From Pixels to Policies: Securing Multi-Agent Systems Against Adversarial Attacks

György Dán

2 October 2025

AI/ML ubiquitous in safety critical systems

Communication networks

Smart grids

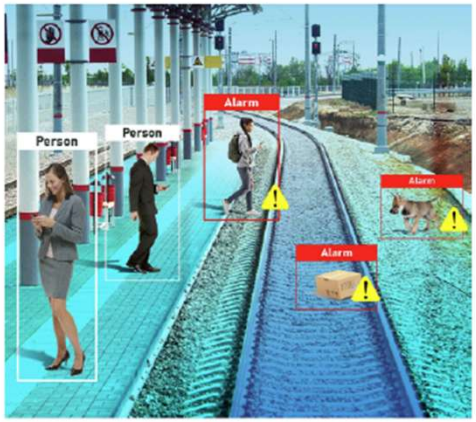
Healthcare

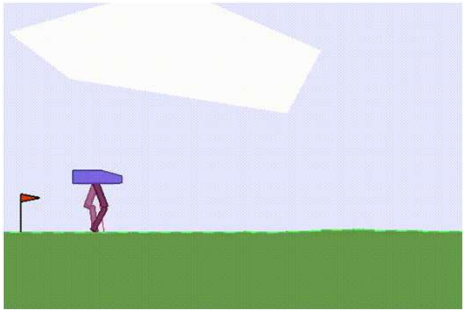
Transportation systems

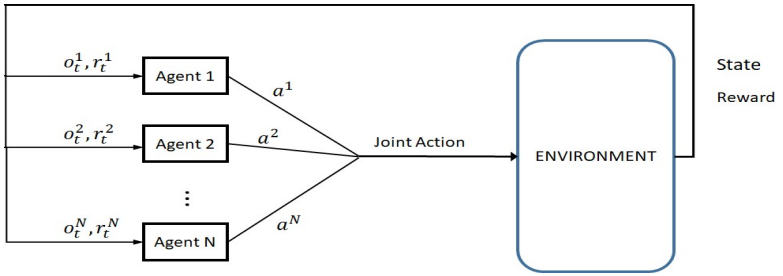
Smart cities and buildings

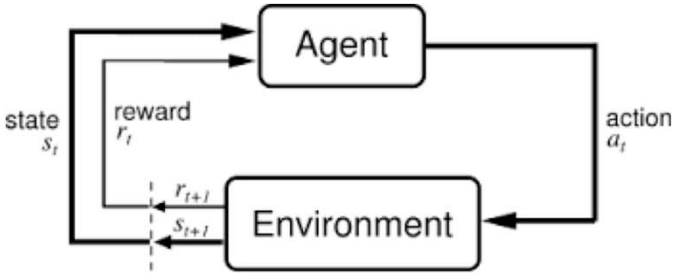
Manufacturing

ML-enabled Perception and Control



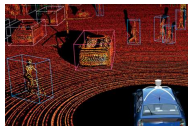
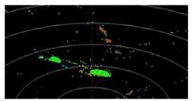


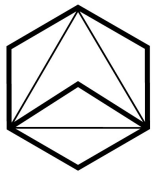




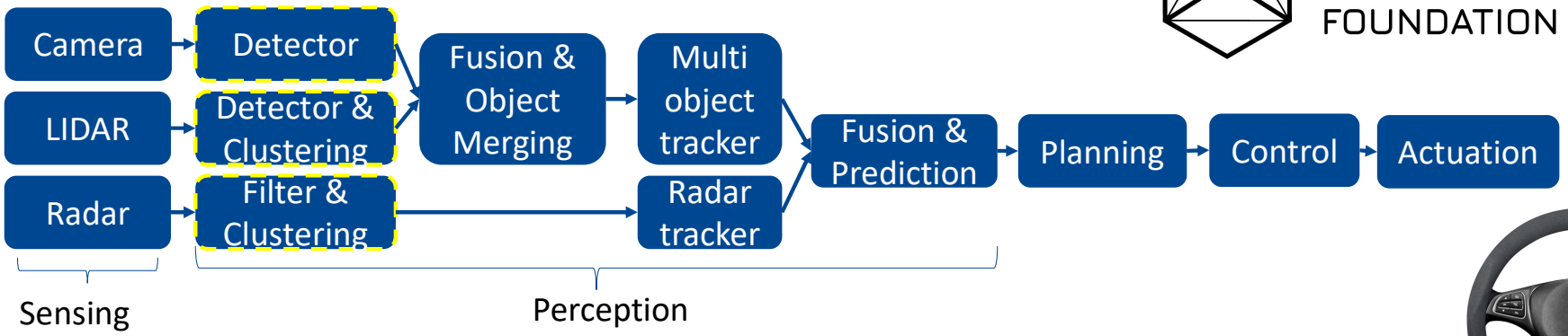
Samvelyan et al, “The StarCraft Multi-Agent Challenge”, NIPS Workshops, 2019

From Perception to Control: End-to-end or Modular

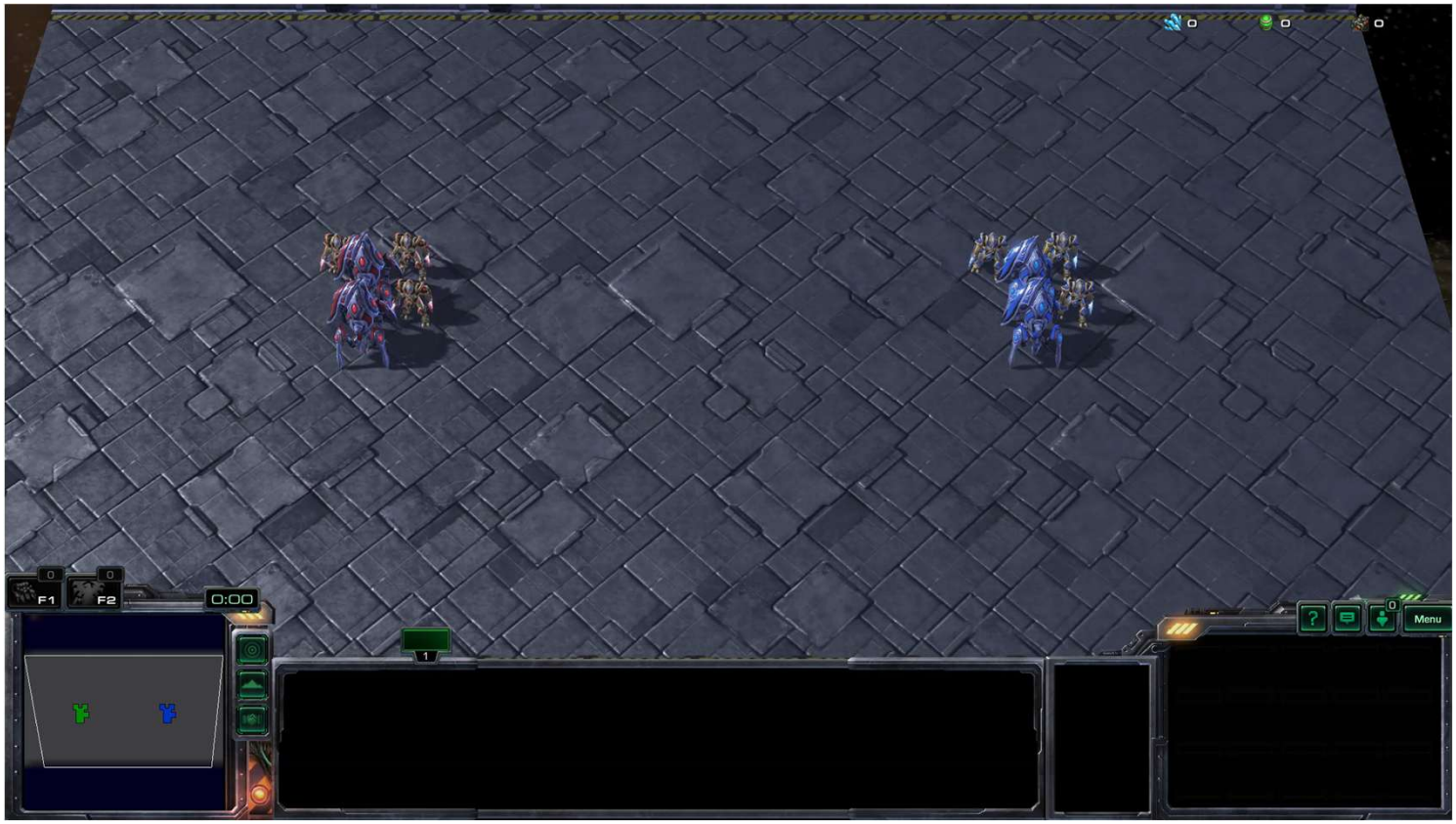


THE AUTOWARE FOUNDATION



TESLA

MARL performs well in many areas



96% win rate
using QMIX

CPS Example

Distributed Voltage Control in AC Microgrids

- Scenario
  - Grid connected inverter based resources
- Objective
  - Minimize frequency and voltage deviation
- Challenge
  - Complex dynamics
  - May not be fully known

https://blog.norcalcontrols.net/power-plant-controls-for-grid-following-grid-forming-ibrs

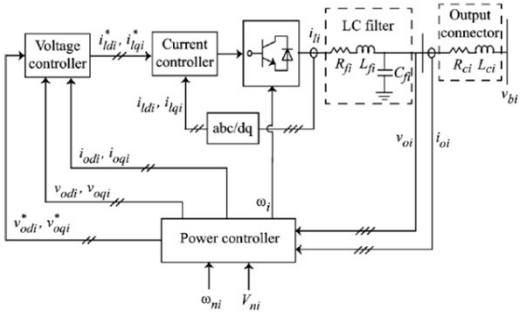
- Hierarchical control
  - Primary droop control
  - Secondary control
    (voltage, frequency, droop control gain)
    - > *PI*
    - > *MPC*
    - > *cMARL*

A. Bidram, et al. “Distributed cooperative secondary control of microgrids using feedback linearization,” IEEE Trans. on Power Systems, vol. 28, no. 3, pp. 3462–3470, 2013
A. Bidram, et al. “A multiobjective distributed control framework for islanded AC microgrids,” IEEE Trans. on Ind. Informatics, vol. 10, no. 3, pp. 1785–1798, 2014
G. Lou, et al. “Distributed MPC-based secondary voltage control scheme for autonomous droop-controlled microgrids,” IEEE Trans. on Sustainable Energy, vol. 8, no. 2, pp. 792–804, 2017
D. Chen, et al “Powernet: Multi-agent deep reinforcement learning for scalable powergrid control,” IEEE Trans. on Power Systems, vol. 37, no. 2, pp. 1007–1017, 2022

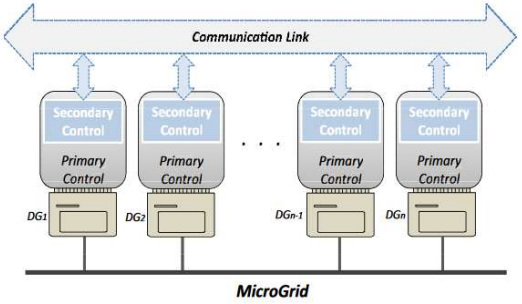
CPS Example

Distributed Voltage Control in AC Microgrids

- Scenario
  - Grid connected inverter based resources
- Objective
  - Minimize frequency and voltage deviation
- Challenge
  - Complex dynamics
  - May not be fully known



- Hierarchical control
  - Primary droop control
  - Secondary control
    (voltage, frequency, droop control gain)
    - > *PI*
    - > *MPC*
    - > *cMARL*

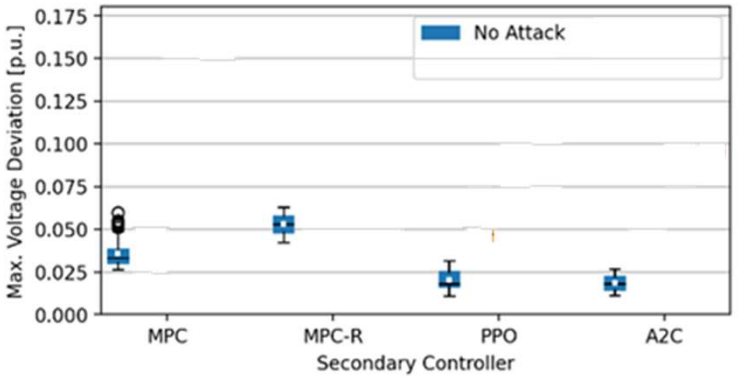


A. Bidram, et al. "Distributed cooperative secondary control of microgrids using feedback linearization," IEEE Trans. on Power Systems, vol. 28, no. 3, pp. 3462–3470, 2013
A. Bidram, et al. "A multiobjective distributed control framework for islanded AC microgrids," IEEE Trans. on Ind. Informatics, vol. 10, no. 3, pp. 1785–1798, 2014
G. Lou, et al. "Distributed MPC-based secondary voltage control scheme for autonomous droop-controlled microgrids," IEEE Trans. on Sustainable Energy, vol. 8, no. 2, pp. 792–804, 2017
D. Chen, et al "Powernet: Multi-agent deep reinforcement learning for scalable powergrid control," IEEE Trans. on Power Systems, vol. 37, no. 2, pp. 1007–1017, 2022

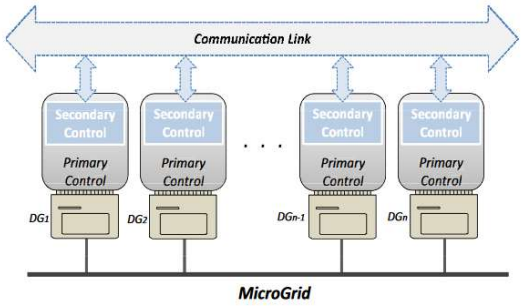
CPS Example

Distributed Voltage Control in AC Microgrids

- Scenario
  - Grid connected inverter based resources
- Objective
  - Minimize frequency and voltage deviation
- Challenge
  - Complex dynamics
  - May not be fully known



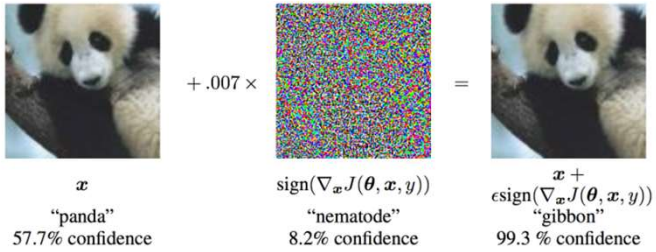
- Hierarchical control
  - Primary droop control
  - Secondary control
    (voltage, frequency, droop control gain)
    - > *PI*
    - > *MPC*
    - > *cMARL*



A. Bidram, et al. "Distributed cooperative secondary control of microgrids using feedback linearization," IEEE Trans. on Power Systems, vol. 28, no. 3, pp. 3462–3470, 2013
A. Bidram, et al. "A multiobjective distributed control framework for islanded AC microgrids," IEEE Trans. on Ind. Informatics, vol. 10, no. 3, pp. 1785–1798, 2014
G. Lou, et al. "Distributed MPC-based secondary voltage control scheme for autonomous droop-controlled microgrids," IEEE Trans. on Sustainable Energy, vol. 8, no. 2, pp. 792–804, 2017
D. Chen, et al "Powernet: Multi-agent deep reinforcement learning for scalable powergrid control," IEEE Trans. on Power Systems, vol. 37, no. 2, pp. 1007–1017, 2022

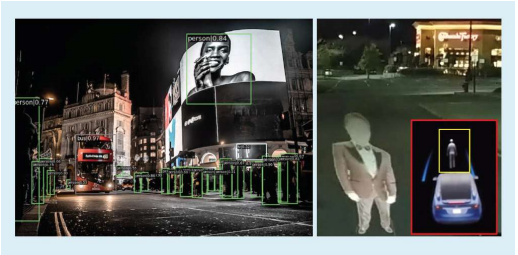
ML Models are Vulnerable in Many Ways

Digital attacks



Physically realizable attacks

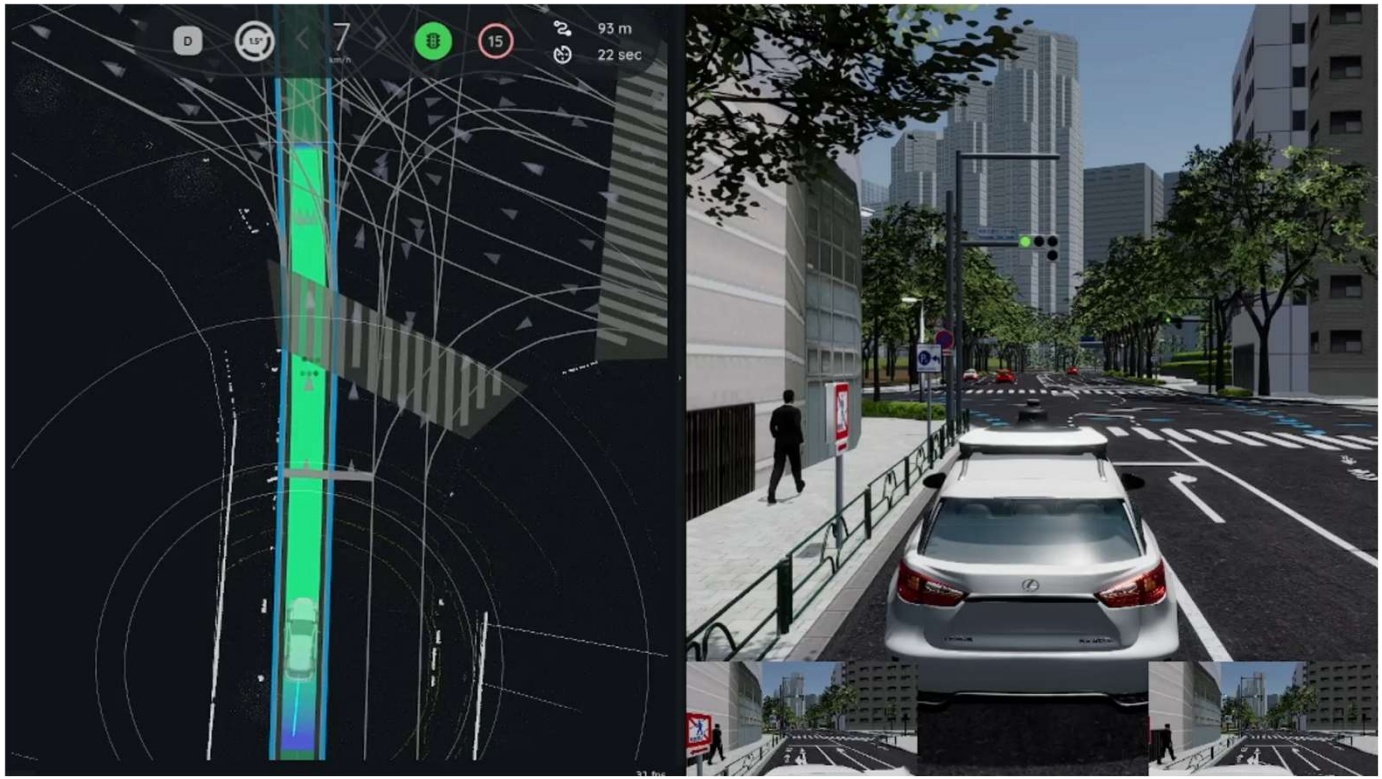
Phantom attacks



Patch attacks

Nassi et al, "Protecting Autonomous Cars from Phantom Attacks", CACM, 2023
Goodfellow et al, "Explaining and Harnessing Adversarial Examples", ICLR 2014
Byrd et al, "SpaNN: Detecting Multiple Adversarial Patches on CNNs by Spanning Saliency Thresholds" IEEE SaTML, 2025

Attacks can Compromise Entire Pipeline



As well as cMARL

CPS Example

Distributed Voltage Control in AC Microgrids

under attack

Objective:

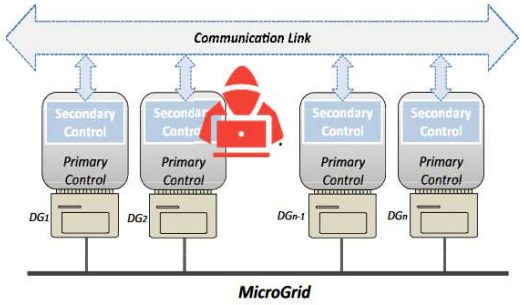
- Minimize frequency and voltage deviation

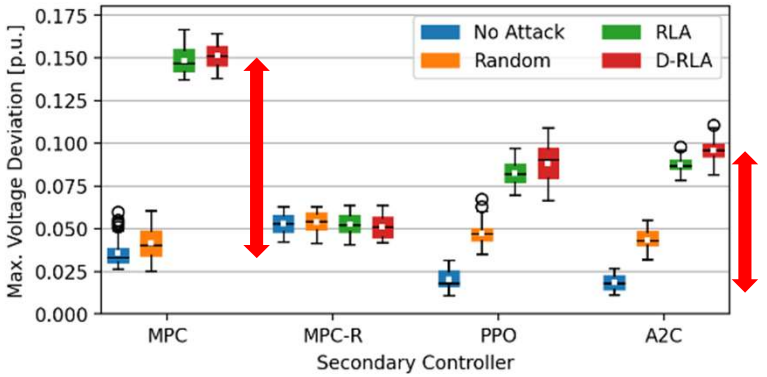
Environment

- Complex dynamics
- May not be fully known

Hierarchical control

- Primary droop control
- Secondary control (power set point, droop control gain)
  - PI
  - MPC
  - cMARL

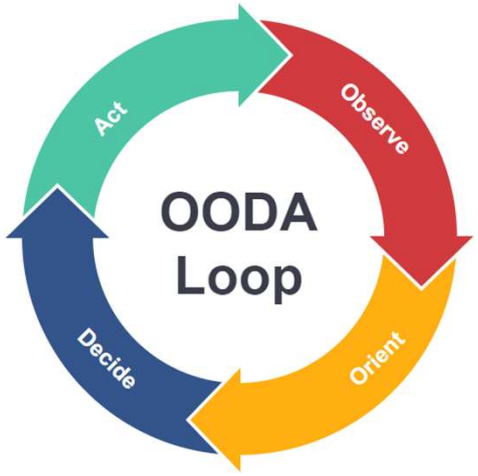


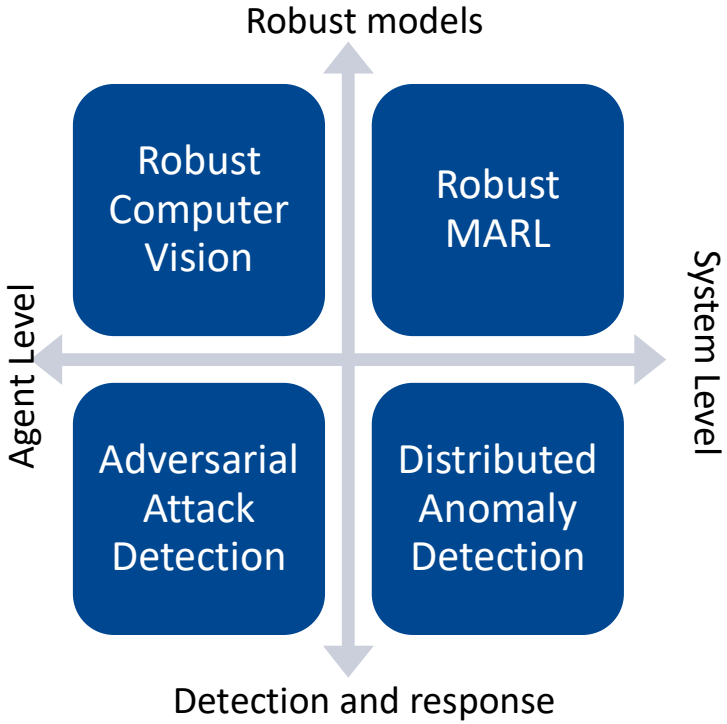


Shereen et al, ``Adversarial Robustness of Multi-agent Reinforcement Learning Secondary Control of Islanded Inverter-based AC Microgrids,'' in Proc. of IEEE SmartGridComm, Oct. 2023

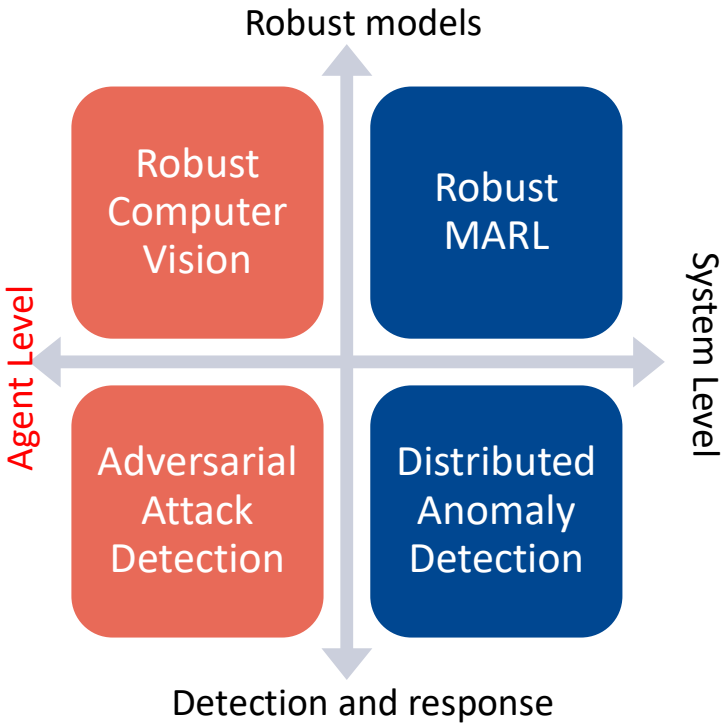
Securing ML-enabled Multi-agent Systems

- Vulnerability assessment
  - Threat model
  - Dynamic/Adaptive adversaries
- Where to defend
  - Agent level
  - System level
- When to defend
  - Design time
    - Formal verification
    - Robust training
    - Sensor fusion
  - Runtime
    - Anomaly detection
    - Response

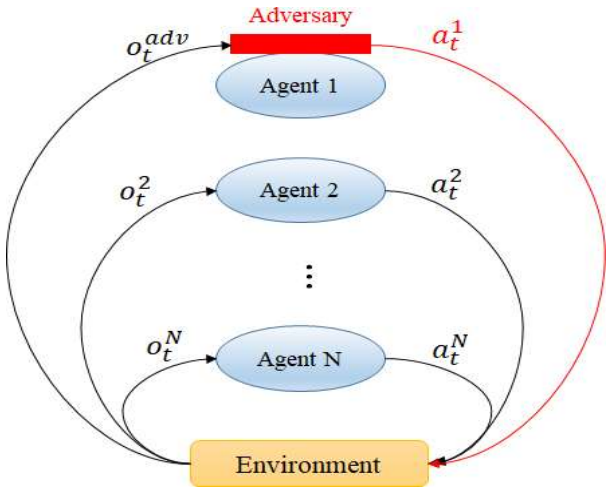




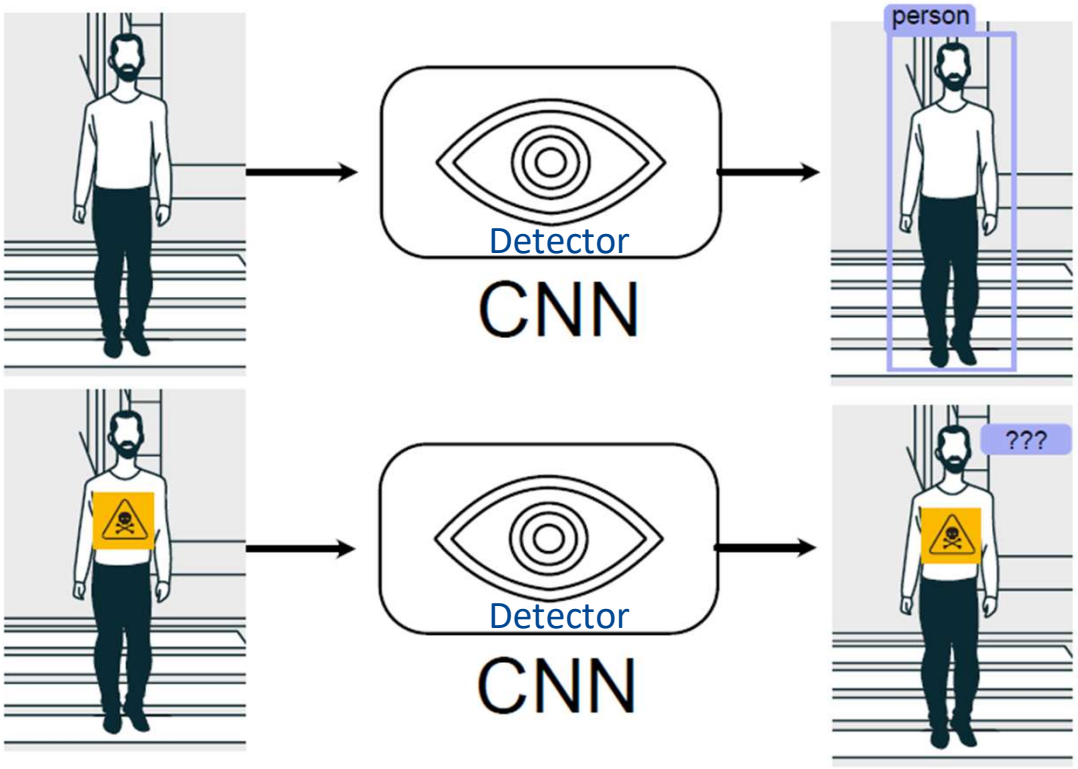
Agenda


Robust models
Robust Computer Vision
Robust MARL
Adversarial Attack Detection
Distributed Anomaly Detection
Agent Level
System Level
Detection and response


Adversarial Patch
Person
Person
Alarm
Alarm

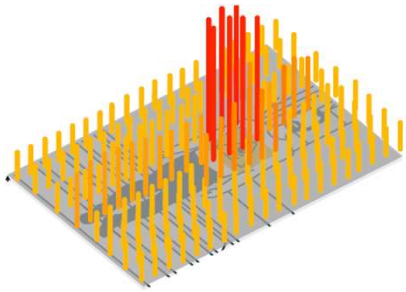

Adversary
o_t^{adv}
a_t^1
Agent 1
o_t^2
Agent 2
a_t^2
o_t^N
Agent N
a_t^N
Environment

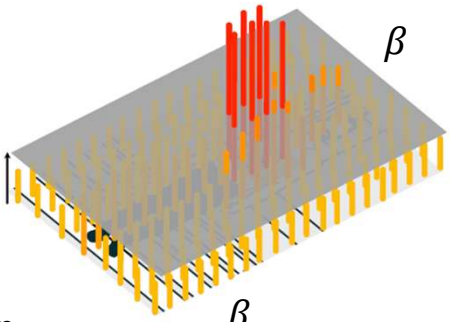
Patch Attack Detection

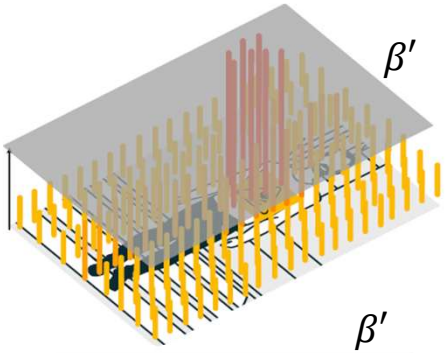


- Detection problem: Is there an adversarial patch in the image?

Existing approaches to detection and recovery



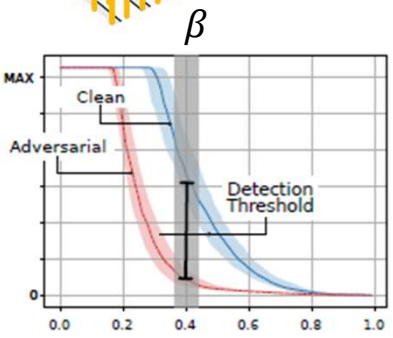


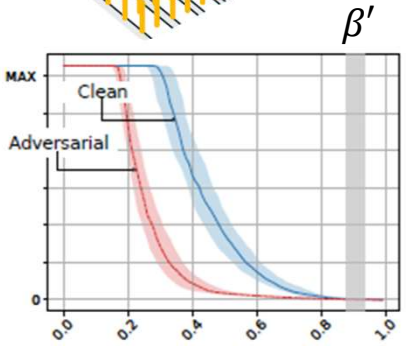


Transform image to feature domain

- Embedding produced by CNN
- Entropy

Threshold to construct saliency map





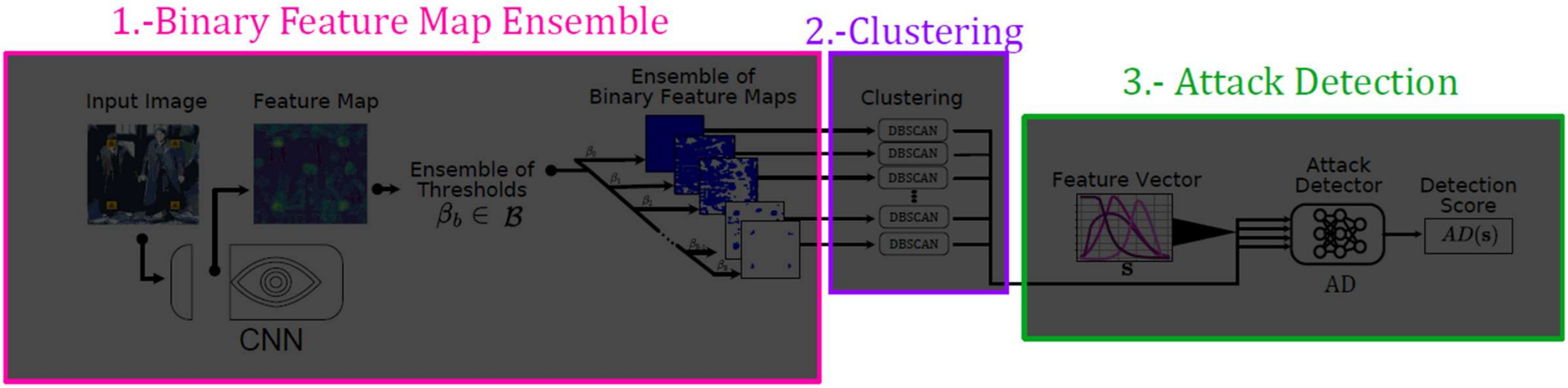
Shortcomings

- Fixed threshold easy to bypass
- Assume single square patch

Han et al. “Real-Time Robust Video Object Detection System Against Physical-World Adversarial Attacks”. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (2023)
Tarchoun et al. “Jedi: Entropy-Based Localization and Removal of Adversarial Patches”. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

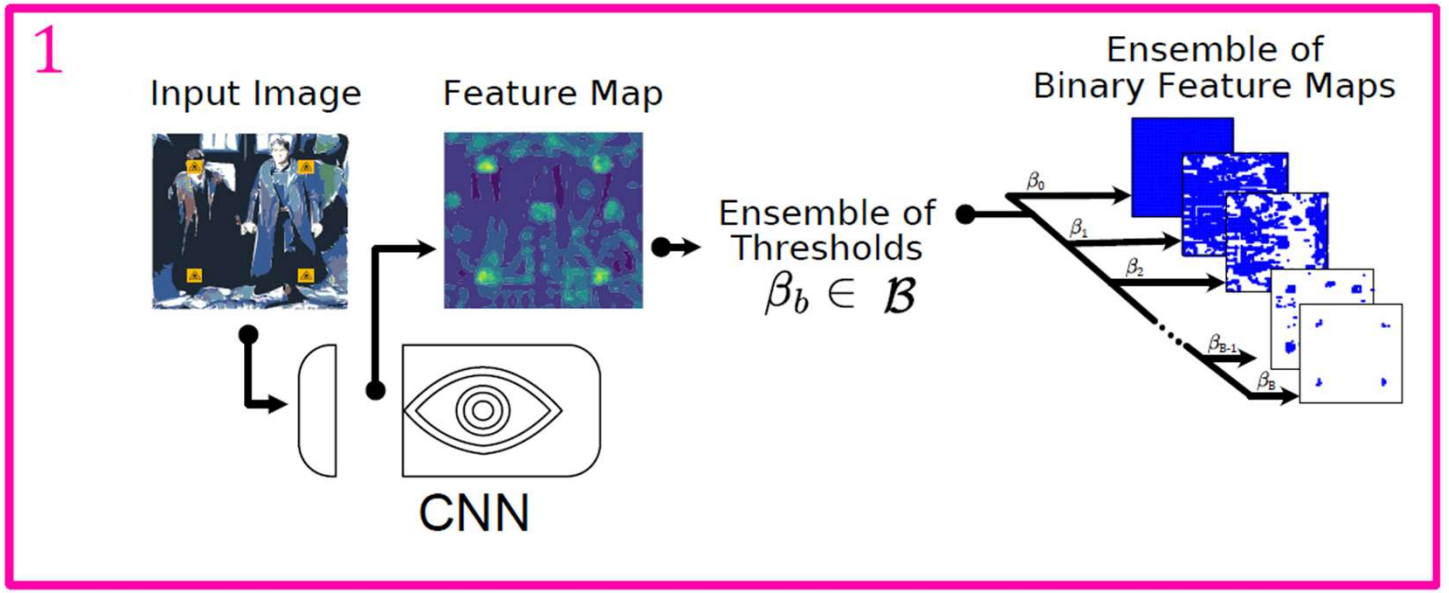
Spanning Saliency Thresholds using SpaNN

- Hypothesis: Attack changes behavior across saliency thresholds
- Approach: Transform image into saliency domain (think of Fourier transform)



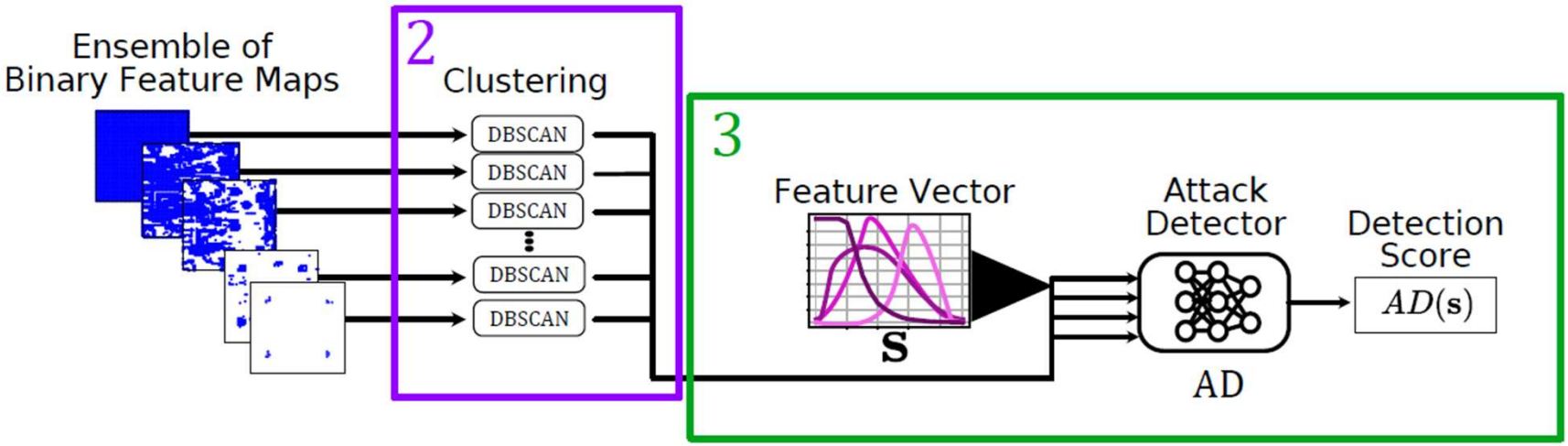
Step 1: Ensemble of Binary Feature Maps

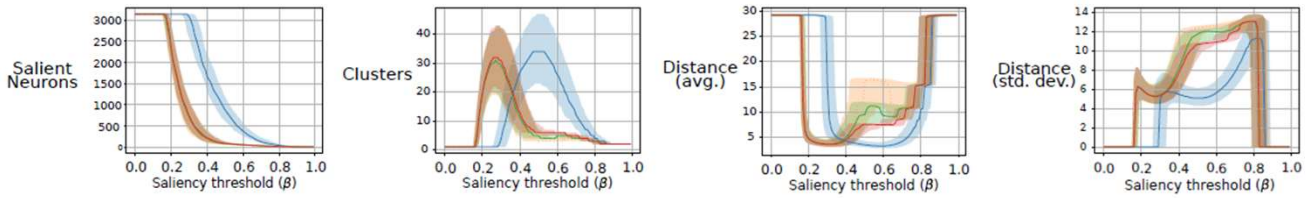
- Use CNN to create feature map
- Ensemble $\mathcal{B}$ of B thresholds
- Binary feature map for each $\beta_b \in \mathcal{B}$



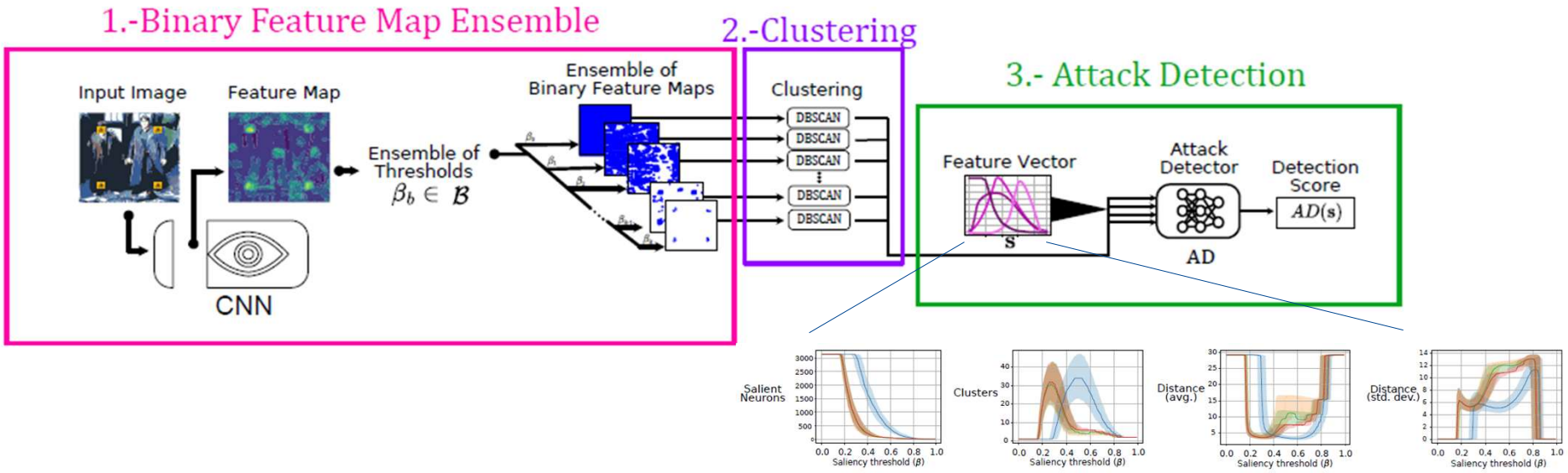
Step 2: Clustering
Step 3: Attack Detection

- Density based clustering of each binary feature map
- Create 4 features per feature map → feature vector $s \in \mathbb{R}^{4B}$
- Use s as input to attack detector AD to compute detection score





SpaNN: Spanning Saliency Threshold based Detection

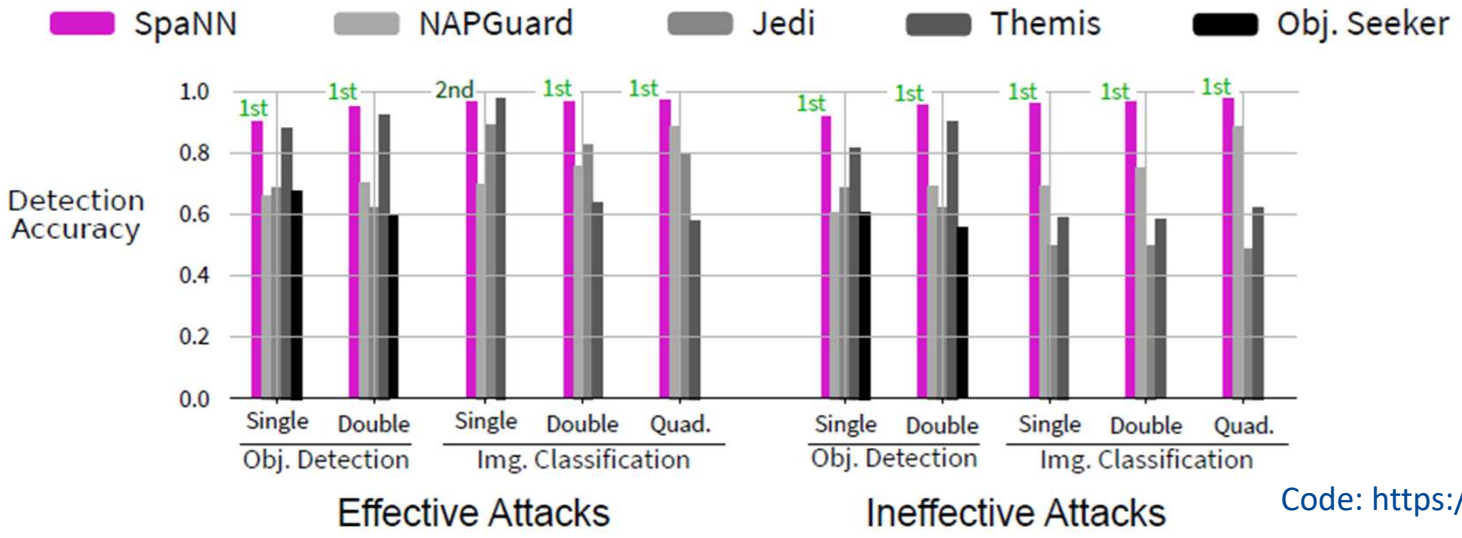


- Key advantages of SpaNN
  - No saliency threshold tuning ⇒ Adaptive attacks less powerful
  - Detection independent of size and number of patches ⇒ Detect multiple patches of arbitrary sizes and shapes
  - Complexity and accuracy depend on ensemble size ⇒ Accuracy vs. overhead tradeoff

Byrd Victorica et al, ``SpaNN: Detecting Multiple Adversarial Patches on CNNs by Spanning Saliency Thresholds'' in Proc. of IEEE Conf. on Secure and Trustworthy Machine Learning (SaTML), Apr. 2025

SpaNN Attack Detection Performance

- Evaluation for attacks on object detection and classification, 2+2 datasets
- SpaNN superior to all baselines
  - Higher detection accuracy
  - Performance insensitive to number of patches
  - Detects attacks that are not effective
  - Resilient to dynamic adversary (patches created to evade the detection scheme)



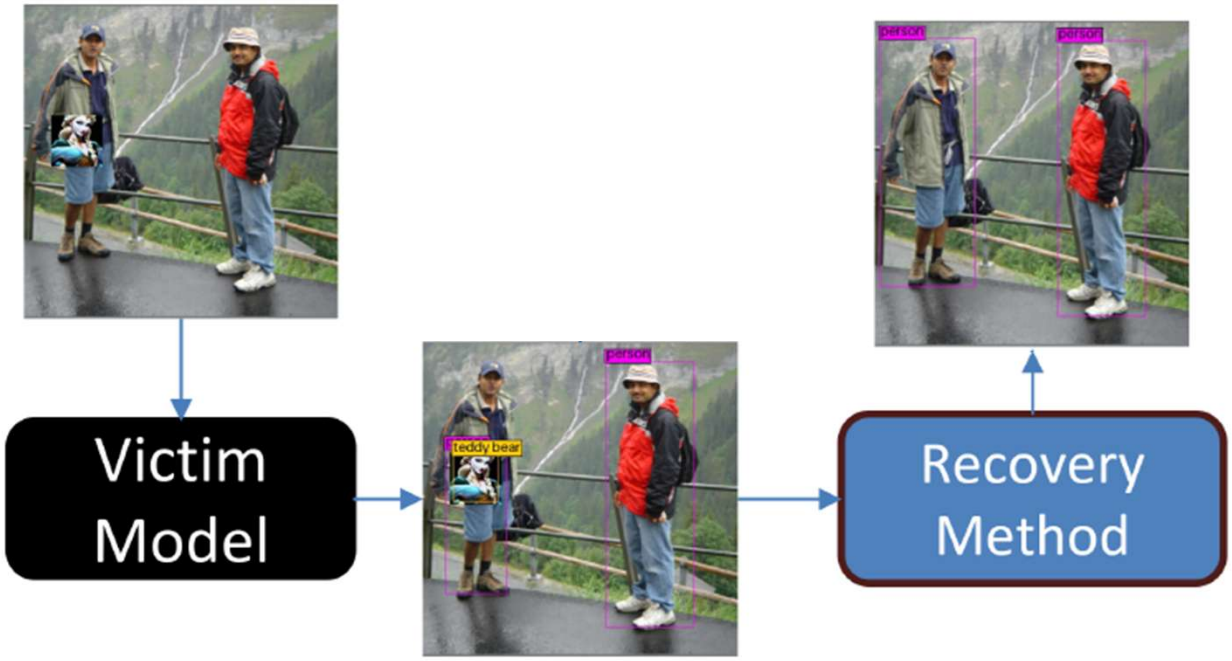
Code: https://github.com/gerkbyrd/SpaNN

Byrd Victorica et al, ``SpaNN: Detecting Multiple Adversarial Patches on CNNs by Spanning Saliency Thresholds'' in Proc. of IEEE Conf. on Secure and Trustworthy Machine Learning (SaTML), Apr. 2025

Going beyond detection: Recovery

- Limitation of existing methods
  - No explicit detection
  - Fixed saliency threshold
    - Cover salient areas exceeding threshold

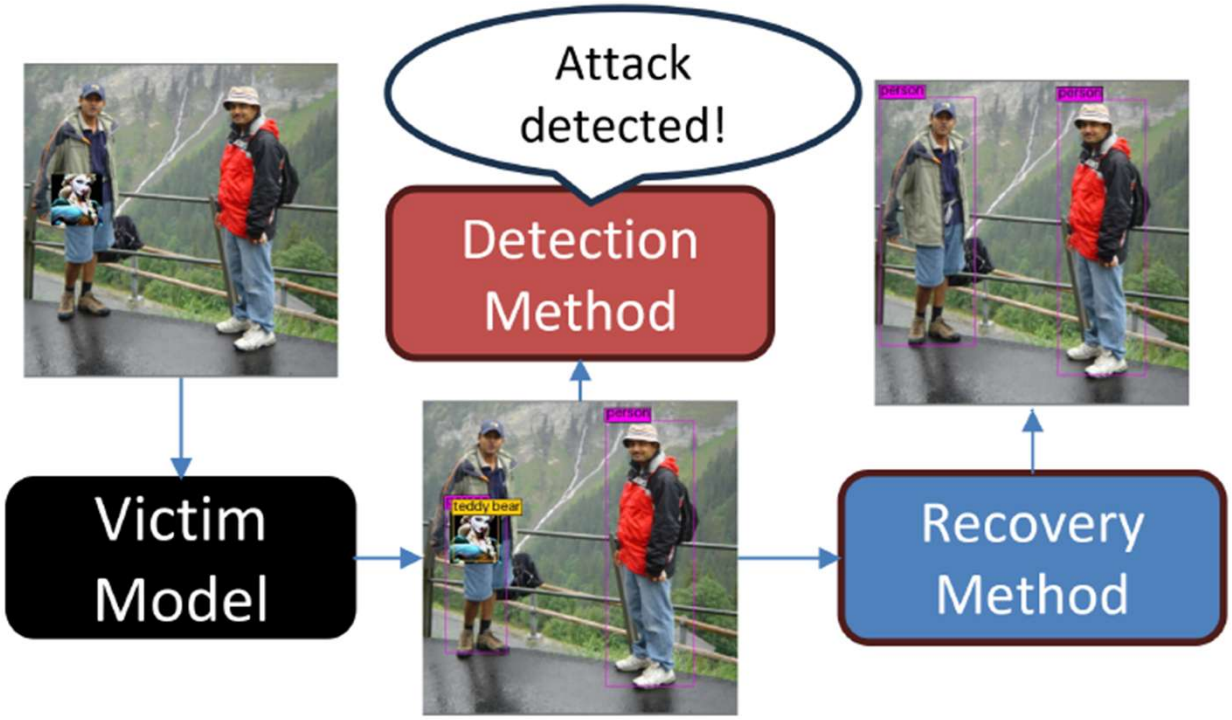
- Alter non-compromised inputs
- Cannot deal with multiple patches
- Adaptive attacks evade recovery



Going beyond detection: Recovery

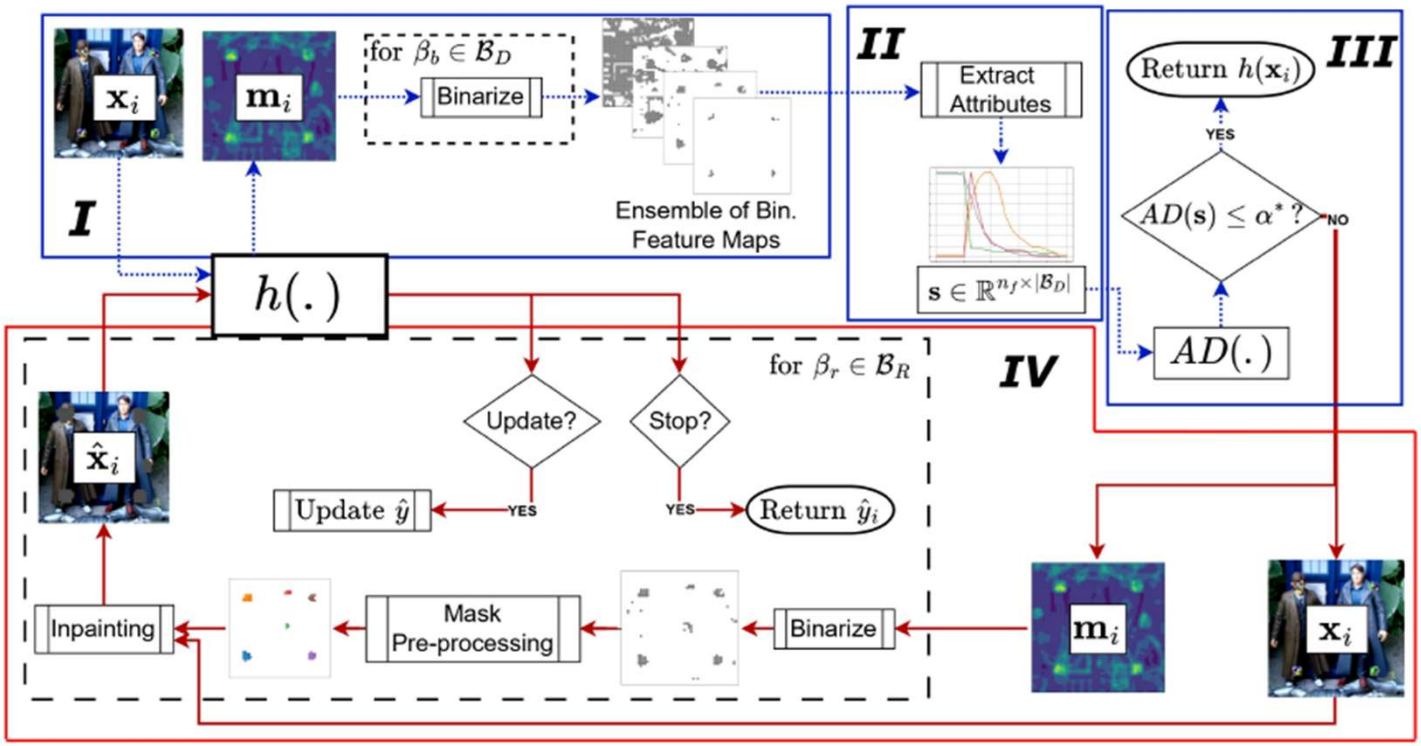
- Limitation of existing methods
  - No explicit detection
  - Fixed saliency threshold
    - Cover salient areas exceeding threshold

- Alter non-compromised inputs
- Cannot deal with multiple patches
- Adaptive attacks evade recovery



Ensemble Saliency for Recovery: Saliutl

- Novel two stage approach
  1. Detection using SpaNN
  2. Recovery by iterative inpainting over saliency thresholds

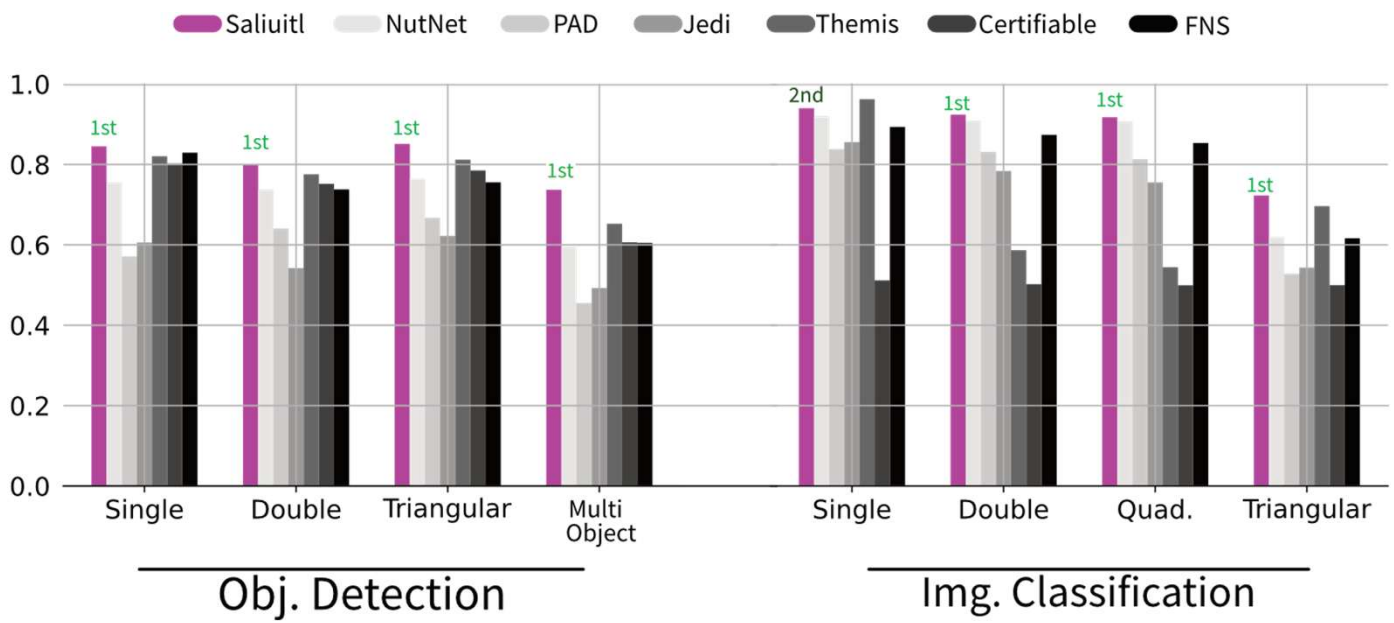


Byrd Victorica et al “Saliuitl: Ensemble Salience Guided Recovery of Adversarial Patches against CNNs” CVPR 2025

Saliutl Recovery Performance

- Improved average precision compared to baselines
- Does not ruin clean images – no spurious detections

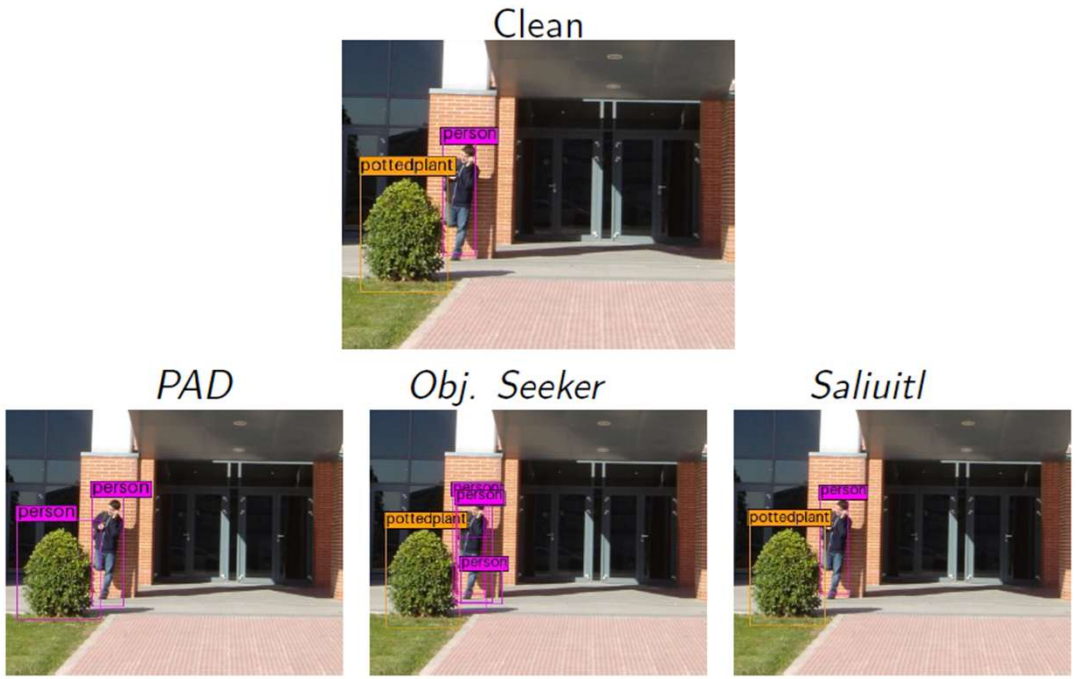
Clean/Adversarial Performance Tradeoff



Code: https://github.com/Saliuitl/Saliuitl/tree/main

Saliutl Recovery on Clean Images

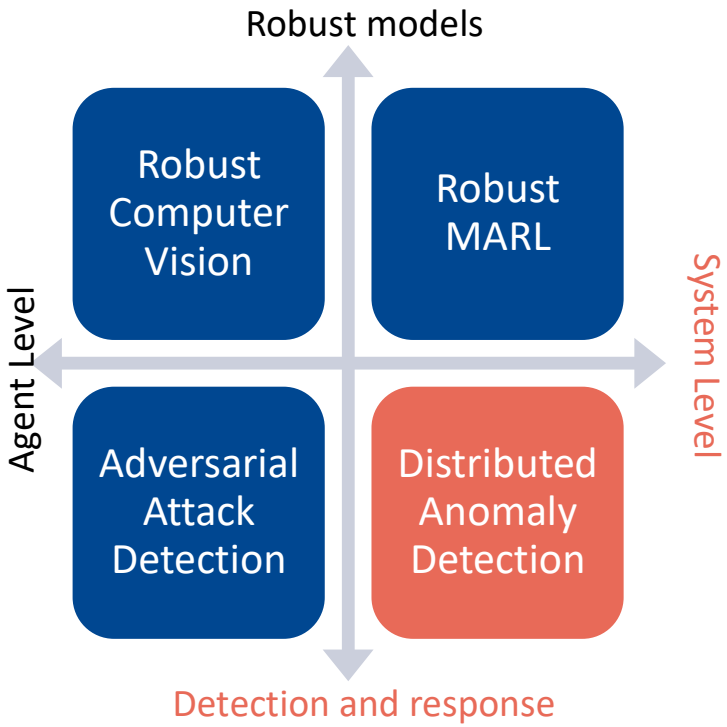
- Improved average precision compared to baselines
- Does not ruin clean images



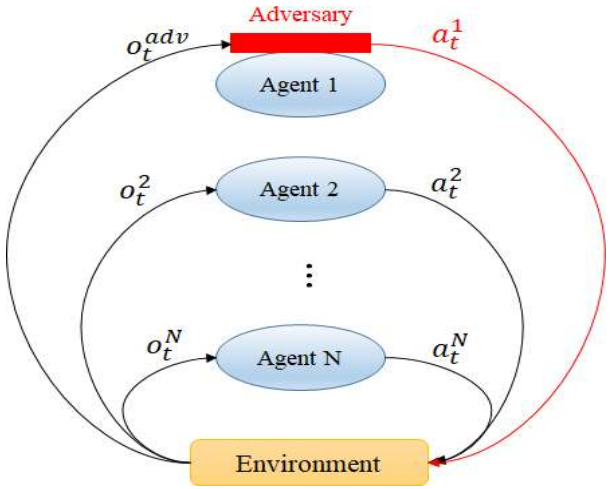
Code: https://github.com/Saliuitl/Saliuitl/tree/main

Byrd Victorica et al “Saliuitl: Ensemble Salience Guided Recovery of Adversarial Patches against CNNs” CVPR 2025

Agenda


Robust models
Robust Computer Vision
Robust MARL
Adversarial Attack Detection
Distributed Anomaly Detection
Agent Level
System Level
Detection and response


Adversarial Patch
Person
Person
Alarm
Alarm


Adversary
o_t^{adv}
a_t^1
Agent 1
o_t^2
Agent 2
a_t^2
o_t^N
Agent N
a_t^N
Environment

Attack Detection Problem

Quickest Detection

- Adversary starts to attack an (unknown) agent v at an unknown time step t_0
- Agents can observe the actions of other agents
- Objective: identify the victim agent(s) as soon as possible after the attack starts
- Lower bound δ_F on mean time between false detections

$$\min \sup_{t_0<\infty} E^{(t_0)}[u_v - t_0 | u_v \geq t_0]$$
$$\text{s.t. } E^{(\infty)}[u_v] \geq \delta_F$$

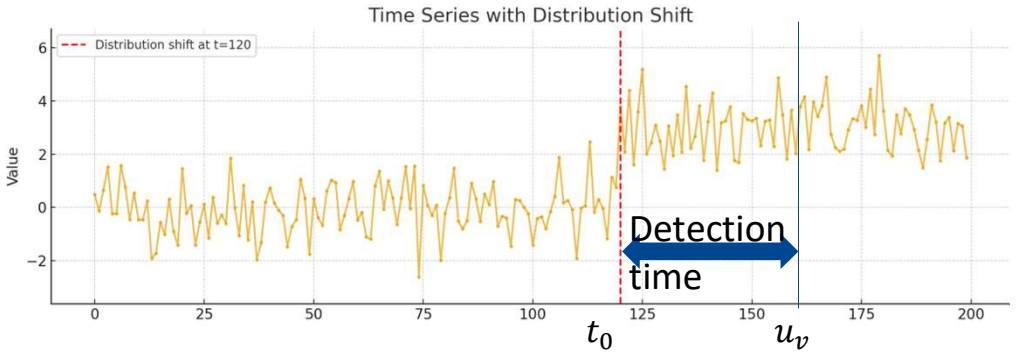
Attack Detection Problem

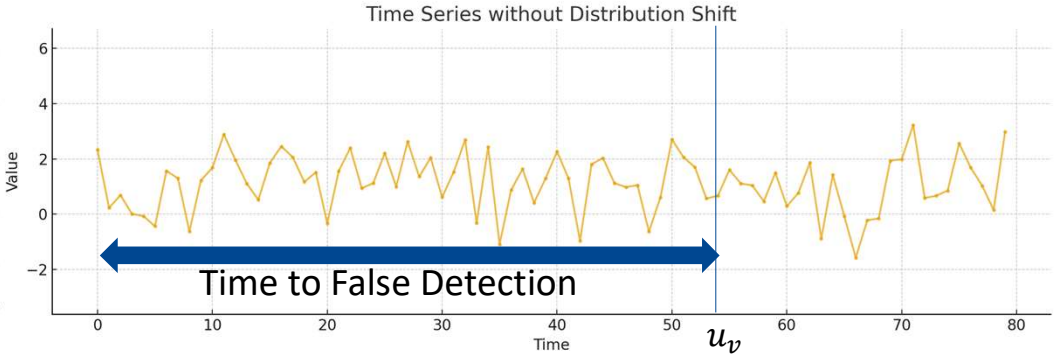
Quickest Detection

- Adversary starts to attack an (unknown) agent v at an unknown time step t_0
- Agents can observe the actions of other agents
- Objective: identify the victim agent(s) as soon as possible after the attack starts
- Lower bound δ_F on mean time between false detections

$$\min \sup_{t_0<\infty} E^{(t_0)}[u_v - t_0 | u_v \geq t_0]$$

$$\text{s.t. } E^{(\infty)}[u_v] \geq \delta_F$$





Attack Detection Problem – Change detection

Quickest Detection

- Adversary starts to attack an (unknown) agent v at an unknown time step t_0
- Agents can observe the actions of other agents
- Objective: identify the victim agent(s) as soon as possible after the attack starts
- Lower bound δ_F on mean time between false detections

$$\min \sup_{t_0<\infty} E^{(t_0)}[u_v - t_0 | u_v \geq t_0]$$

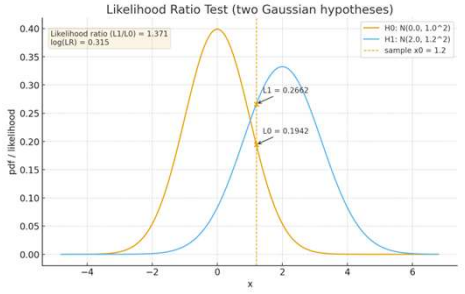
$$\text{s.t. } E^{(\infty)}[u_v] \geq \delta_F$$

- Change detection interpretation
  - $A_t \sim f_o\ (t < t_0), A_t \sim f_1\ (t \geq t_0)$
  - $A_1^t = (A_1, \dots, A_t)$
- CUSUM
  - Log-likelihood ratio: $s_t = ln \frac{f_1(A_t)}{f_0(A_t)}$
  - Decision function $g_0 = 0, \quad g_t = (g_{t-1} + s_t)^+$
  - Detection rule: $t_a = \min\{t: g_t \geq \beta\}$

Hypothesis about underlying distribution

Sequence of observed actions of other agent

$$E[u_v - t_0] \sim \frac{1}{KL(f_0 || f_1)}$$



Detection Problem – Change detection

Quickest Detection

- Adversary starts to attack an (unknown) agent v at an unknown time step t_0
- Agents can observe the actions of other agents
- Objective: identify the victim agent(s) as soon as possible after the attack starts
- Lower bound δ_F on mean time between false detections

$$\min \sup_{t_0<\infty} E^{(t_0)}[u_v - t_0 | u_v \geq t_0]$$
$$\text{s.t. } E^{(\infty)}[u_v] \geq \delta_F$$

- Change detection interpretation
  - $A_t \sim f_o\ (t < t_0), A_t \sim f_1\ (t \geq t_0)$
  - $A_1^t = (A_1, \dots, A_t)$
- CUSUM
  - Log-likelihood ratio: $s_t = ln \frac{f_1(A_t)}{f_0(A_t)}$
  - Decision function $g_0 = 0, \quad g_t = (g_{t-1} + s_t)^+$
  - Detection rule: $t_a = \min\{t: g_t \geq \beta\}$

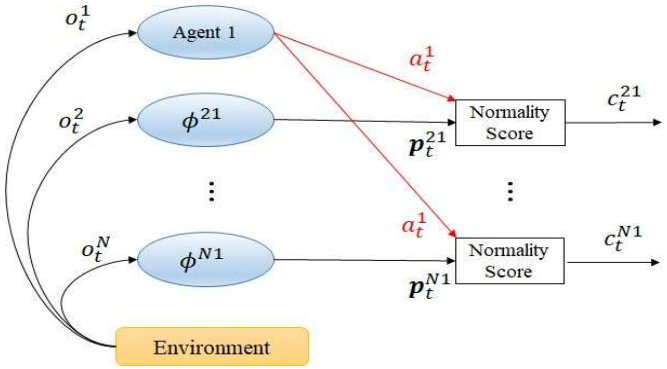
Hypothesis about underlying distribution

Sequence of observed actions of other agent

Challenge:
- f_1 depends on the attack
- f_0 and f_1 depend on the state

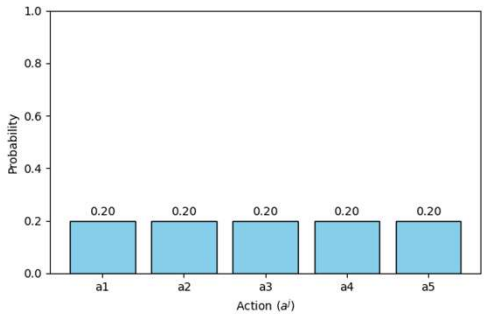
Distributed detection for discrete action sets

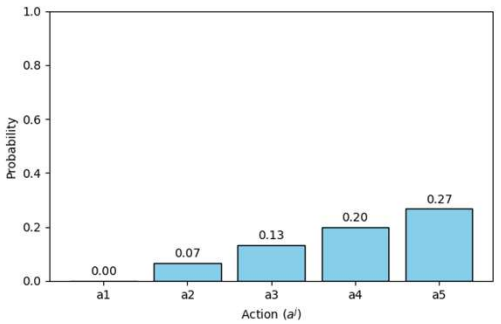
- Discrete action set: $A_i \in \mathbb{N}^d$
- Idea: Characterize normal behavior of agents as seen by other agents
  - conditioned on local observation
- Detection scheme:
  - Predict the categorical distribution of actions based on local observations
  - Compute (ab)normality score
    - $z_t^{ij} \triangleq \log\left(\frac{p_t^{ij}(a^j)}{\max_{a^j} p_t^{ij}(a^j)}\right)$
    - $c_t^{ij} \triangleq \begin{cases} \frac{1}{t}\sum_{l=1}^{t} z_l^{ij}, & 1 \le t < w \\ \frac{1}{w}\sum_{l=t-w+1}^{t} z_l^{ij}, & t \ge w \end{cases}$
  - Detection rule: $c_t^{ij} < \beta^{ij}$



ϕ^{ij}: predictor of agent *i* for action distribution of agent *j*

p_t^{ij} :predicted distribution of actions of agent *j* by agent *i*

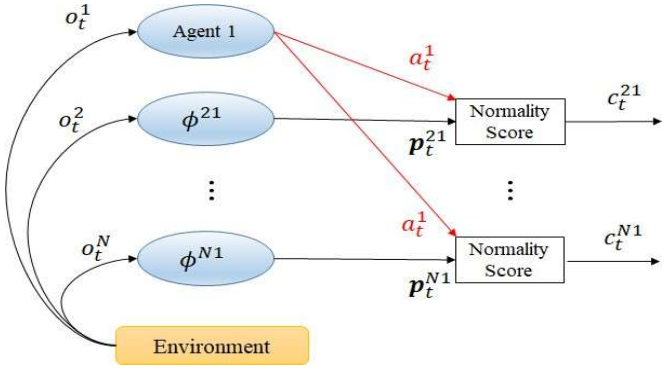




Kazari et al., ``Decentralized Anomaly Detection in Cooperative Multi-Agent Reinforcement Learning'', in Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI), Aug. 2023

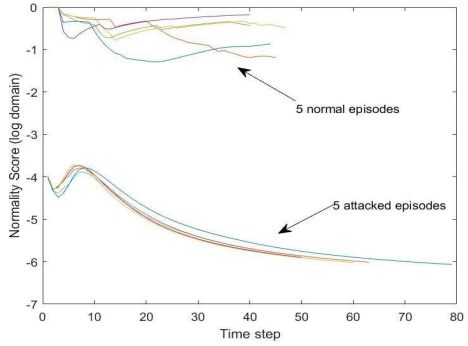
Distributed detection for discrete action sets

- Discrete action set: $A_i \in \mathbb{N}^d$
- Idea: Characterize normal behavior of agents as seen by other agents
  - conditioned on local observation
- Detection scheme:
  - Predict the categorical distribution of actions based on local observations
  - Compute (ab)normality score
    - $z_t^{ij} \triangleq \log\left(\frac{p_t^{ij}(a^j)}{\max_{a^j} p_t^{ij}(a^j)}\right)$
    - $c_t^{ij} \triangleq \begin{cases} \frac{1}{t}\sum_{l=1}^{t} z_l^{ij}, & 1 \le t < w \\ \frac{1}{w}\sum_{l=t-w+1}^{t} z_l^{ij}, & t \ge w \end{cases}$
  - Detection rule: $c_t^{ij} < \beta^{ij}$



ϕ^{ij}: predictor of agent *i* for action distribution of agent *j*

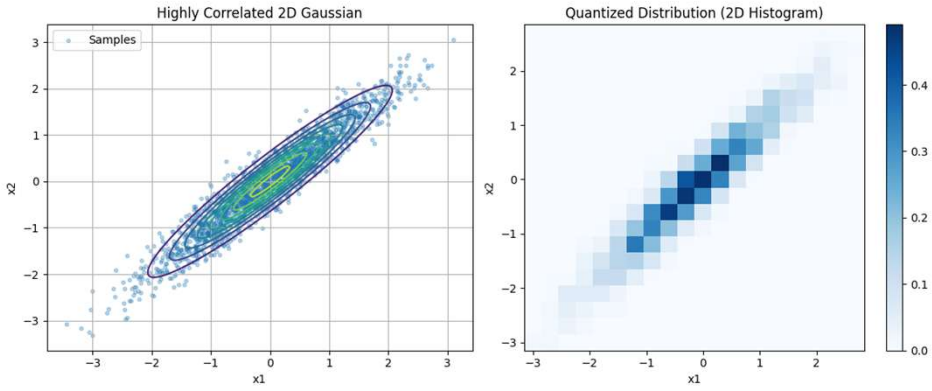
p_t^{ij} :predicted distribution of actions of agent *j* by agent *i*

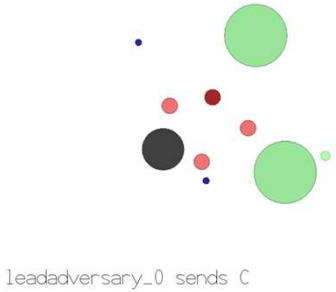


Kazari et al., ``Decentralized Anomaly Detection in Cooperative Multi-Agent Reinforcement Learning'', in Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI), Aug. 2023

How to deal with continuous action sets?

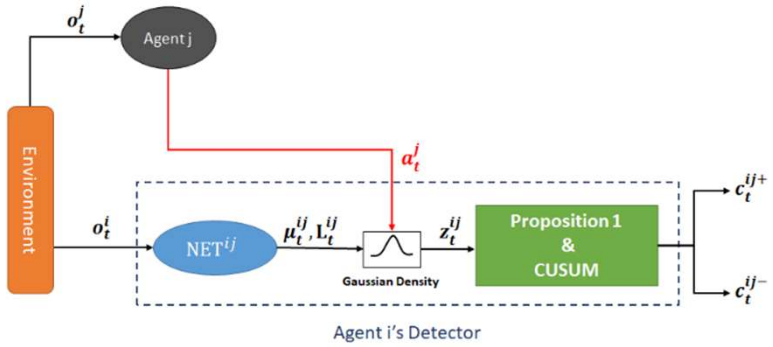
- Continuous action set $A_i \in \mathbb{R}^d$
- Discretization of continuous action set scales poorly
  - q^d actions with q bins per dimension
  - qd actions if assuming independence





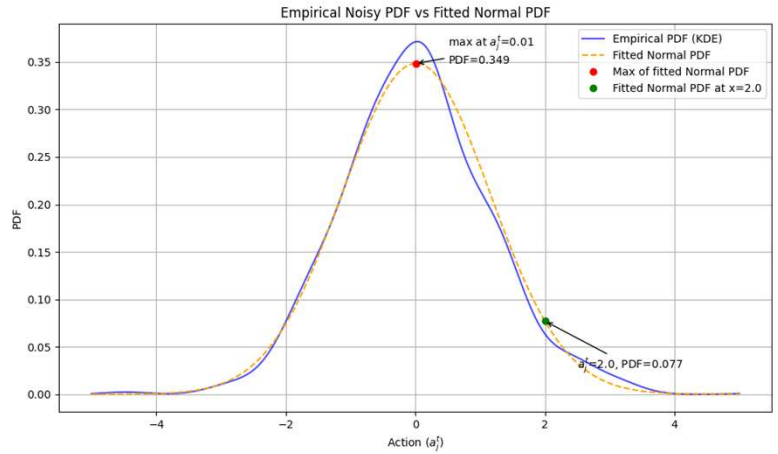
Detection for continuous action sets
Parametrized Gaussian CUSUM

- Approximate action distribution
  - $f_t^{ij}(a_j) \sim \mathcal{N}(\mu_t^{ij}, \sigma_t^{ij})$
  - Multivariate Gaussian distribution predicted based on past observations τ_t^{ij}
- Normality score
  - $z_t^{ij} = \log\left(\frac{f^{ij}(a_j^t|\tau_t^{ij})}{\max_a f^{ij}(a|\tau_t^{ij})}\right)$



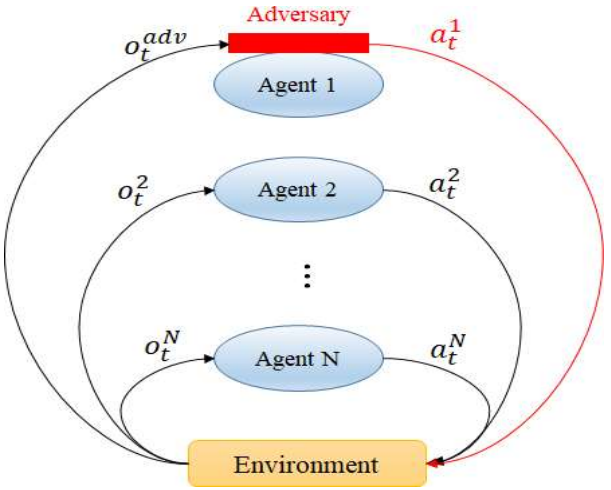
Agent i's Detector

- **Result:** Closed form expression of mean and stdev of normality score without anomaly
  - $m_z^j = E[z_t^{ij}] = -\frac{d^j}{2}, \qquad \sigma_z^j = \sqrt{E\left[\left(z_t^{ij} - m_z^j\right)^2\right]} = d/2$
- Maintain CUSUM statistics
  - $c_t^{ij} = \max\left\{0, c_{t-1}^{ij+} + \frac{z_t^{ij} - m_z^j}{\sigma_z^j} - w\right\}$
  - $c_t^{ij} = \max\left\{0, c_{t-1}^{ij-} - \frac{z_t^{ij} - m_z^j}{\sigma_z^j} - w\right\}$
- Detection thresholds β^{ij} , β^{ij-}



Kazari et al., `` Distributed Detection of Adversarial Attacks in Multi-Agent Reinforcement Learning with Continuous Action Space'', in Proc. of European Conference on Artificial Intelligence (ECAI), 2025

Fooling the Detector: Dynamic Adversary



- Attacker has access to
  - Predictor p_t^{iv} or f_t^{iv}
  - Thresholds β^{iv}
- Attack π^{adv} on agent v is **expectedly undetectable** if
  - $\mathrm{E}\left[c_t^{iv}\left(\pi^{adv}\right)\right] \geq \beta^{iv} \;\; \forall t > 0, \forall i \neq v$
- Attack policy can be obtained by solving a non-Markovian problem

$$\max E\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_r^{adv}\right]$$

$$s.t. \qquad E\left[c_t^{iv}\right] \geq \beta^{iv} \qquad \forall t, \forall i \neq v$$

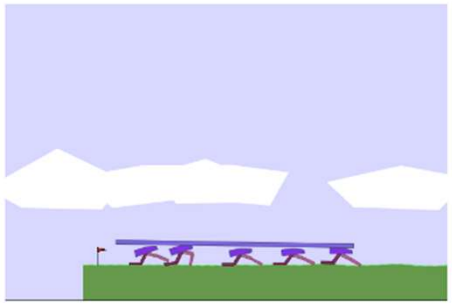
- Solution
  - Extend state space and relax constraint to obtain Markovian problem that upper bounds the problem
  - Exploit zero duality gap and minimize dual using gradient descent over Lagrangian

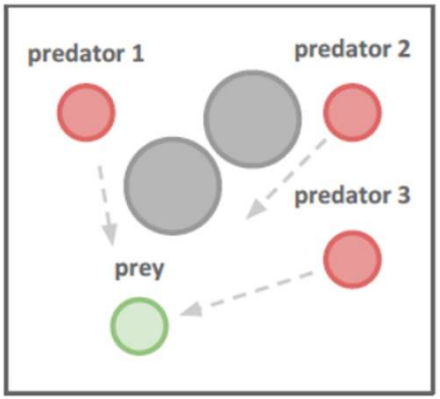
Kazari et al., ``Decentralized Anomaly Detection in Cooperative Multi-Agent Reinforcement Learning'', in Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI), Aug. 2023

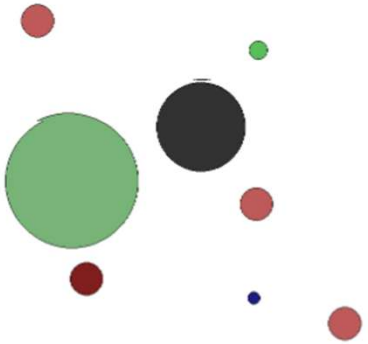
Detection Performance

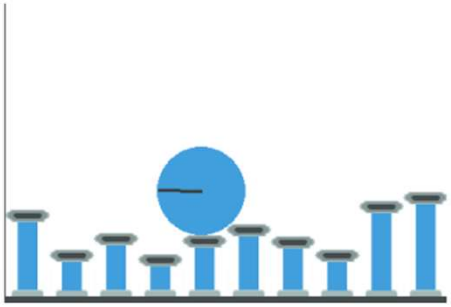
Average episodic rewards

	Multiwalker	Tag	World Comm	Pistonball
No Attack	-12.7	101.8	37.6	228.6
ACT	-107.6	64.9	26.7	83.1
RAND	-75.6	68.1	30.4	202.1
Grad	-42.7	90.4	34.7	215.1
DYN1	-96.9	65.1	27.8	95.5
DYN2	-89.6	69.2	30.1	139.3









Detection Performance

Average episodic rewards

	Multiwalker	Tag	World Comm	Pistonball
No Attack	-12.7	101.8	37.6	228.6
ACT	-107.6	64.9	26.7	83.1
RAND	-75.6	68.1	30.4	202.1
Grad	-42.7	90.4	34.7	215.1
DYN1	-96.9	65.1	27.8	95.5
DYN2	-89.6	69.2	30.1	139.3

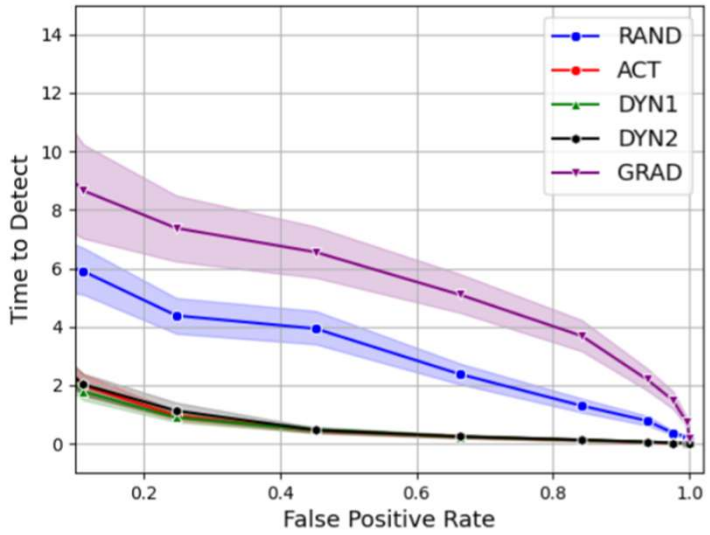
	Multiwalker		**Tag**		**World Comm**		**Pistonball**	
Attack Types	*PGC*	*Discrete*	*PGC*	*Discrete*	*PGC*	*Discrete*	*PGC*	*Discrete*
ACT	**0.996**	0.972	**0.993**	0.948	**0.995**	0.821	**0.999**	0.758
RAND	**0.995**	0.855	0.843	**0.893**	0.677	**0.713**	**0.997**	0.970
GRAD	**0.674**	0.566	0.653	**0.858**	0.884	**0.913**	**0.581**	0.554
DYN1	**0.929**	0.818	**0.988**	0.964	**0.992**	0.754	**0.907**	0.711
DYN2	**0.954**	0.788	**0.968**	0.944	**0.912**	0.707	**0.876**	0.658

ROC AUC

Time to Detection

Average episodic rewards

	Multiwalker	Tag	World Comm	Pistonball
No Attack	-12.7	101.8	37.6	228.6
ACT	-107.6	64.9	26.7	83.1
RAND	-75.6	68.1	30.4	202.1
Grad	-42.7	90.4	34.7	215.1
DYN1	-96.9	65.1	27.8	95.5
DYN2	-89.6	69.2	30.1	139.3



Tag

Recovery via Detection and Response

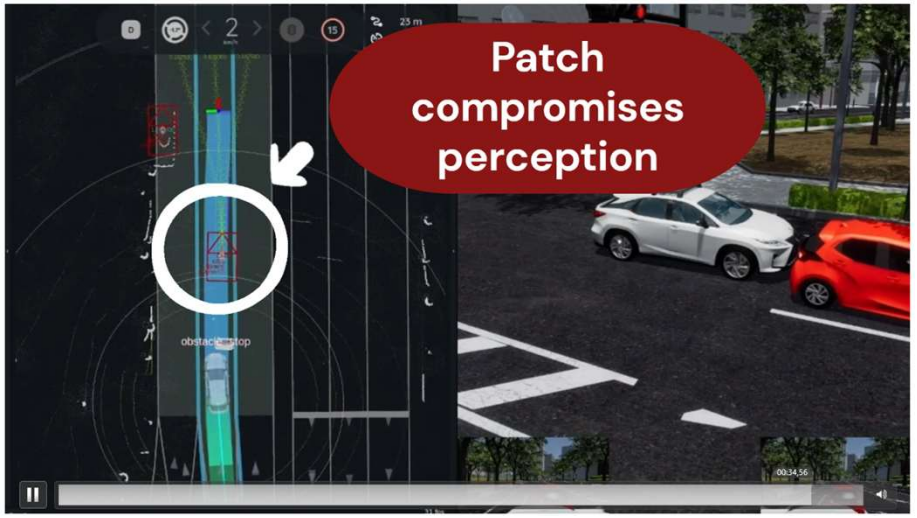
Method	Win rate
EIR-MAPPO	0.35
GenM	0.22
RAP	0.2
MAPPO	0.0
D&R	**0.36**
Oracle	0.75

Conclusion

- ML vulnerabilities are a threat to the safety of autonomous systems

- Defense in depth for ML-enabled CPS
  - Agent level detection and robustification
    - SpaNN and Saliutl for patch attacks
  - System level detection and robustification
    - Distributed detection and response
  - Runtime defense at design time

- Test, verify, and secure ML in every layer
  - A vulnerable AI is worse than a useless one



References

- Kazari et al., ``Decentralized Anomaly Detection in Cooperative Multi-Agent Reinforcement Learning'', in Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI), Aug. 2023
- Shereen et al, ``Adversarial Robustness of Multi-agent Reinforcement Learning Secondary Control of Islanded Inverter-based AC Microgrids,'' in Proc. of IEEE SmartGridComm, Oct. 2023
- Ezzeldin Shereen, Kiarash Kazari, György Dán, ``A Reinforcement Learning Approach to Undetectable Attacks against Automatic Generation Control,'' IEEE Trans. on Smart Grids, vol. 15., no. 1., Jan. 2024
- *Kiarash Kazari, Aris Kanellopulos, György Dán, ``Quickest Detection of Adversarial Attacks against Correlated Equilibria,'' in Proc. of AAAI Conference on Artificial Intelligence (AAAI), Feb 2025*
- Byrd Victorica et al, ``SpaNN: Detecting Multiple Adversarial Patches on CNNs by Spanning Saliency Thresholds'' in Proc. of IEEE Conf. on Secure and Trustworthy Machine Learning (SaTML), Apr. 2025
- Byrd Victorica et al "Saliuitl: Ensemble Salience Guided Recovery of Adversarial Patches against CNNs". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2025
- Kazari et al., `` Distributed Detection of Adversarial Attacks in Multi-Agent Reinforcement Learning with Continuous Action Space'', in Proc. of European Conference on Artificial Intelligence (ECAI), Oct. 2025

Thank you




From Pixels to Policies: Securing Multi-Agent Systems Against Adversarial Attacks

György Dán

2 October 2025