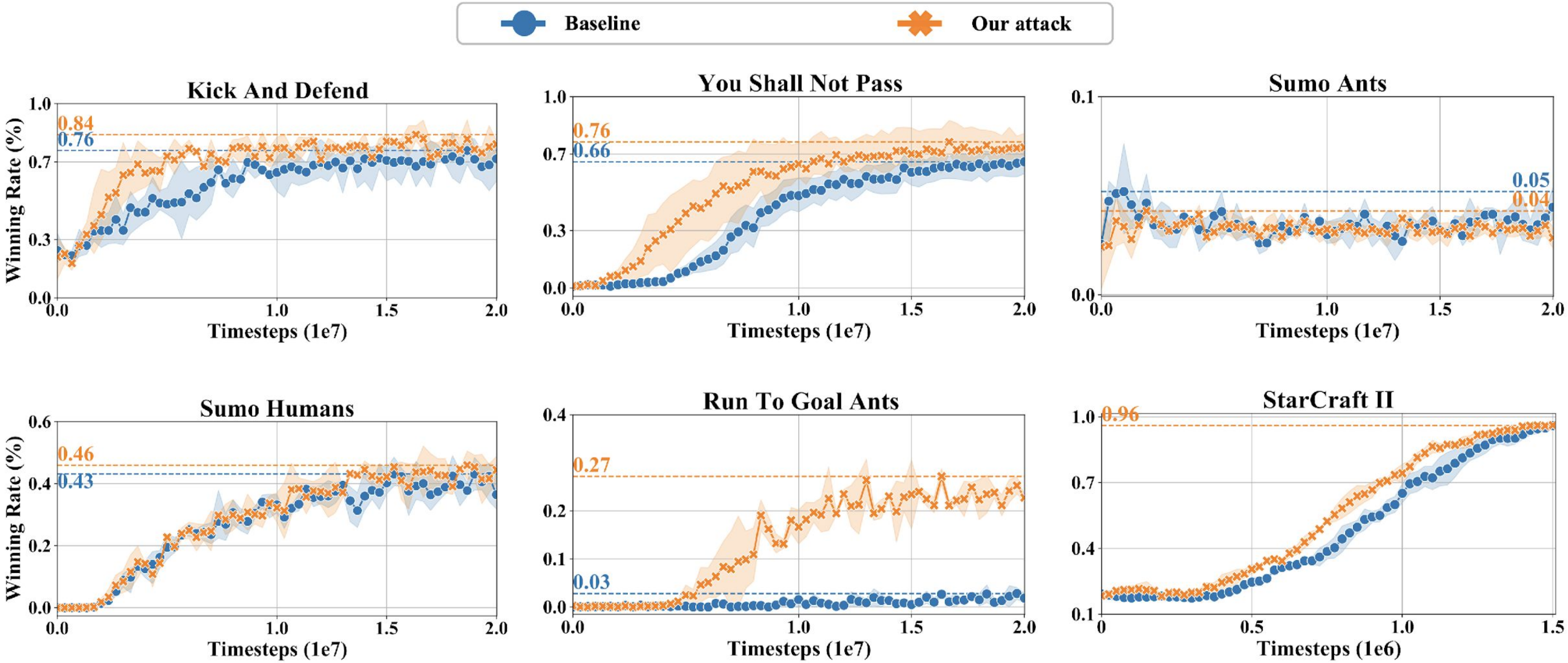
$$\frac{\text{the number of winning rounds}}{\text{the number of total rounds}} \times 100\%$$

Non-loss rate:

$$\frac{\text{the number of non-loss rounds}}{\text{the number of total rounds}} \times 100\%$$

Evaluation Metrics

Winning Rates of Adversarial Policies



The average winning rates of our method is 7.3% higher than that of the state-of-the-art method.

Adversarial Training

Games	Ours (%)		Baseline (%)		Regular (%)	
	Before	After	Before	After	Before	After
K-A-D	16.0	51.0	23.0	41.0	48.0	30.0
Y-S-N-P	18.0	76.0	28.0	49.0	55.0	37.0
S-A	89.0	89.0	94.0	96.0	55.0	50.0
S-H	58.0	75.0	57.0	56.0	67.0	41.0
R-T-G-A	70.0	72.0	97.0	96.0	58.0	59.0
SC II	3.0	76.0	2.0	79.0	68.0	94.0

Non-loss rates of victim agents

- Re-training a victim against a fixed adversarial policy, which is called adversarial training of DRL
- Adversarial training helps defend against adversarial policies

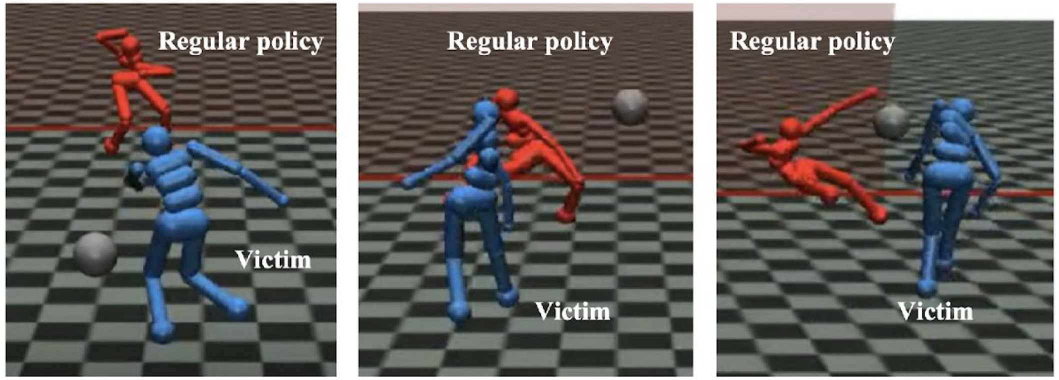
Ablation Study

Games	Baseline (%)	Victim-*un*aware (%)	Victim-aware (%)
Kick-And-Defend	76.0	79.0 **(+ 3.0)**	84.0 **(+ 8.0)**
You-Shall-Not-Pass	66.0	71.0 **(+ 5.0)**	76.0 **(+ 10.0)**
Sumo-Ants	5.0	4.0 **(- 1.0)**	4.0 **(- 1.0)**
Sumo-Humans	43.0	44.0 **(+ 1.0)**	46.0 **(+ 3.0)**
Run-To-Goal-Ants	3.0	10.0 **(+ 7.0)**	27.0 **(+ 24.0)**
StarCraft II	96.0	96.0 **(+ 0.0)**	96.0 **(+ 0.0)**

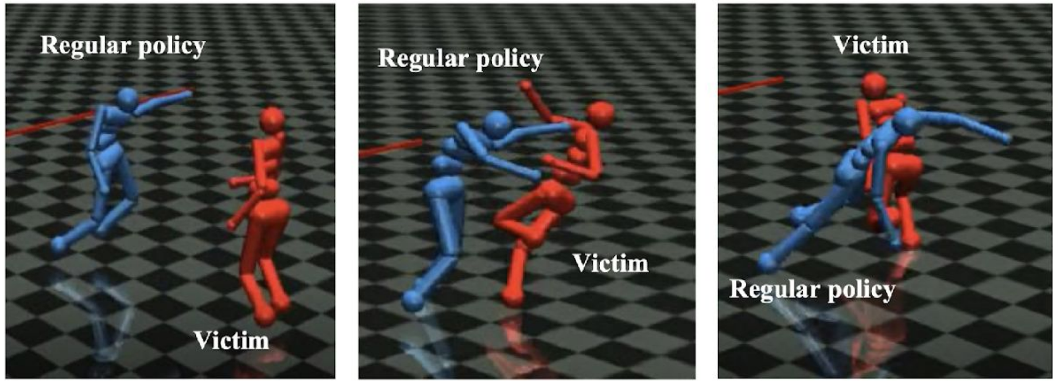
Winning rates of adversarial policies

- The victim-aware exploration leads to stronger adversarial policies, which improves the averaged performance by 5.8% compared to the victim-unaware method
- Our proposed method does not improve the state-of-the-art approach solely using the curiosity mechanism

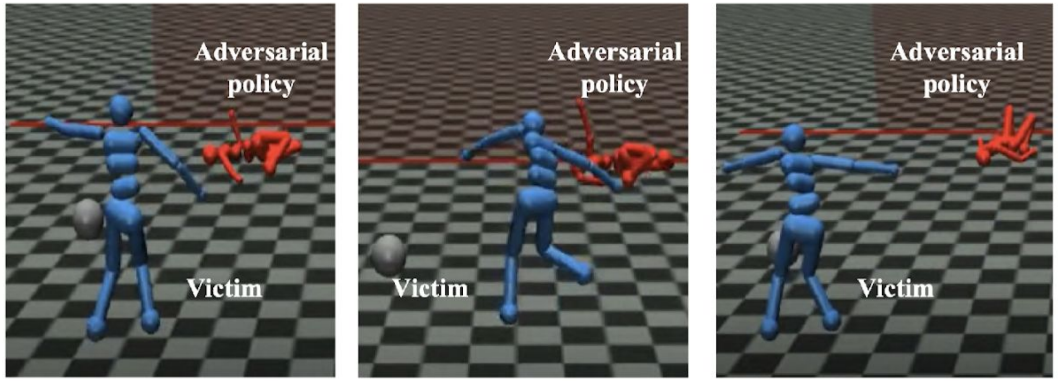
Illustrative Examples



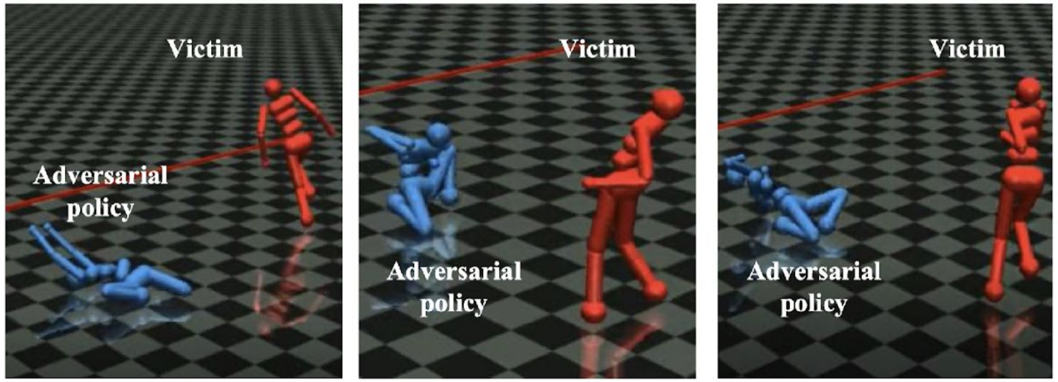
(a) Kick-And-Defend, the victim agent against a regular policy.



(b) You-Shall-Not-Pass, the victim agent against a regular policy.



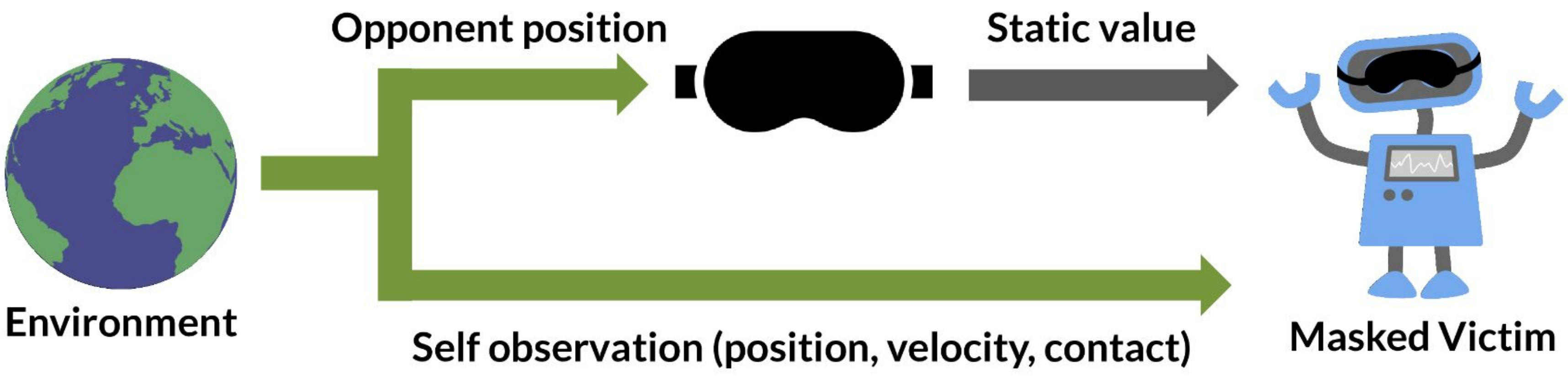
(c) Kick-And-Defend, the victim agent against our adversarial policy.



(d) You-Shall-Not-Pass, the victim agent against our adversarial policy.

Compared with the behaviors of regular agents learning to run, kick or block, the adversarial agents never stand up and lie under the ground in some strange poses, but still win the games.

Observation Masking

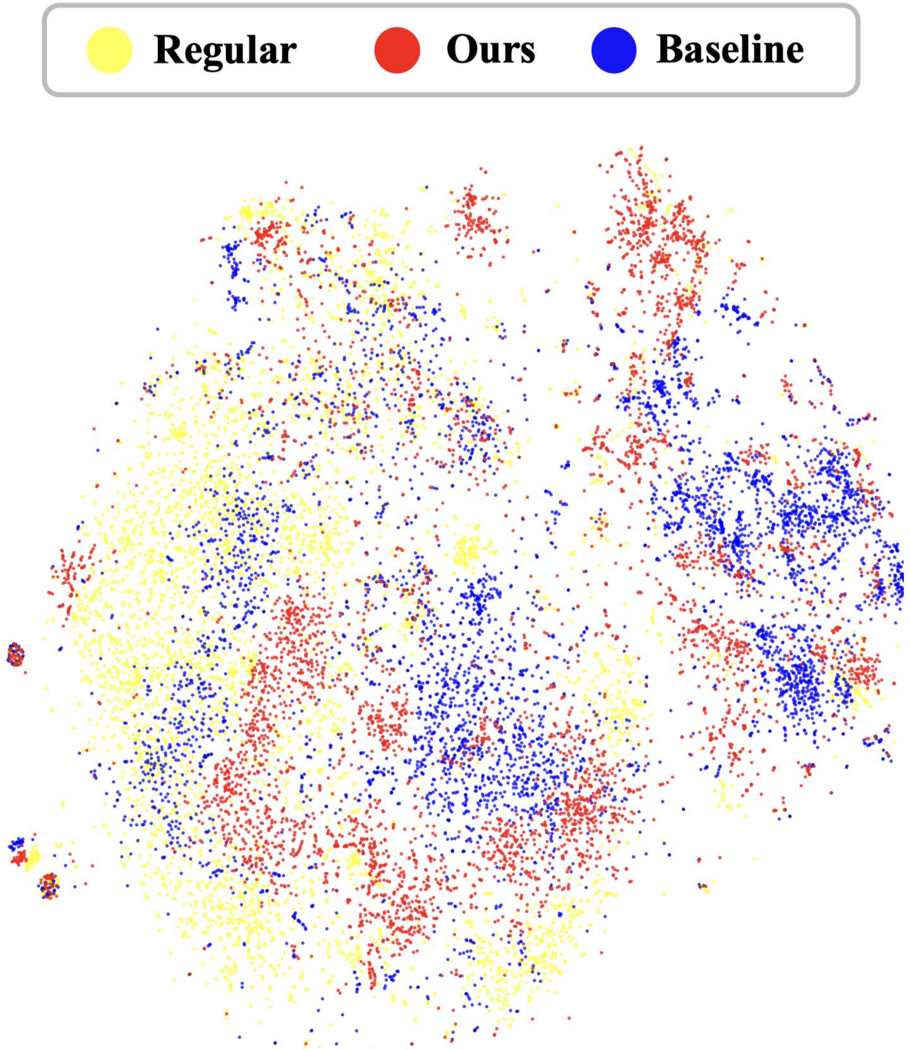


Games	Ours (%)		Baseline (%)	
	Before	After	Before	After
Kick-And-Defend	16.0	90.0	23.0	94.0
You-Shall-Not-Pass	18.0	40.0	28.0	99.0
Sumo-Ants	89.0	91.0	94.0	93.0
Sumo-Humans	58.0	70.0	57.0	76.0
Run-To-Goal-Ants	70.0	73.0	97.0	99.0
StarCraft II	3.0	14.0	2.0	18.0

non-loss rates of victims

Masking observation can defend against adversarial policies, suggesting that our adversarial policy succeeds by manipulating the victim's observations but not physically interfering with the victim

Policy Network Activation Analysis



The activation triggered by our adversarial policies distributes differently from the other two, indicating that our method uncovers different vulnerabilities.

Contributions and Future Work

Contributions:

- We present a novel curiosity-driven and victim-aware approach to attack DRL agents in two-player games
- The obtained adversarial policies outperform the current state-of-the-art results

Future Work

- Explore adversarial policies beyond two-player environments, the complexity of which exponentially increases with the number of players
- Develop more effective techniques to defend against adversarial policies

Hope it inspires!

Questions are welcome 😊!

Paper

Artifact