Improving Resilience and Trust in AI Driven Cyber Physical Systems

Dr. Apurva Narayan
University of Western Ontario,
1151 Richmond St | London, ON | N6A 3K7
Canada
apurva.narayan@uwo.ca

Intelligent Data Science Lab

Empowering the next generation of Cyber Physical Systems through the power of Data and AI

SETTING THE STAGE

Systems are Safety-Critical

Taylor Swift

Taylor Swift used facial recognition software to detect stalkers at LA concert

The Rose Bowl venue didn't inform concert-goers that their image might be collected at a special kiosk showing Taylor Swift rehearsal clips

Laura Snapes

Thu 13 Dec 2018 12.58 GMT

Most viewed


I reversed my type 2 diabetes. Here's how I did it
Neil Barsky


Live Russia-Ukraine war live: Poland will demand EU restores permits for Ukrainian truckers as first vehicles cross border


Former Hong Kong activist Agnes Chow flees territory for Canada


Cop28 president says there is 'no science' behind demands for phase-out of fossil fuels


My plane crashed in the Andes. Only the unthinkable kept me and the other starving survivors alive

When Deep Learning Goes Wrong!

- What humans may perceive as imperceptible changes can produce misclassification errors that can be exploited.
- Deployed systems often have incentives to manipulate the output. This incentive can be financial, privacy-preserving, or malicious.

Changing facial recognition results using crafted glasses.

From: A General Framework for Adversarial Examples with Objectives, Sharif (2019)

15 Aug 2019

Engineering and IT

Media contact
Stephanie Juleff
sjuleff@unimelb.edu.au
+61 466 023 039

Find an Expert >

Related news

Myki privacy de-railed: Travellers' movements and identities at risk by public release of "anonymised data"

Detailed personal information can be garnered from the data on travellers' Myki cards.

Carnegie Mellon University

News

Stories | Experts

News › Stories › Archives › 2023 › July › Researchers Discover New Vulnerability in Large Language Models

July 28, 2023

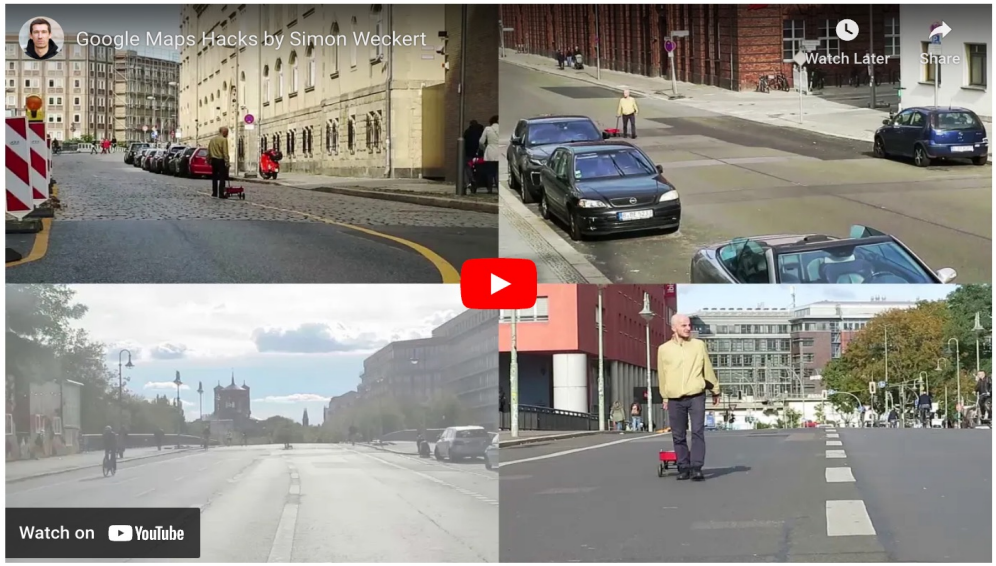
Researchers Discover New Vulnerability in Large Language Models

By: Ryan Noone

Business | Oddities | Technology | Traffic Lab

A man walked down a street with 99 phones in a wagon. Google Maps thought it was a traffic jam.

Feb. 10, 2020 at 6:32 am | *Updated Feb. 10, 2020 at 5:50 pm*

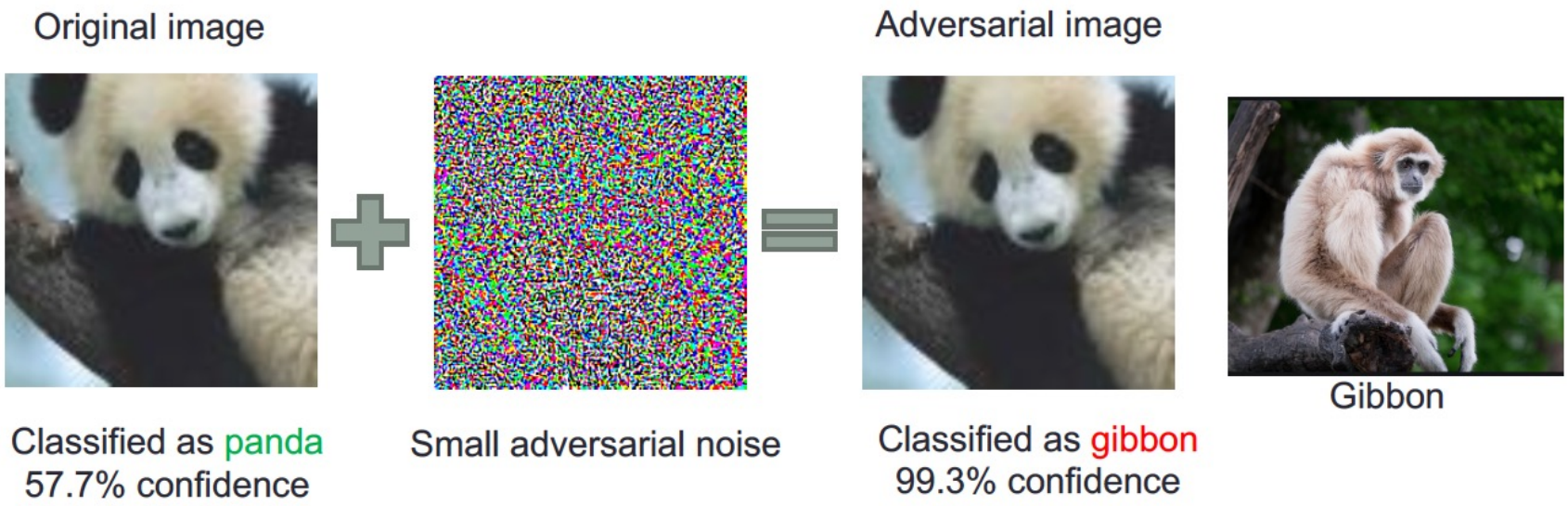


AI Failure can be catastrophic!

Robust, Secure and Trustworthy functioning of machine learning is the foundation of autopilot systems and other safety-critical problems.

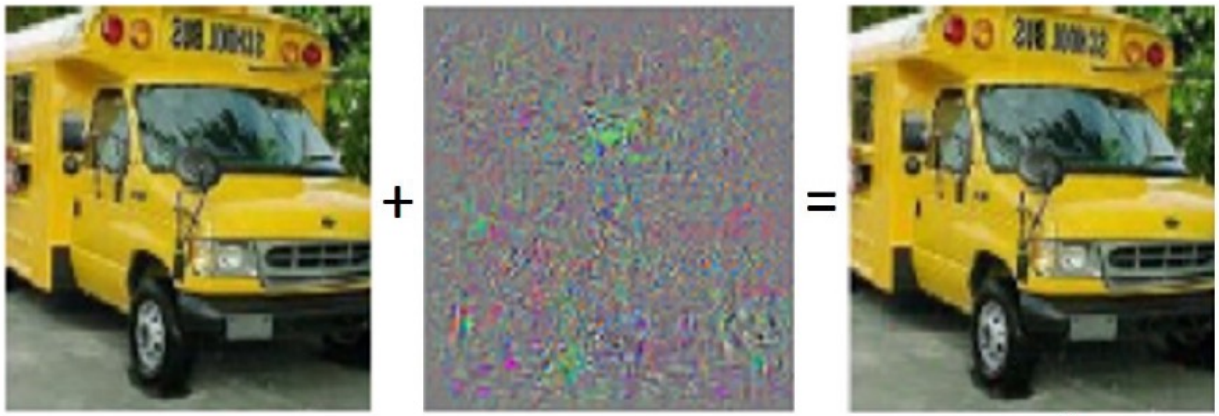
ADVERSARIAL ATTACKS

Adversarial Attacks on AI Models



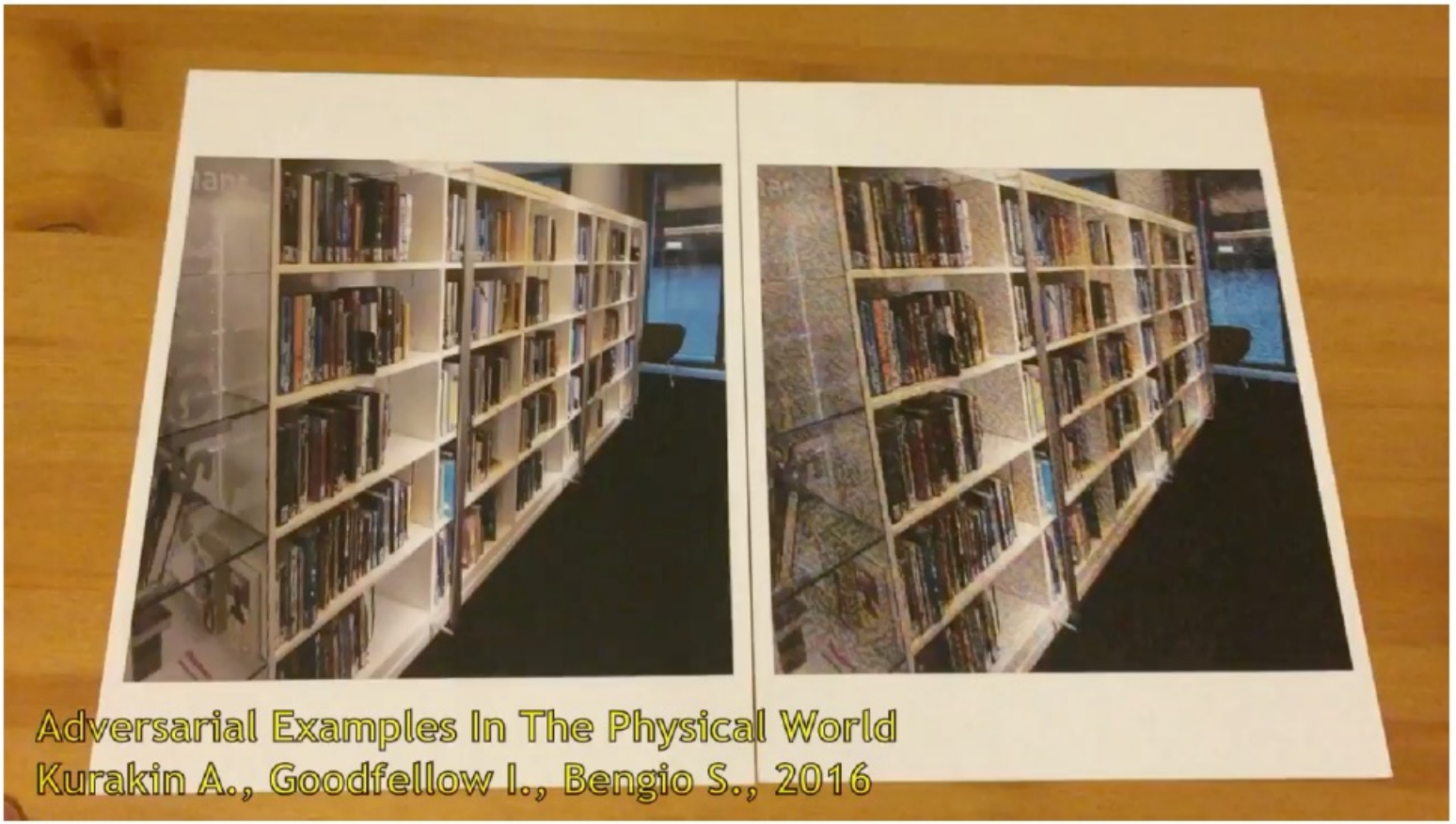
Picture from: Goodfellow et al. (2014) – Explaining and Harnessing Adversarial Examples

Adversarial Attacks on AI Models



Picture from: Szagedy et al. (2014) – Intriguing Properties of Neural Networks

Adversarial Attacks on AI Models



Practical Adversarial Attacks – Patch Attacks

Lab (Stationary) Test

Physical road signs with adversarial perturbation under different conditions

Cropping, Resizing

Stop Sign → Speed Limit Sign

Field (Drive-By) Test

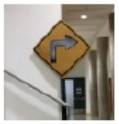
Video sequences taken under different driving speeds

Sample Per K Frames, Cropping, Resizing

Stop Sign → Speed Limit Sign

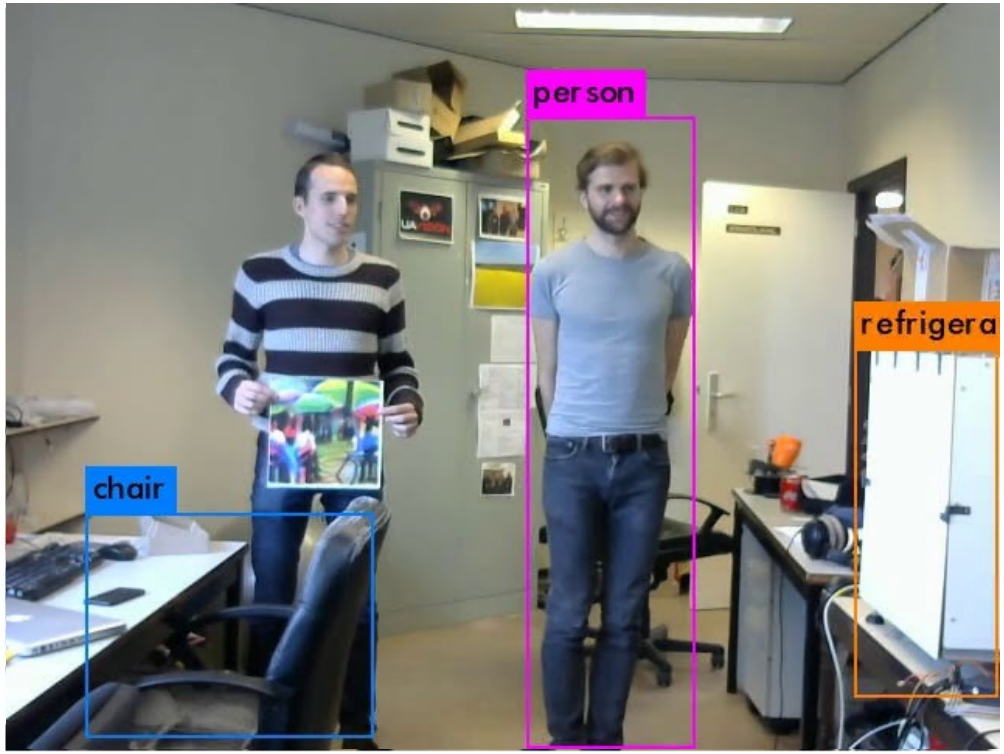
Picture from: Eykholt (2017) - Robust Physical-World Attacks on Deep Learning Visual Classification

Practical Adversarial Attacks – Patch Attacks

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Picture from: Eykholt (2017) - Robust Physical-World Attacks on Deep Learning Visual Classification

Practical Adversarial Attacks – Patch Attacks

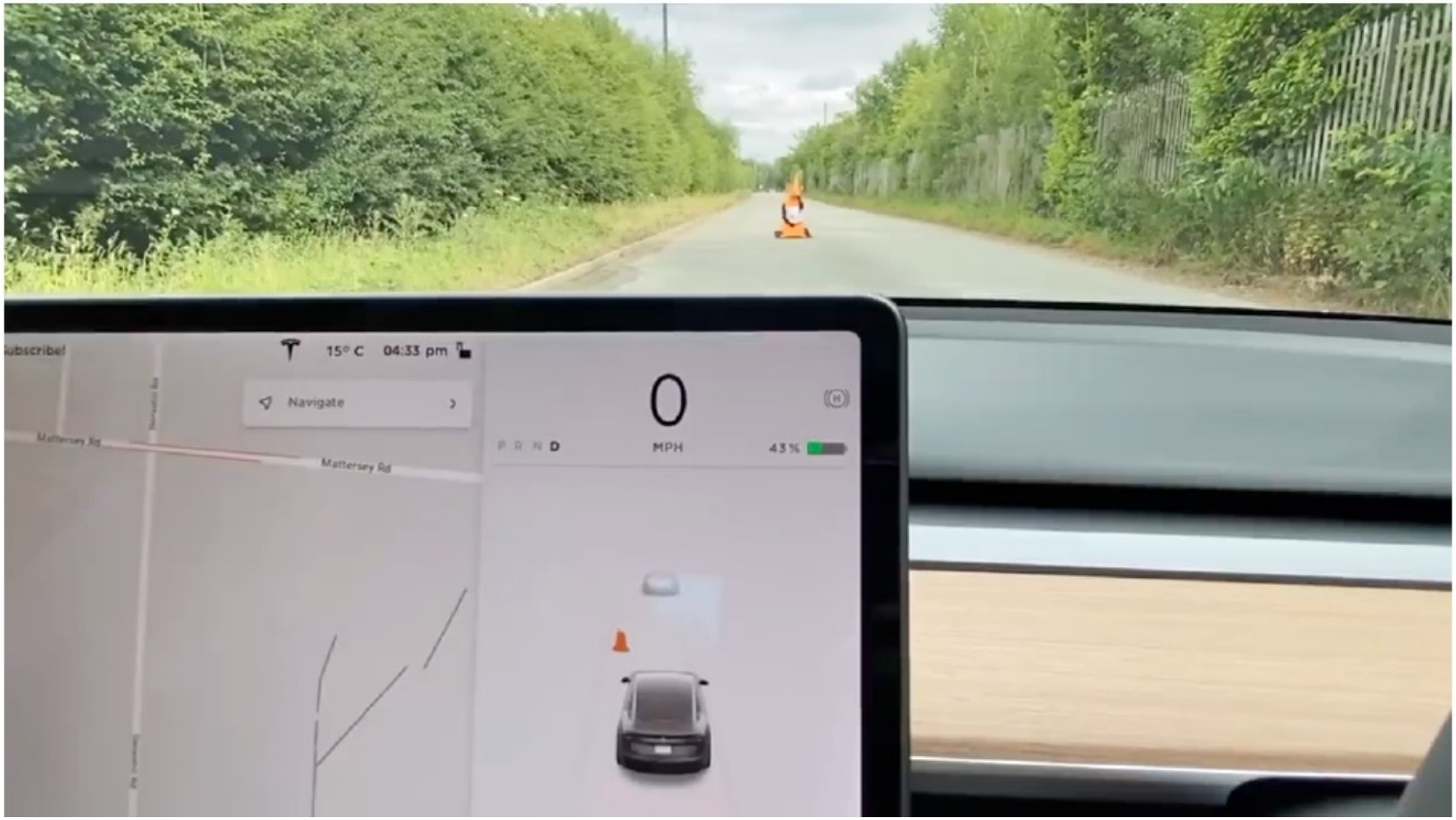


Thys et al. (2019) - Fooling automated surveillance cameras: adversarial patches to attack person detection

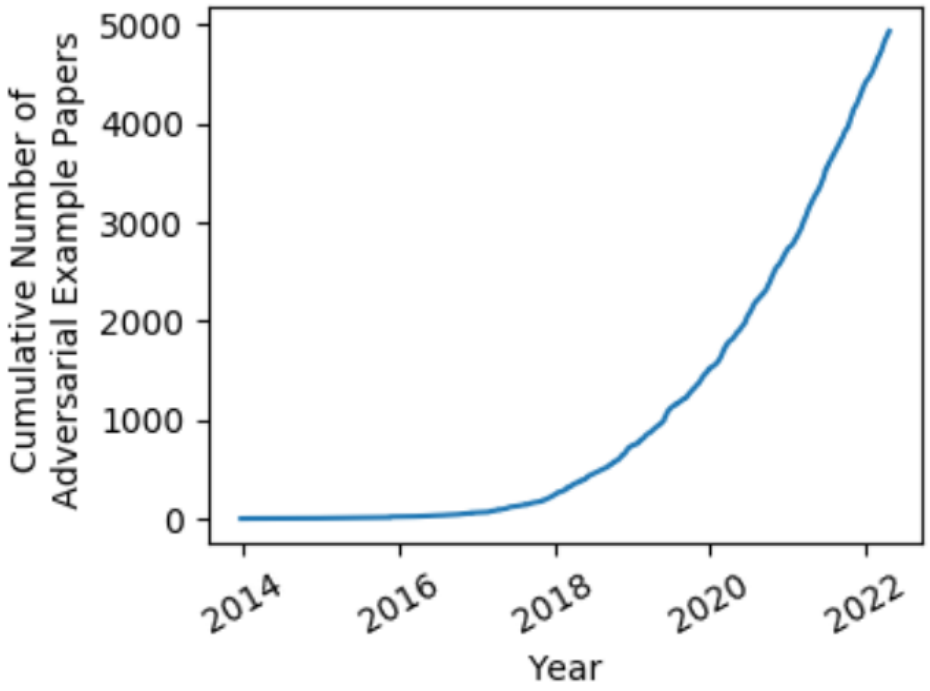
Practical Adversarial Attacks – Patch Attacks

A General Framework for Adversarial Examples with Objectives, ACM Transactions on Privacy and Security, 2019

Practical Adversarial Attacks



Increasing race in Adversarial Machine Learning

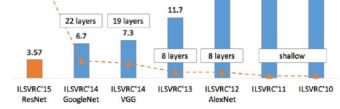


https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

Significant Events

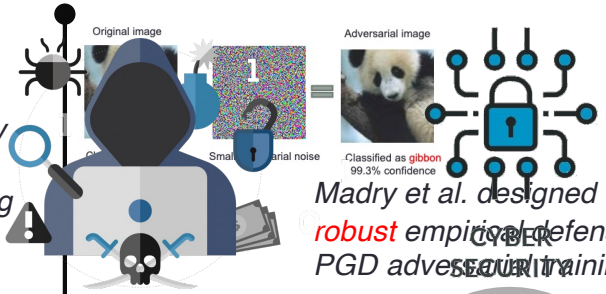
Before 2014, ML algorithms mainly focus on increasing accuracy, and few techniques focused on keeping the ML models secure and robust

Pre 2014



ILSVRC (ImageNet)

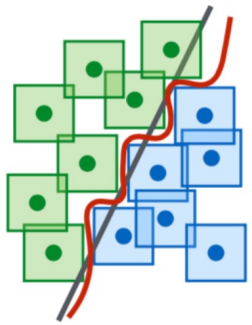
Post 2014



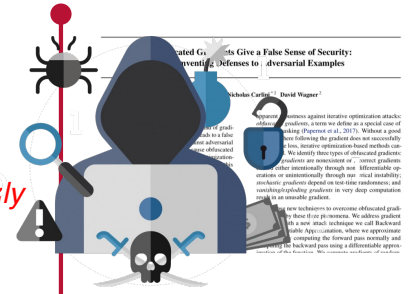
Szagedy et al. and Goodfellow et al. showed the first adversarial examples of panda, which started the area of adversarial ML

2017

Madry et al. designed the first truly robust empirical defense --- PGD adversarial training



2018

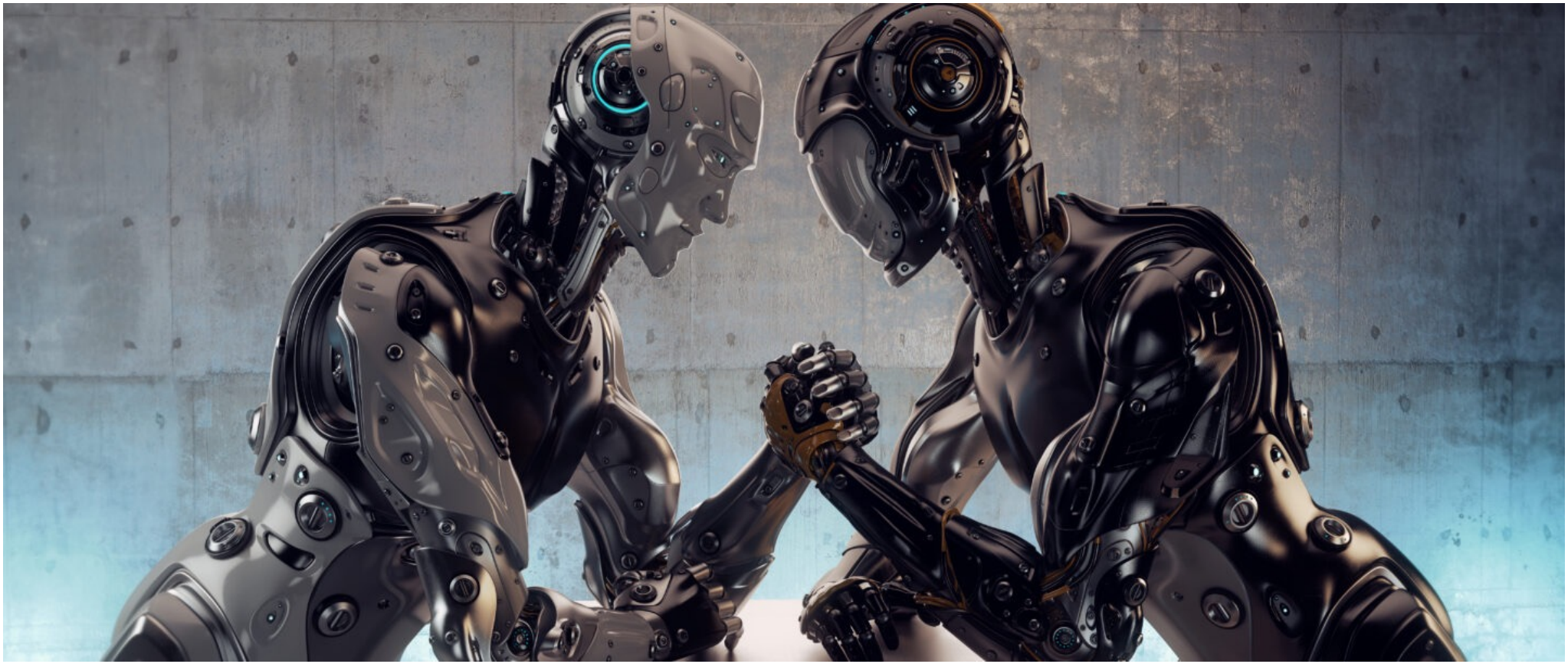


Carlini et al. totally broke 6 peered-reviewed defenses in ICLR'18 and won ICML'18 best paper award, where PGD-AT was the only defense that was partially robust against their attacks

2020

Croce et al. designed AutoAttack and established the first defense benchmark --- RobustBench

RobustBench
A standardized benchmark for adversarial robustness



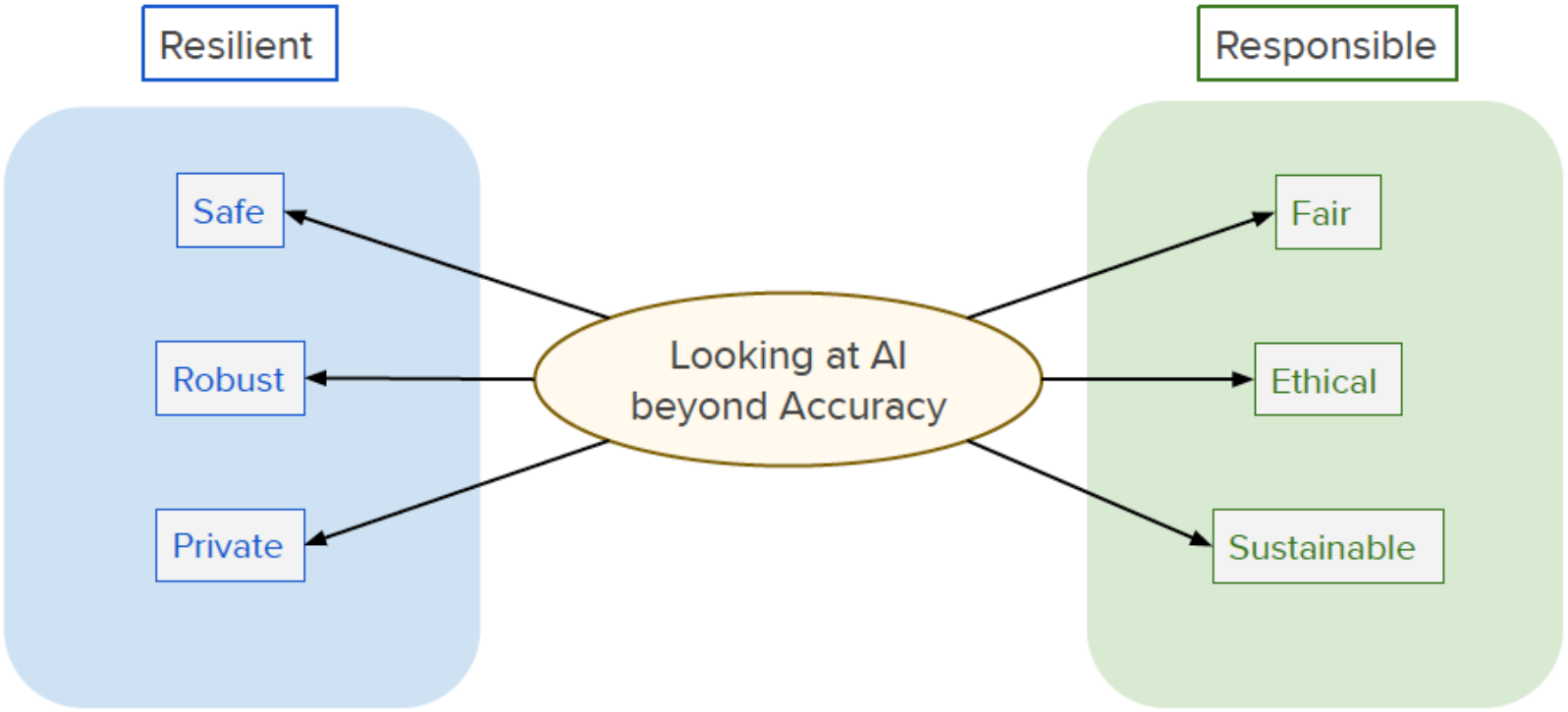
There always exists arms race between attackers and defenders

The gap between AI Development & Deployment

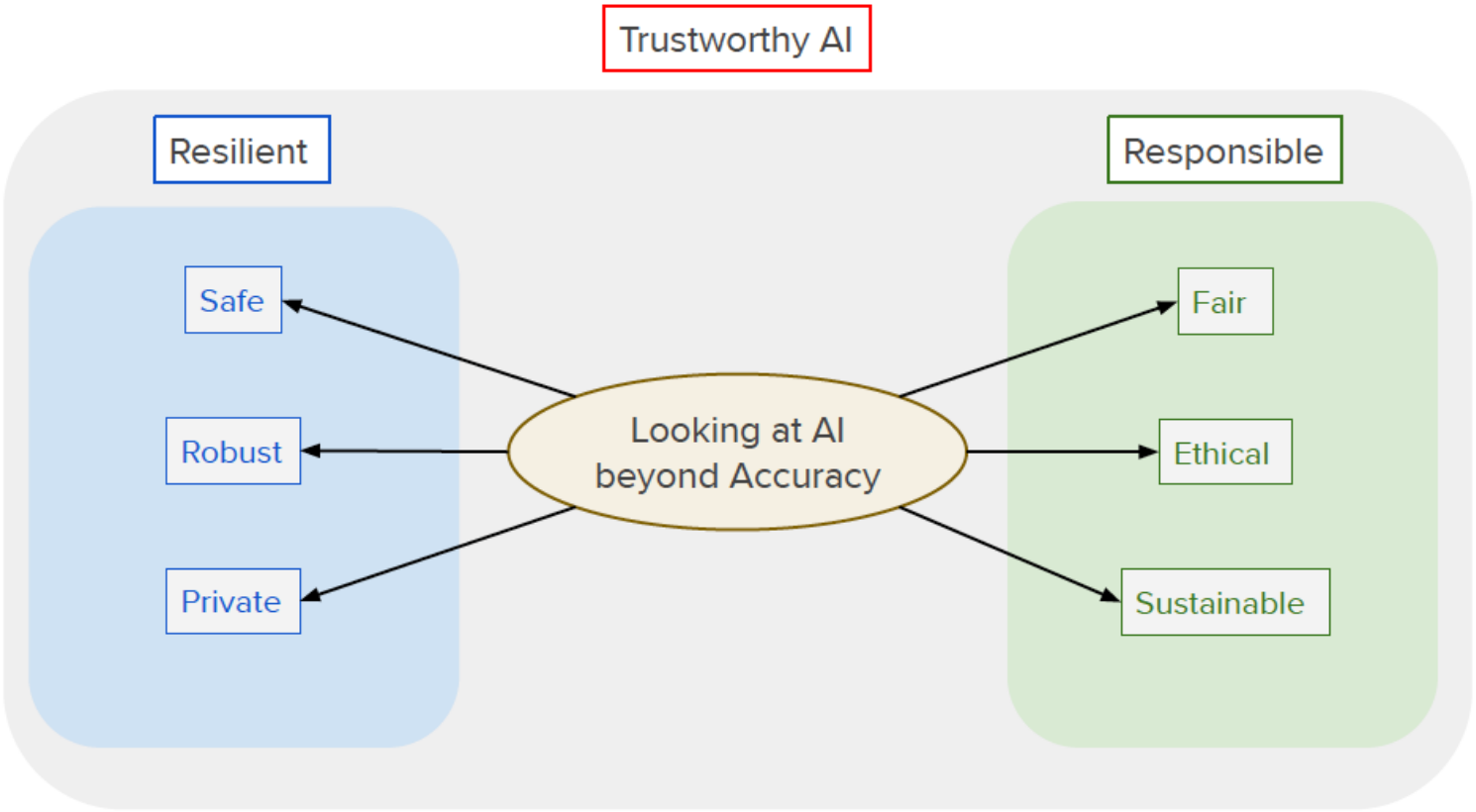
How we develop AI

How we deploy AI

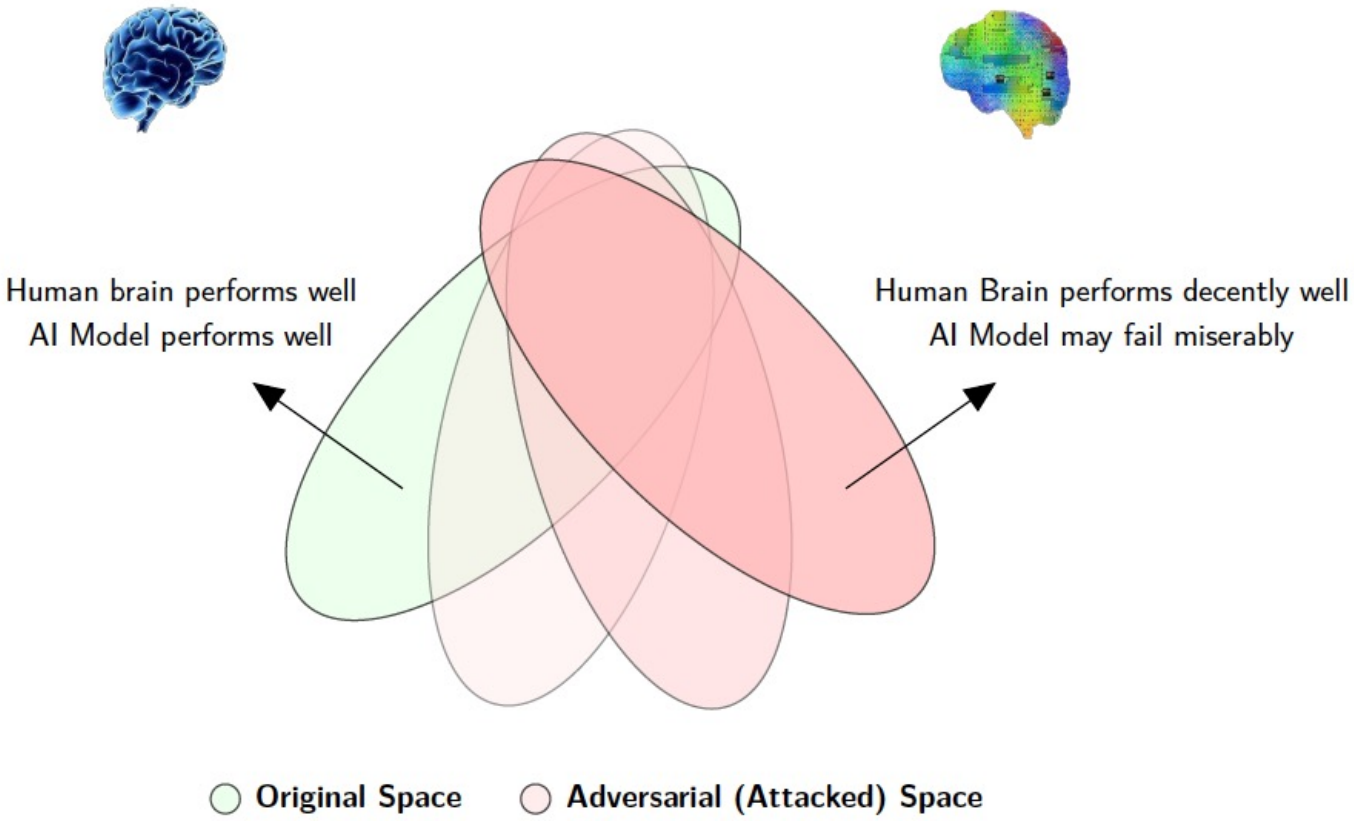
Beyond Accuracy in AI



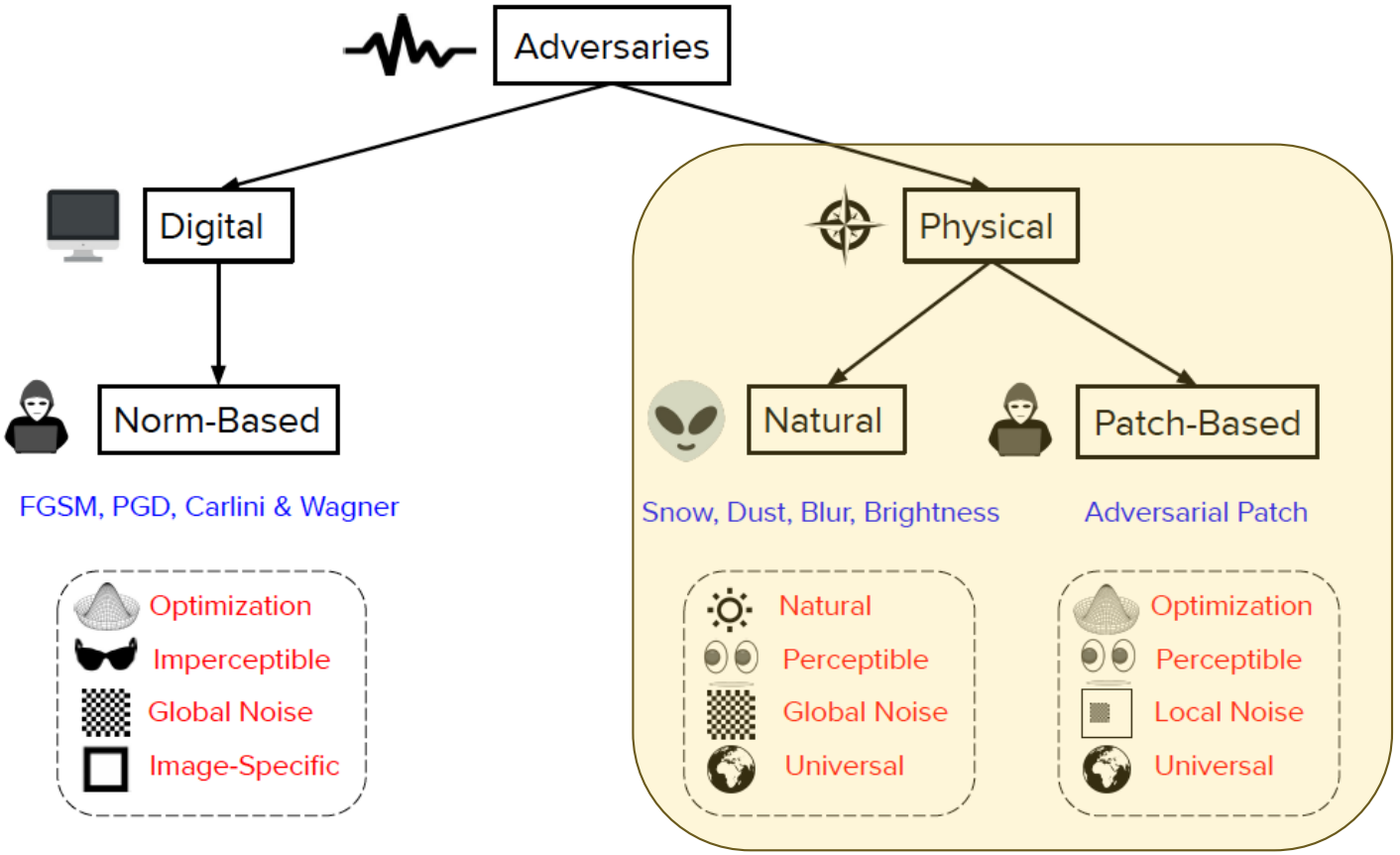
Beyond Accuracy in AI



Machines are Weaker...



Our Focus for this Talk



Resilience

against physical natural corruptions.

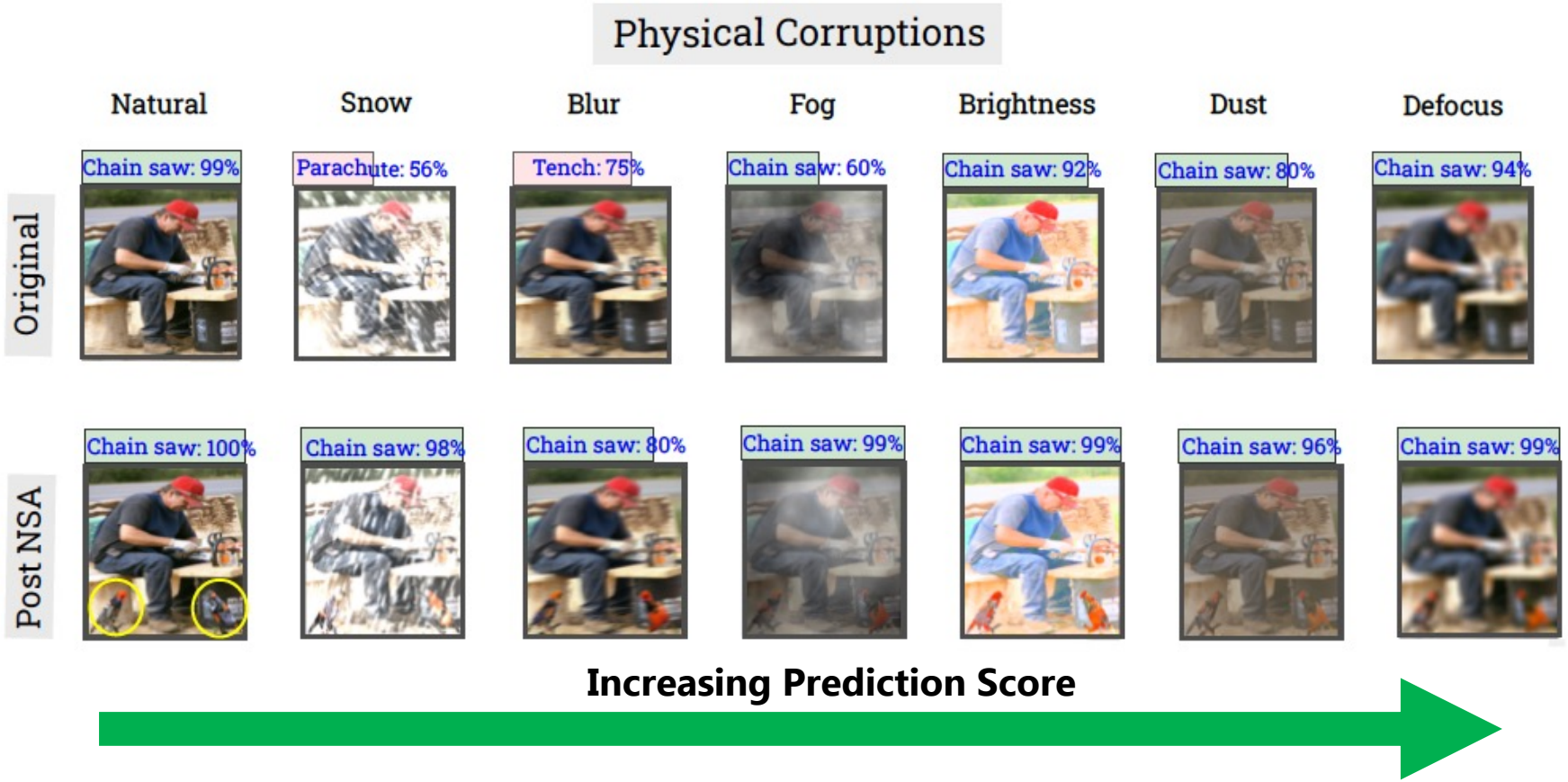
Naturalistic Support Artifacts (NSA)

Motivation

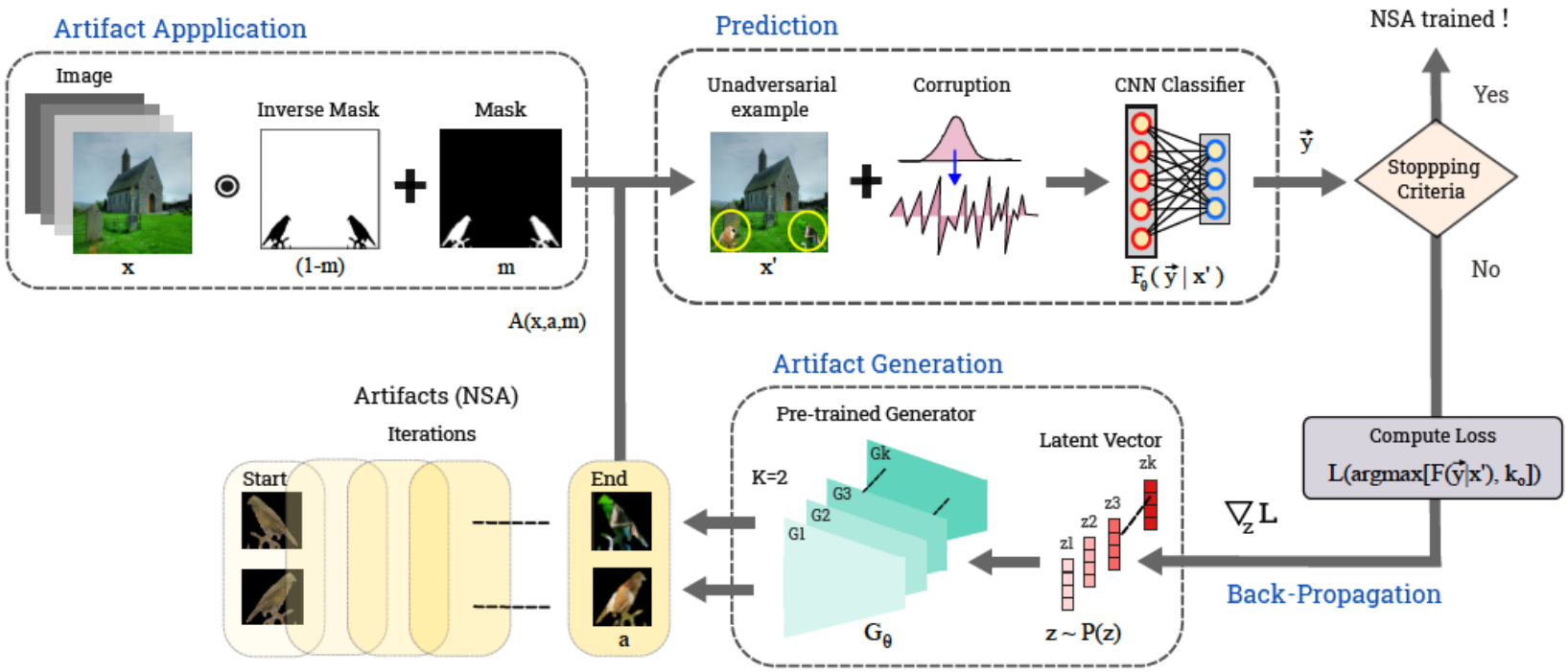
- Natural disturbances are quite common than adversarial attacks.
- No need of model training using image augmentations.
- Useful in cases where model parameters are inaccessible.

Image with naturalistic artifacts

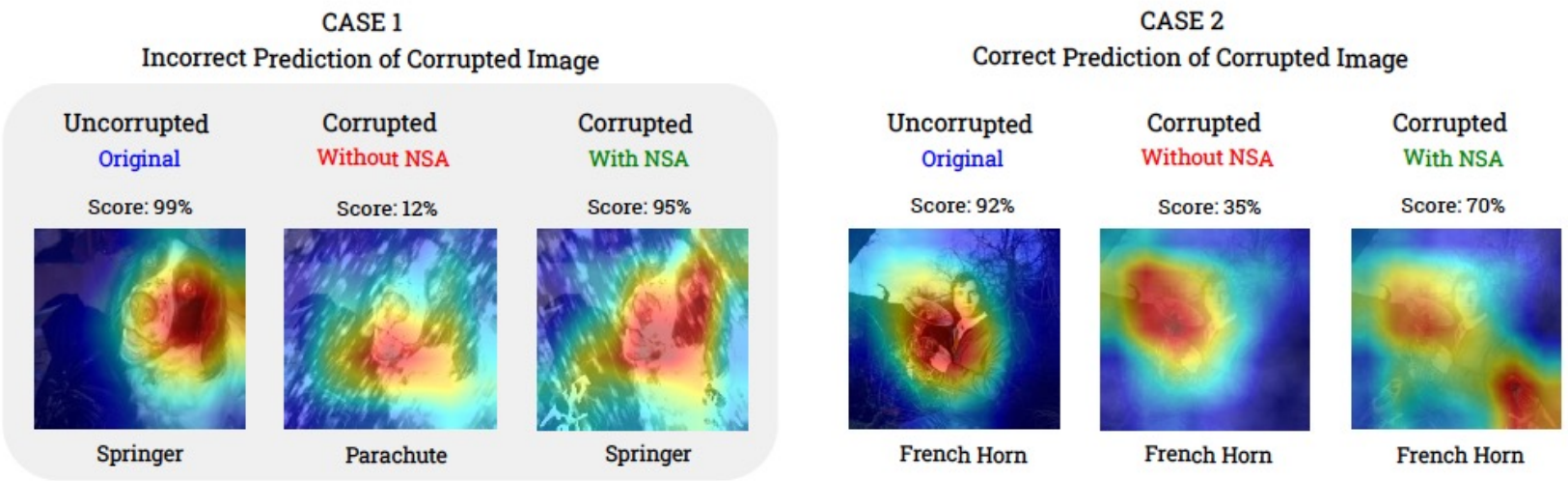
Naturalistic Support Artifacts (NSA)



Naturalistic Support Artifacts (NSA) Framework



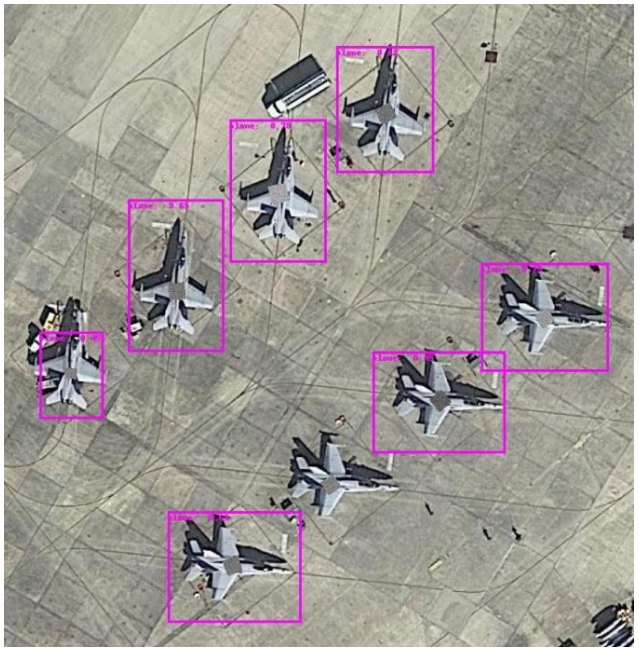
Naturalistic Support Artifacts - Visualization

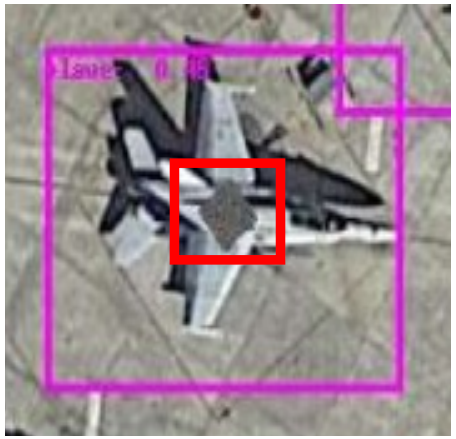


Multi-Patch Attack

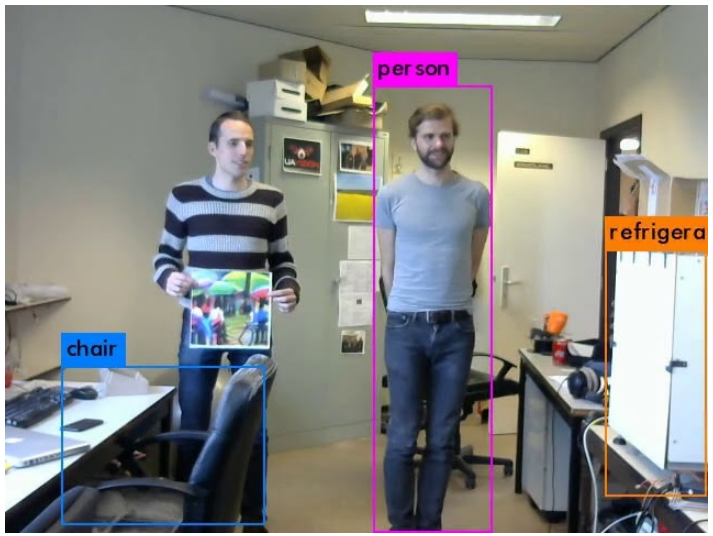
harnessing true potential of patch attacks.

Adversarial Patch Attacks





Adversarial Patch Attacks – Security



Adversarial Patch Attacks – Autonomous Driving

Lab (Stationary) Test

Physical road signs with adversarial perturbation under different conditions

Cropping, Resizing

Stop Sign → Speed Limit Sign

Field (Drive-By) Test

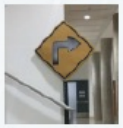
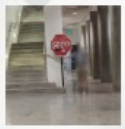
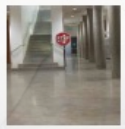
Video sequences taken under different driving speeds

Sample Per K Frames, Cropping, Resizing

Stop Sign → Speed Limit Sign

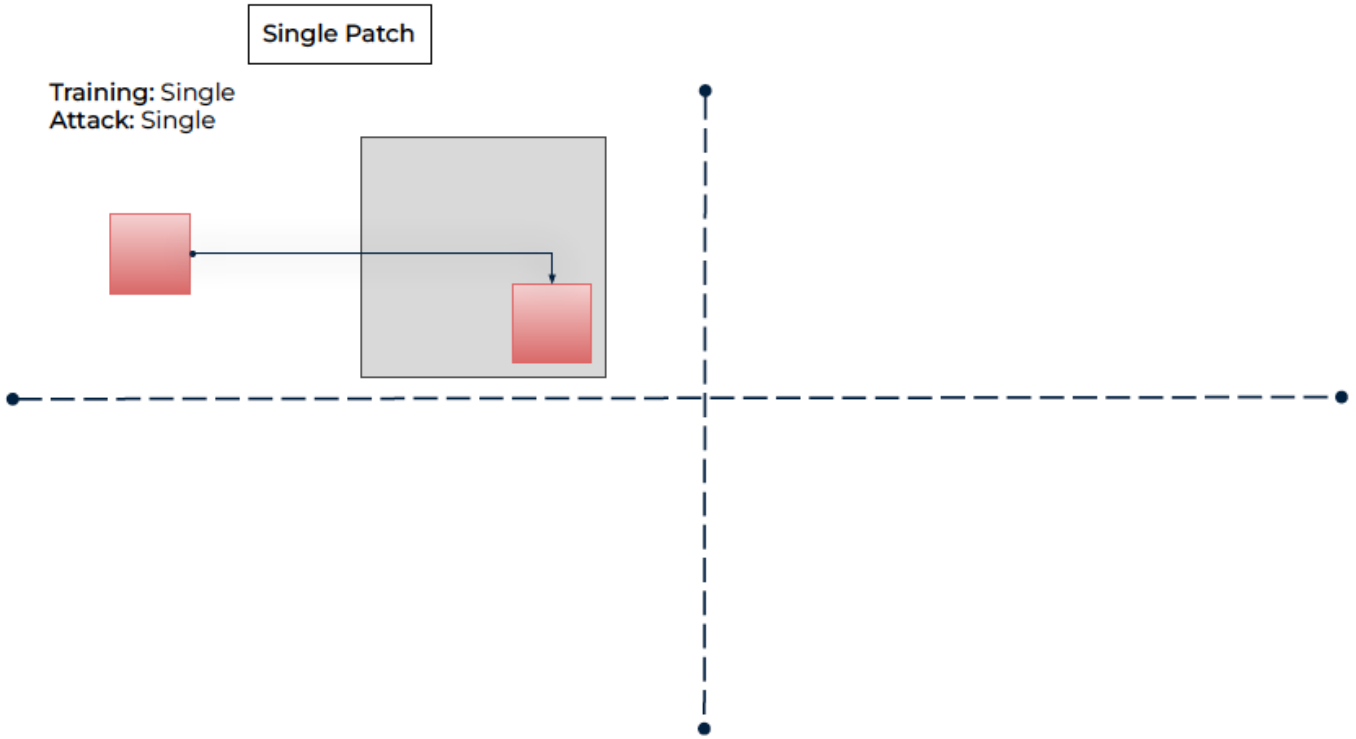
Picture from: Eykholt (2017) - Robust Physical-World Attacks on Deep Learning Visual Classification

Adversarial Patch Attacks – Autonomous Driving

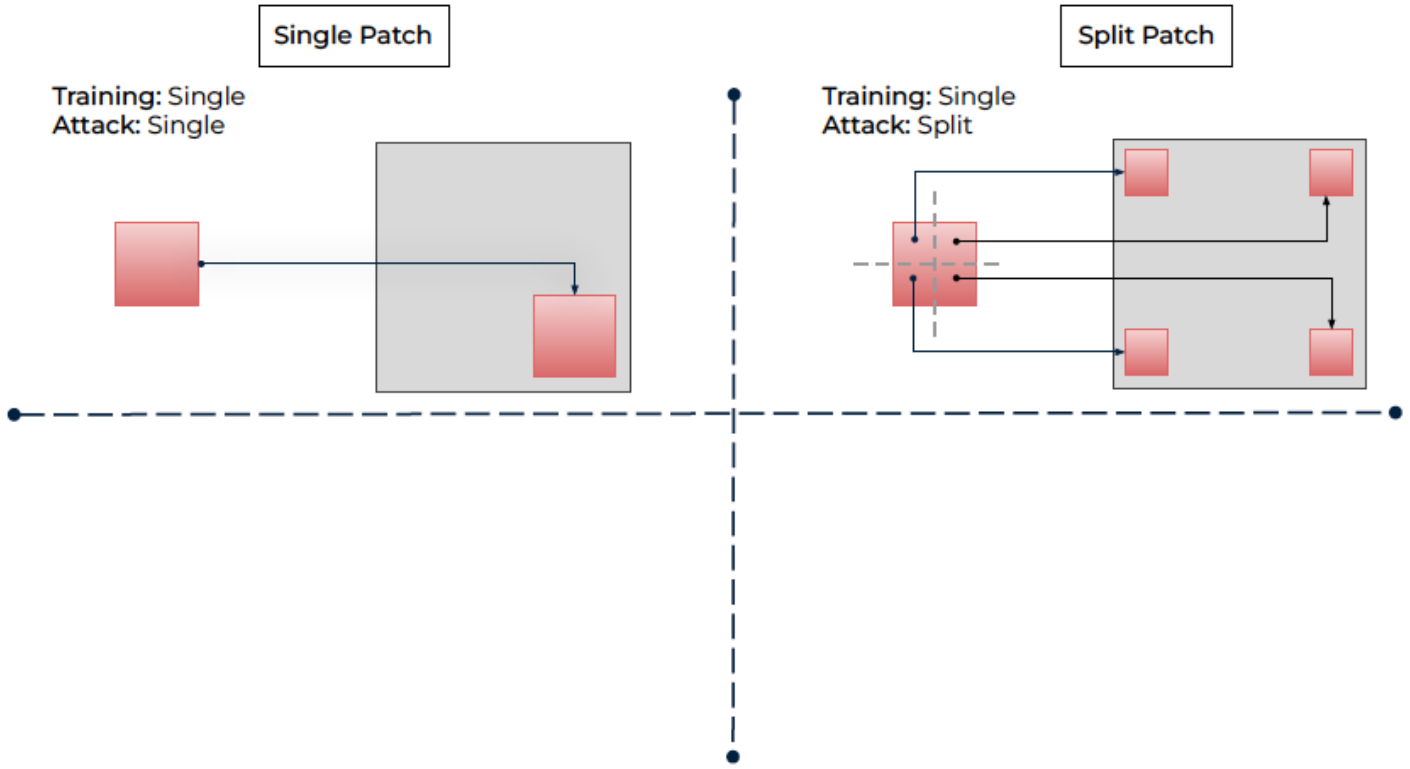
Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Picture from: Eykholt (2017) - Robust Physical-World Attacks on Deep Learning Visual Classification

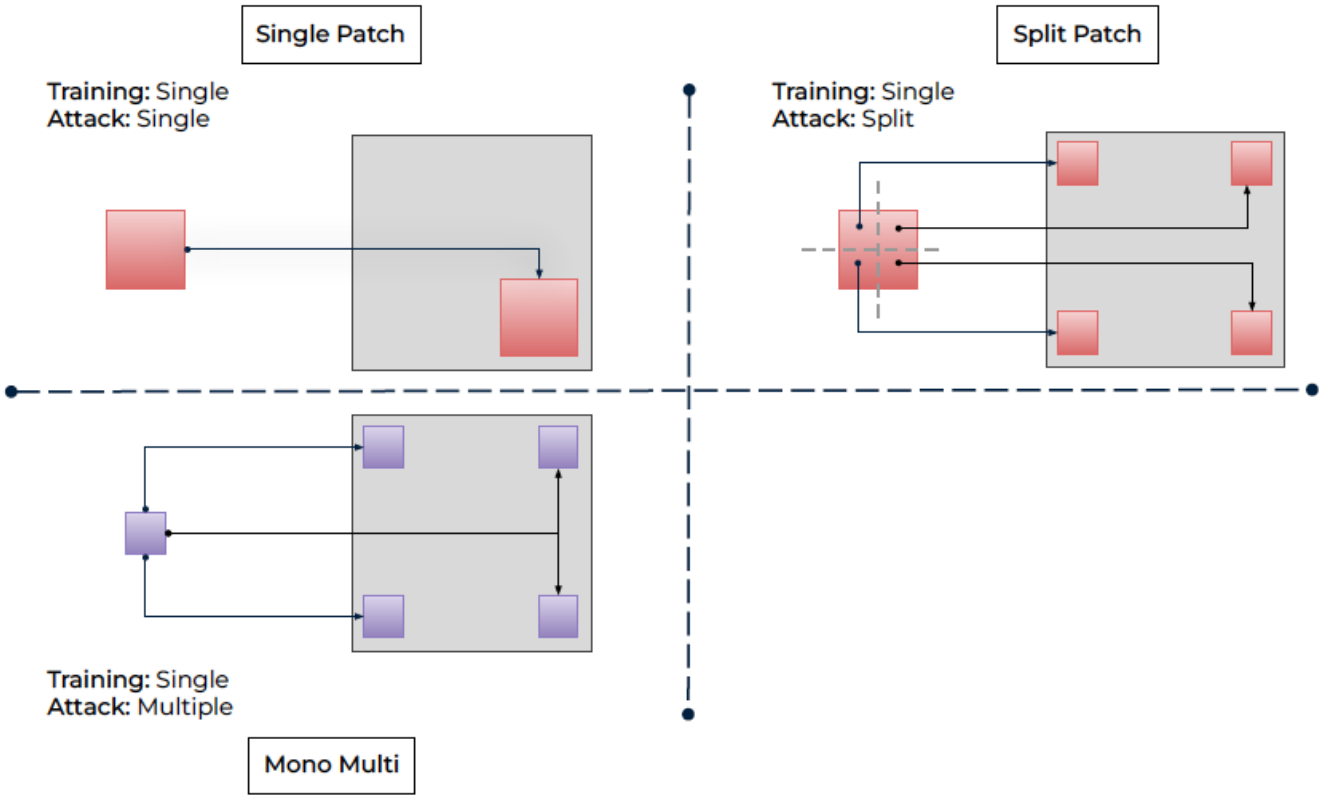
Types of Multi-Patch Attacks



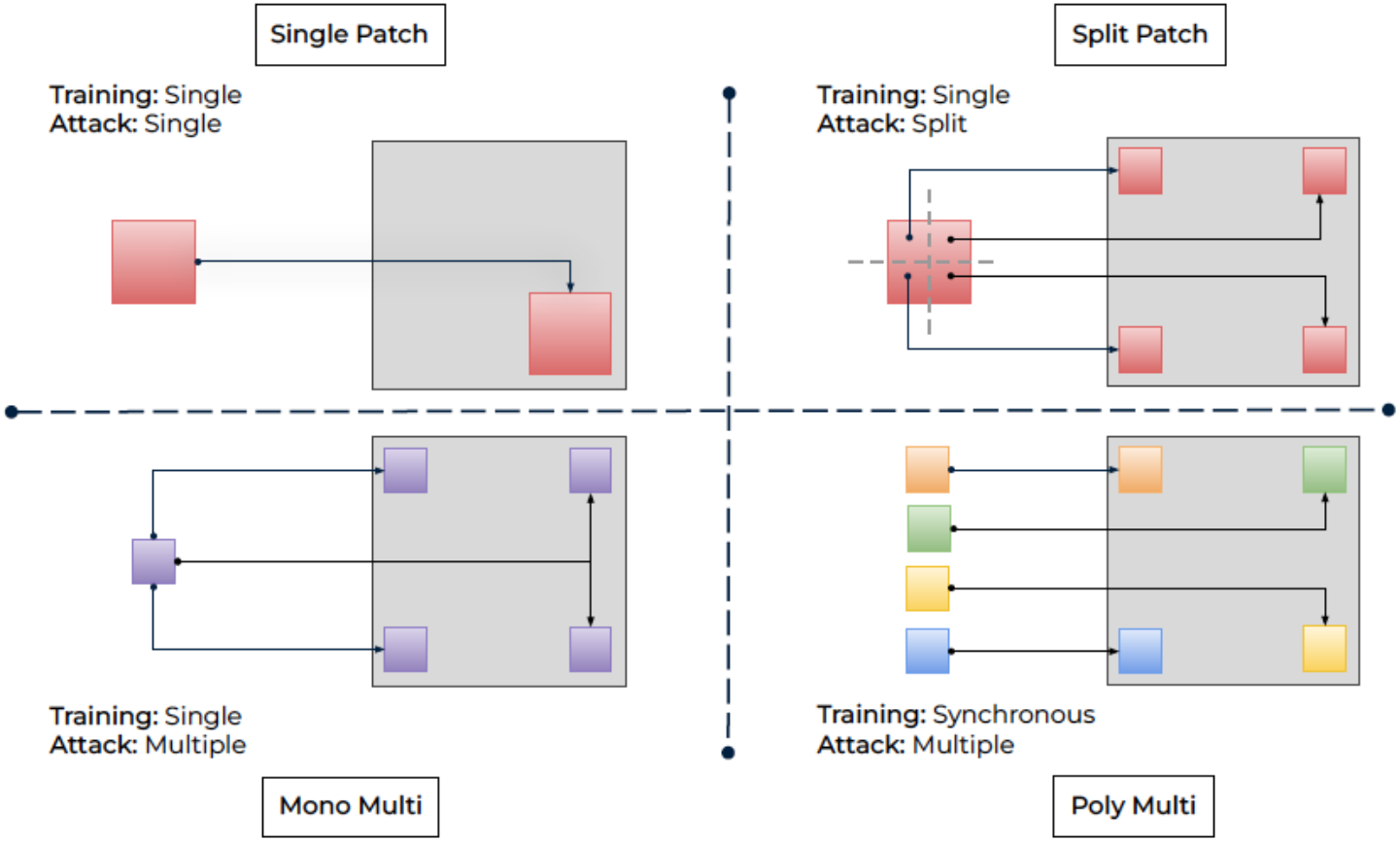
Types of Multi-Patch Attacks



Types of Multi-Patch Attacks



Types of Multi-Patch Attacks



Adversarial Multi-Patch Training

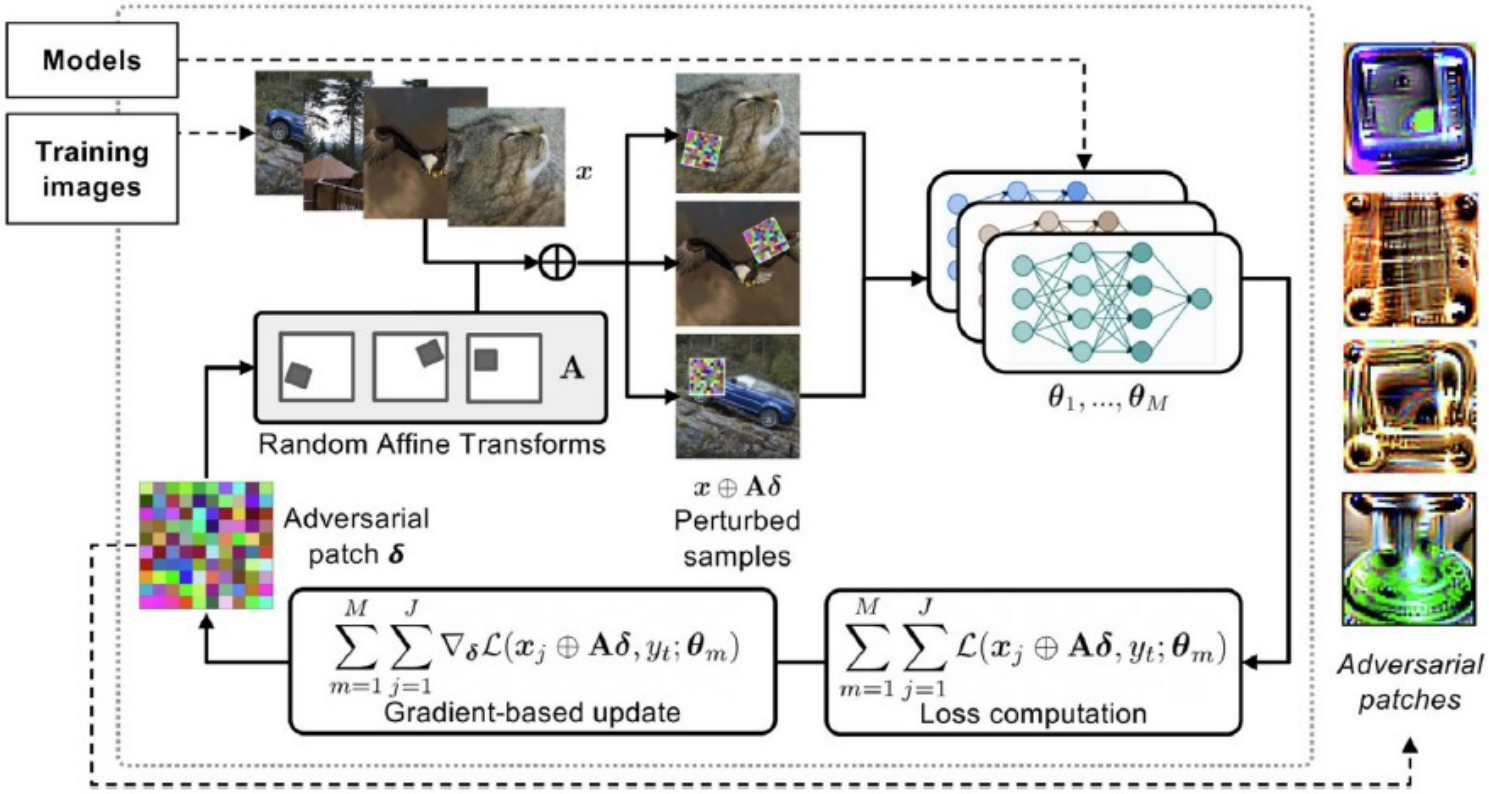


Image Source: Pintor et al., ImageNet-Patch: A dataset for benchmarking machine learning robustness against adversarial patches, Pattern Recognition, 2023

Results

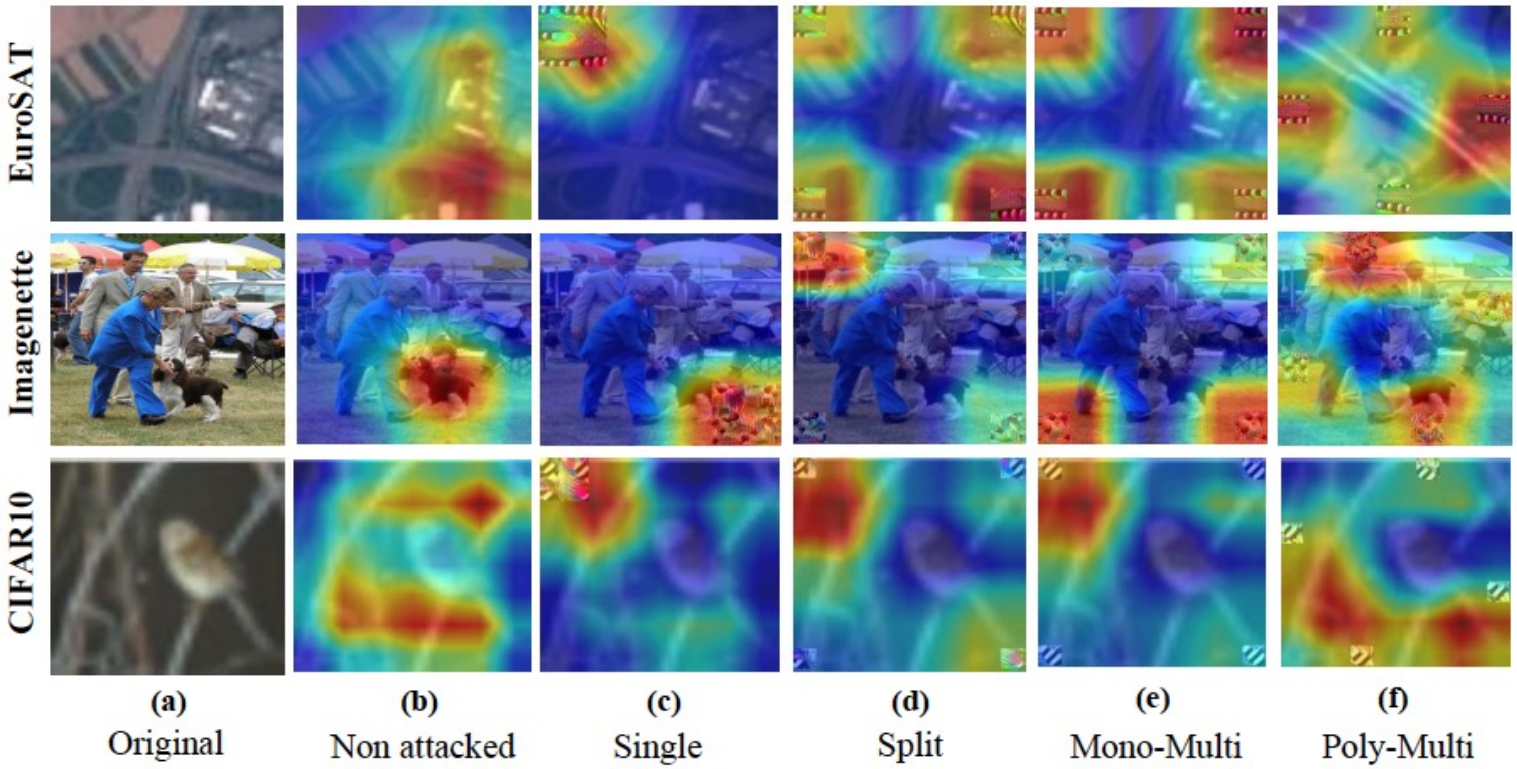
Attack Type	Perturbation Area: 12%						Perturbation Area: 8%						Perturbation Area: 4%					
	EuroSAT		Imagenette		CIFAR10		EuroSAT		Imagenette		CIFAR10		EuroSAT		Imagenette		CIFAR10	
	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.
Single	37.4	92.1	10.0	100	24.5	99.6	48.1	90.0	10.2	100	72.3	85.7	85.2	81.9	16.0	100	72.7	56.1
Split	56.8	63.2	10.6	100	64.9	43.6	87.1	44.5	30.3	100	80.9	36.8	96.1	33.6	70.7	99.2	89.0	25.0
Mono-Multi	28.2	94.4	10.0	100	34.4	68.6	58.5	58.0	10.6	100	52.3	30.7	77.2	45.8	22.5	100	76.2	15.9
Poly-Multi	25.4	94.2	0.0	100	14.6	96.0	43.1	80.0	0.0	100	21.2	77.3	71.5	81.6	2.2	100	52.0	46.1

Experimental results with VGG16 architecture. The clean accuracy of VGG16 is 96.8% on EuroSAT, 99.8% on Imagenette, and 91.1% on CIFAR10.

Attack Type	Perturbation Area: 12%						Perturbation Area: 8%						Perturbation Area: 4%					
	EuroSAT		Imagenette		CIFAR10		EuroSAT		Imagenette		CIFAR10		EuroSAT		Imagenette		CIFAR10	
	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.	Adv.	Tar.
Single	11.1	99.7	10.4	100	13.1	99.8	24.4	99.7	26.3	100	23.4	97.8	37.6	96.6	83.8	100	52.0	81.7
Split	30.2	92.6	14.7	100	30.7	89.1	44.7	70.2	23.9	100	46.4	67.3	76.5	58.4	72.5	100	67.5	42.1
Mono-Multi	18.3	90.2	10.0	100	16.1	99.8	32.9	80.9	10.0	100	28.4	87.3	46.4	71.4	32.2	100	37.7	72.6
Poly-Multi	0.7	99.1	0.0	100	0.0	100	3.4	97.1	0.0	100	0.2	100	30.7	93.8	0.2	100	11.4	99.0

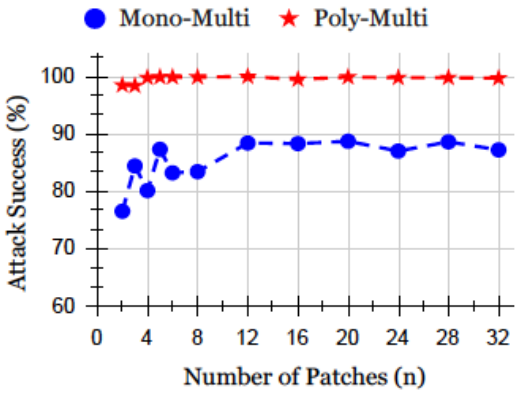
Experimental results with ResNet18 architecture. The clean accuracy is 96.7% on EuroSAT, 99.8% on Imagenette, and 93.3% on CIFAR10

Results - Visuals



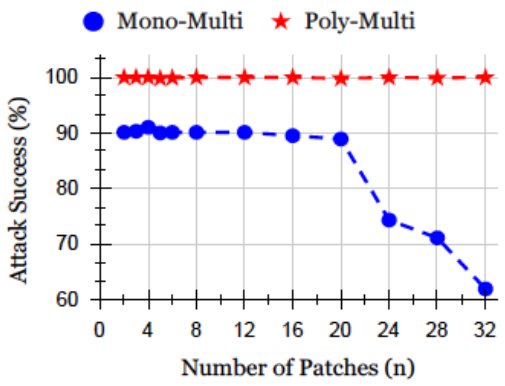
Grad-CAM visualization presenting ResNet18's perception of clean, single patch and Multi-Patch attacked images from EuroSAT, Imagenette, and CIFAR10 datasets.

Patch Number Trade-off



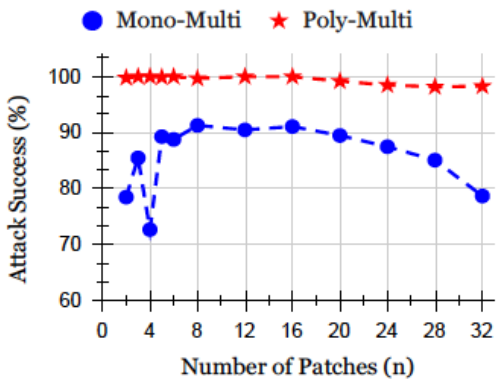
(a)

EuroSAT



(b)

Imagenette



(c)

CIFAR10

Abhijith Sharma, Yijun Bian, Vatsal Nanda, Phil Munz, and Apurva Narayan. 2023. Vulnerability of CNNs against Multi-Patch Attacks. In Proceedings of the 2023 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems (SaT-CPS '23). Association for Computing Machinery, New York, NY, USA, 23–32. https://doi.org/10.1145/3579988.3585054

Performance Against SOTA

Defense Name	Original		Single Patch		Multi Patch	
	w/o. Def	w. Def	w/o. Def	w. Def	w/o. Def	w. Def
	Perturbation 12%					
Patch Cleanser	93.9	80.1	0.2	43.7	0.0	**0.0**
LGS	99.7	90.3	12.0	52.9	0.0	**0.1**
	Perturbation 8%					
Patch Cleanser	93.9	85.2	16.3	57.4	0.0	**0.0**
LGS	99.7	90.3	17.4	33.0	0.0	**0.3**
	Perturbation 4%					
Patch Cleanser	93.9	88.3	71.8	78.9	0.9	**1.0**
LGS	99.7	90.3	82.2	56.8	1.1	**16.3**

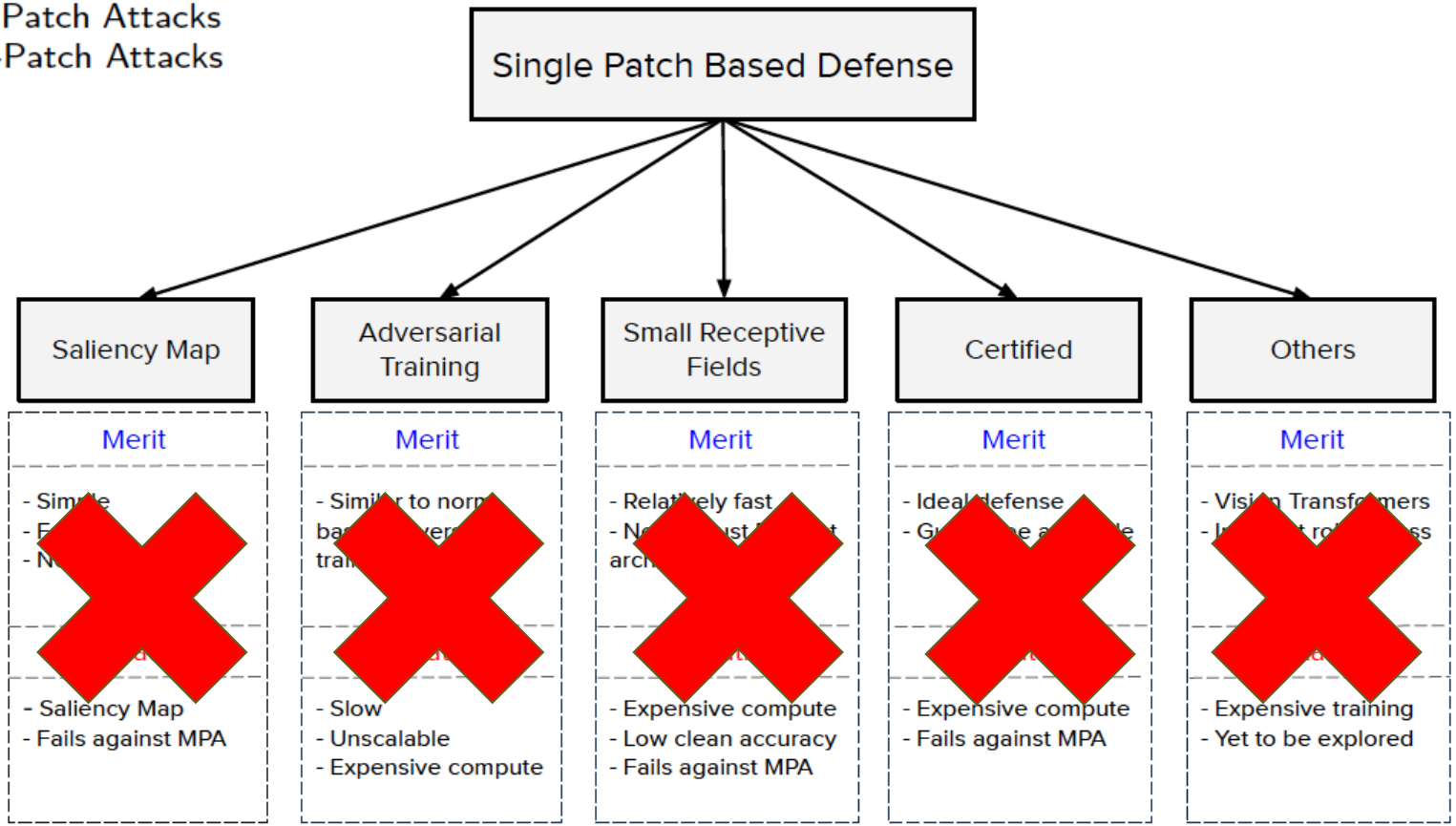
Experimental results with ResNet 18 architecture.

Defense Name	Original		Single Patch		Multi Patch	
	w/o. Def	w. Def	w/o. Def	w. Def	w/o. Def	w. Def
	Perturbation 12%					
Patch Cleanser	94.4	77.6	0.0	50.1	0.0	**0.0**
LGS	99.7	90.2	0.0	7.1	0.0	**0.1**
	Perturbation 8%					
Patch Cleanser	94.4	80.4	0.0	61.5	0.0	**0.0**
LGS	99.8	90.2	0.2	2.2	0.2	**0.5**
	Perturbation 4%					
Patch Cleanser	94.4	83.0	4.9	72.9	12.3	**4.2**
LGS	99.8	90.2	5.1	**6.8**	11.3	22.3

Experimental results with VGG16 architecture.

What are the defenses against patch attacks ?

SPA: Single Patch Attacks
MPA: Multi-Patch Attacks

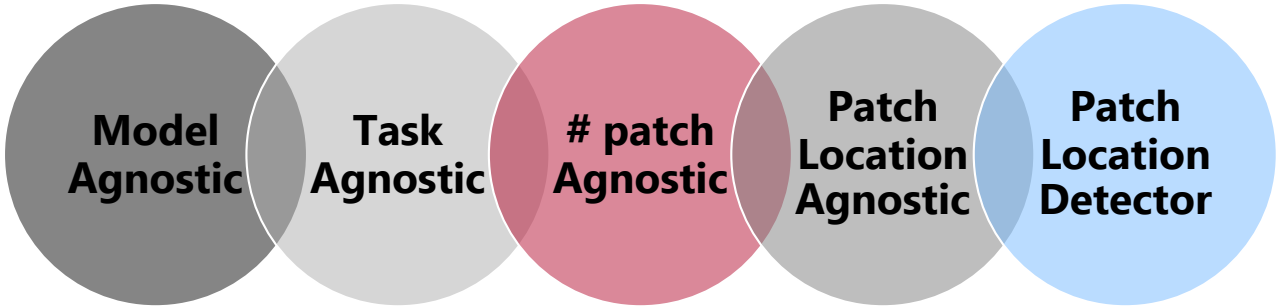


Total Variation Based Image Resurfacing

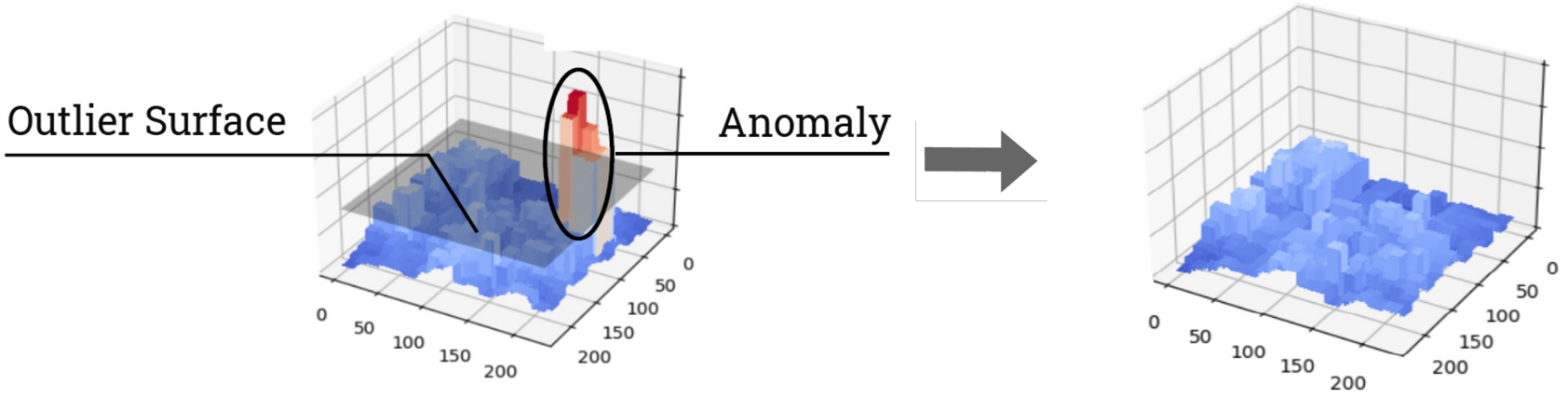
The total variation loss/score is a metric to determine the complexity of an image to its spatial variation in pixel values.

$$\mathcal{TV}(\mathcal{X}) = \sum_{i,j \in \mathcal{N}} ||x_i - x_j||_p^q$$

TVR: Total Variation Based Image Resurfacing
Outliers in the TV loss across the image landscape are removed to resurface the image

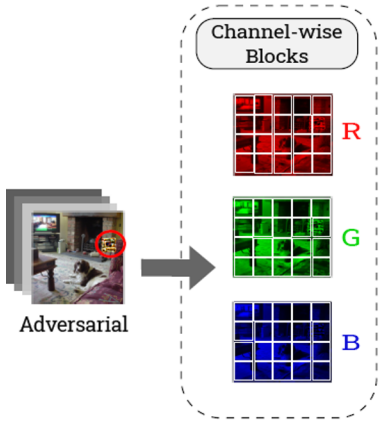


Total Variation Based Image Resurfacing

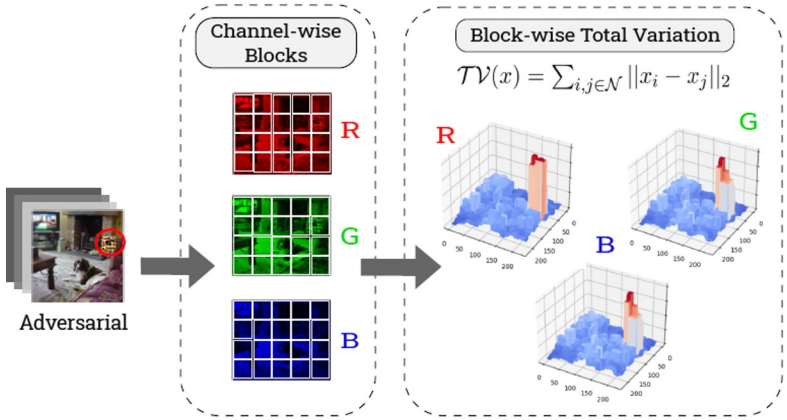


Outliers in the TV loss across the image landscape are removed to resurface the image.

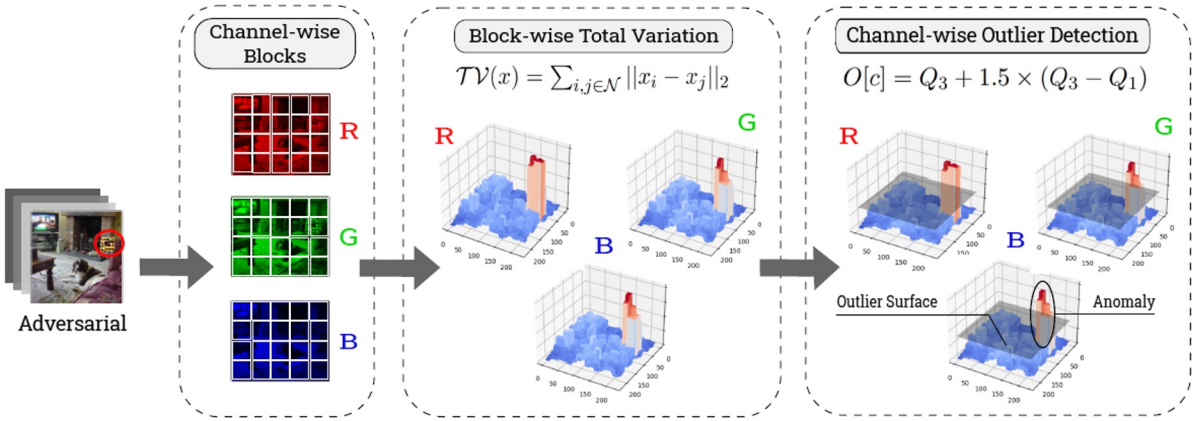
TVR - Framework



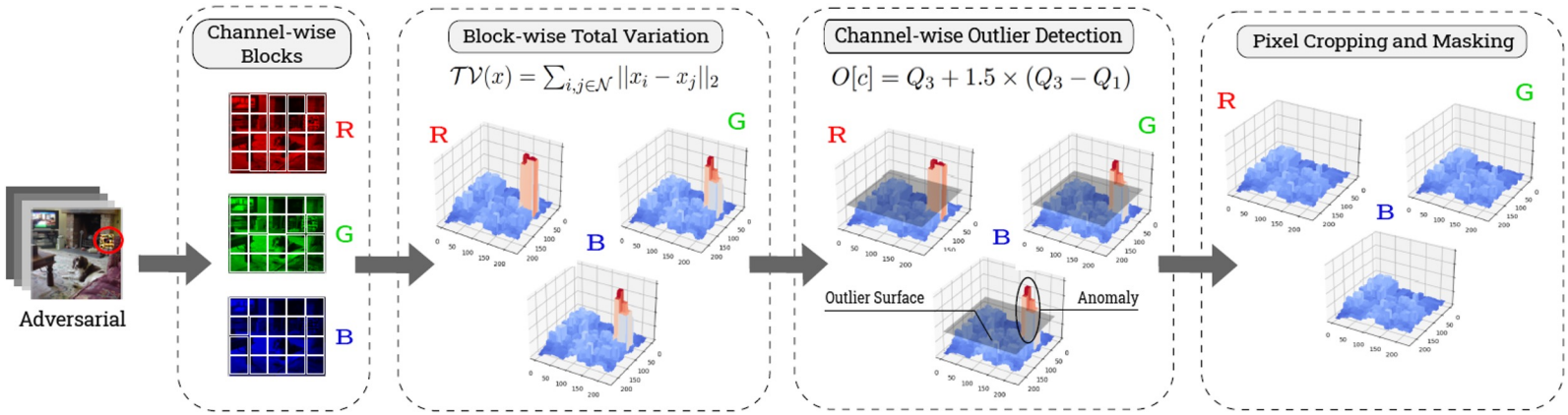
TVR - Framework



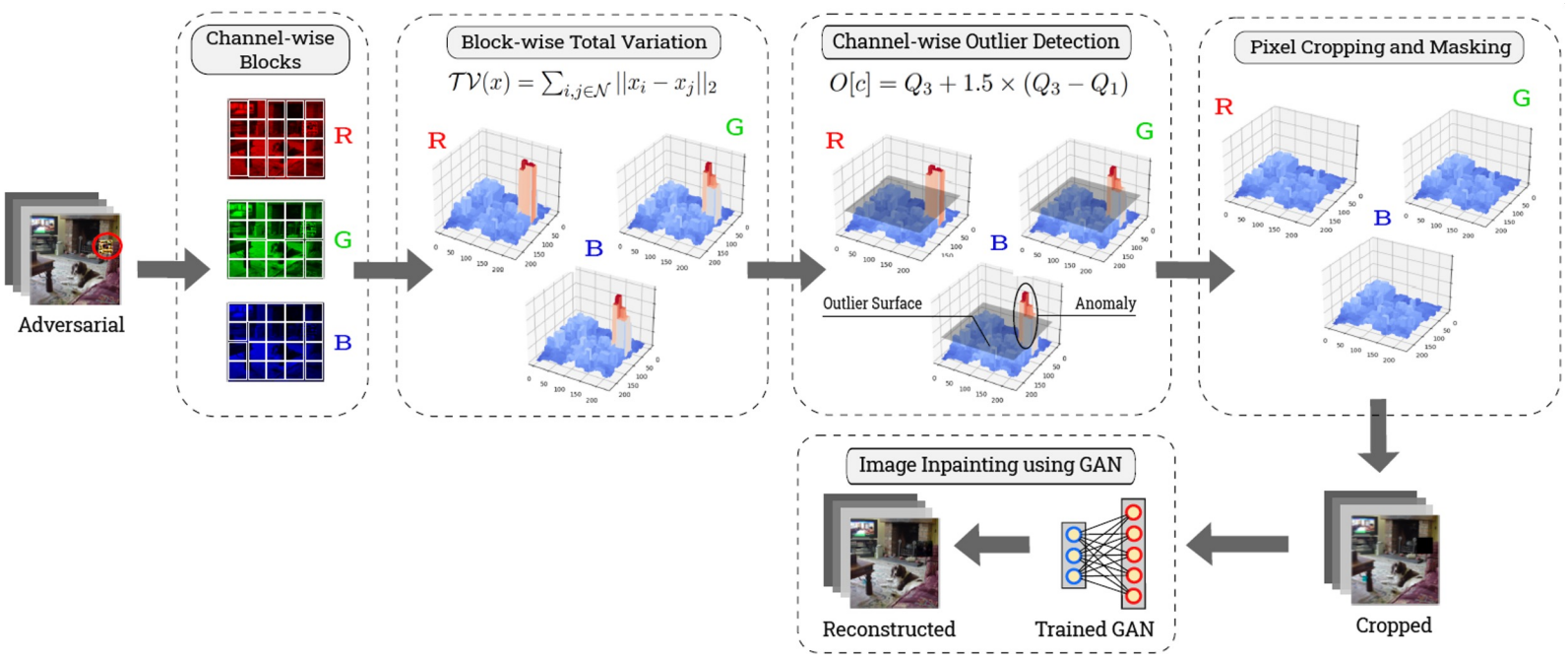
TVR - Framework



TVR - Framework



TVR - Framework



Results

Model	Naive			TVR		
	Natural	Adversarial	Target	Natural	Adversarial	Target
AlexNet	86	39.8	14.4	86	75.6	0.2
ResNet18	96	69.1	25.3	96	91.1	0
SqueezeNet	86	50.9	28.4	86	72	0.7
VGG16	96	71.3	12.9	94	89.8	0.2
GoogleNet	94	74.4	6.9	94	84.9	0
Inception v3	92	62.7	13.6	88	81.8	0
Ensemble	**89**	**53.3**	**22.7**	**89.3**	**79.6**	**0.3**
Overall	**92**	**61.4**	**16.9**	**90.7**	**82.5**	**0.2**

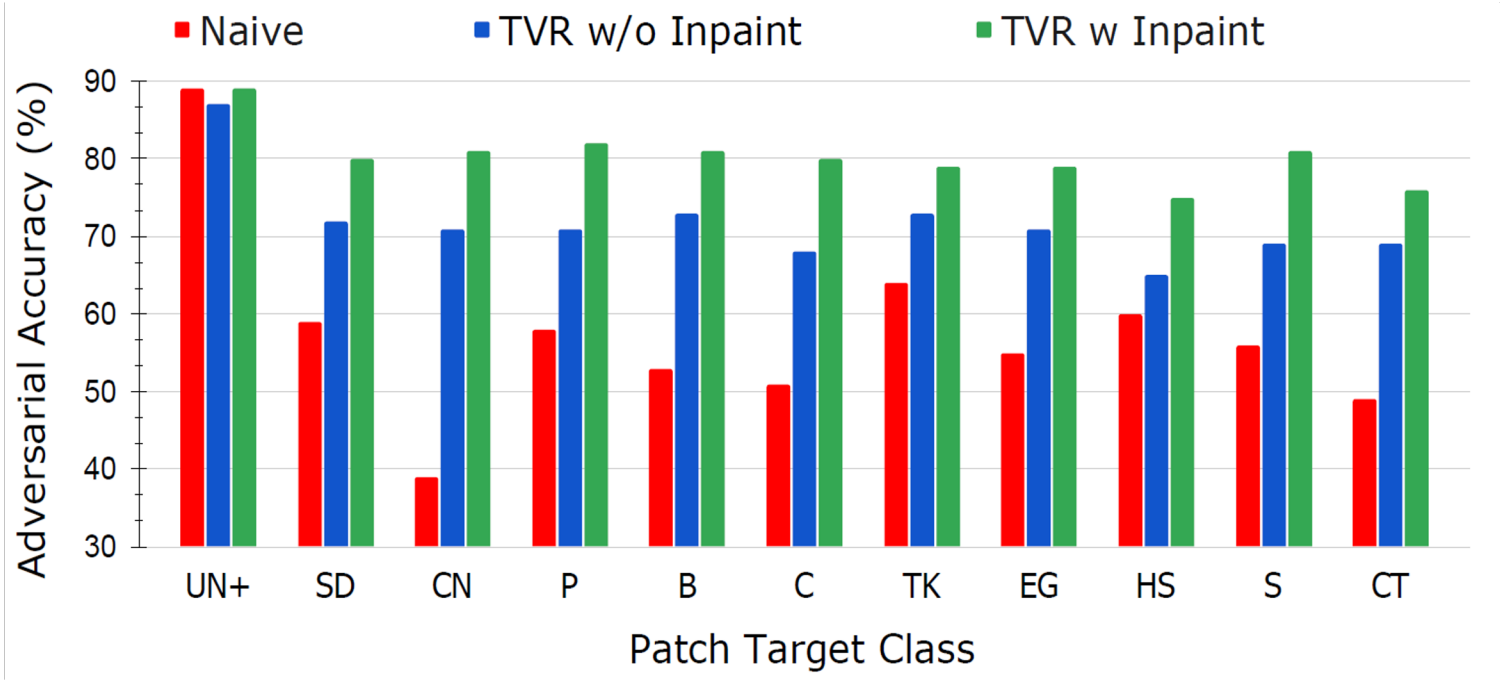
Dataset: ImageNet Patch – Benchmark, Adversarial Patch: Size 5-6%

How do we do against other defenses ?

	Defense Method	Clean	Single	Multi-2	Multi-3	Multi-4
AlexNet	Naive	**65**	31	4	2	0
	LGS [23]	61	**56**	41	18	17
	PatchCleanser [40]	56	53	30	8	2
	TVR (ours)	56	56	**40**	**27**	**20**
ResNet18	Naive	**78**	57	5	0	0
	LGS [23]	68	65	56	22	4
	PatchCleanser [40]	74	46	16	4	1
	TVR (ours)	76	**72**	**52**	**42**	**23**
SqueezeNet	Naive	**57**	18	0	0	0
	LGS [23]	52	**48**	30	9	4
	PatchCleanser [40]	54	20	4	3	0
	TVR (ours)	50	**49**	**42**	**22**	**12**
VGG16	Naive	**75**	39	4	4	0
	LGS [23]	73	64	42	24	11
	PatchCleanser [40]	75	44	17	6	2
	TVR (ours)	74	**66**	**46**	**43**	**24**
GoogleNet	Naive	**74**	54	29	20	10
	LGS [23]	73	54	46	37	20
	PatchCleanser [40]	72	53	30	8	2
	TVR (ours)	73	**62**	**58**	**42**	**25**
Inception	Naive	**76**	50	38	25	16
	LGS [23]	74	62	47	42	29
	PatchCleanser [40]	70	55	31	10	8
	TVR (ours)	72	**65**	**60**	**48**	**33**

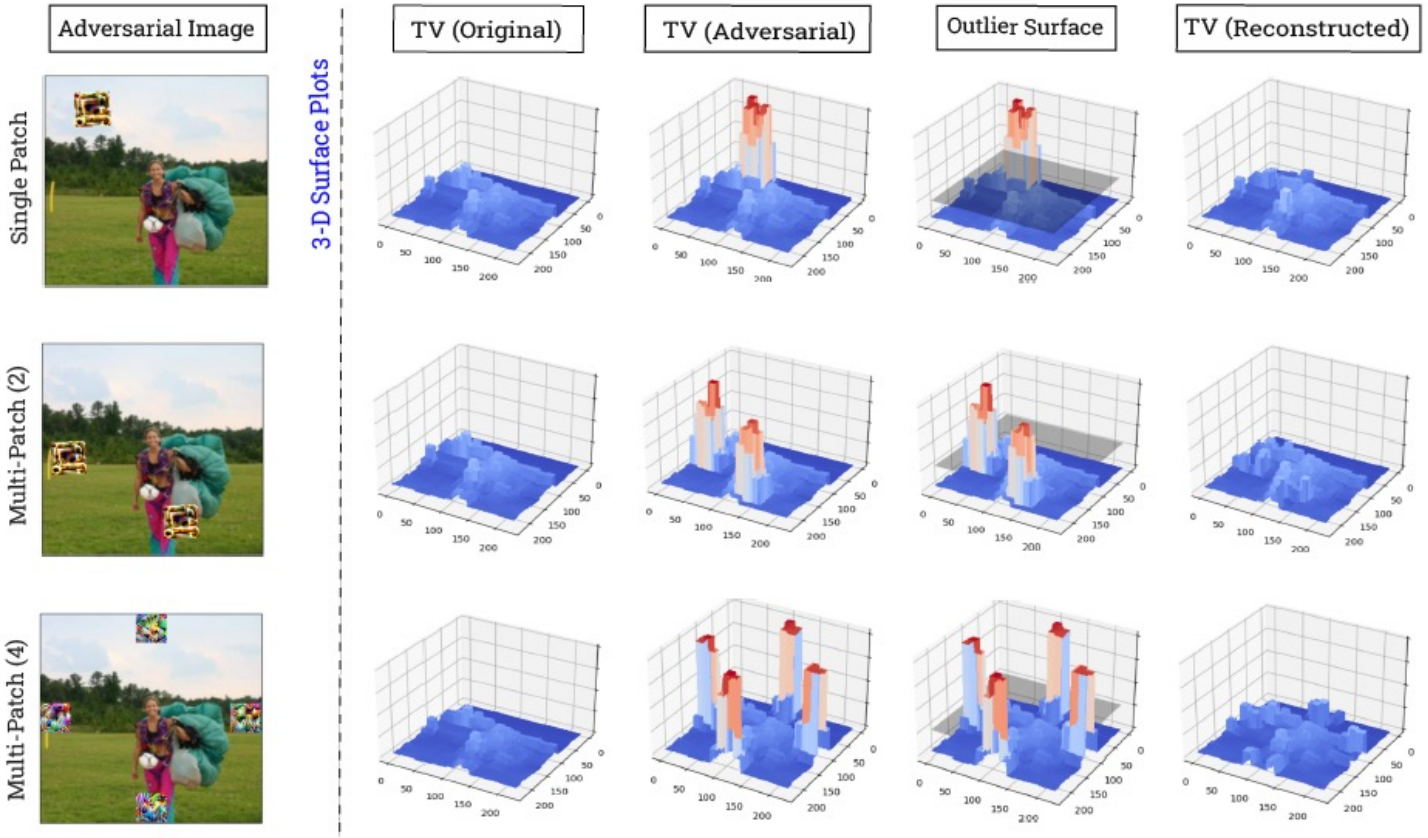
Target Class = Banana, Test images = 100

Image Inpainting Analysis

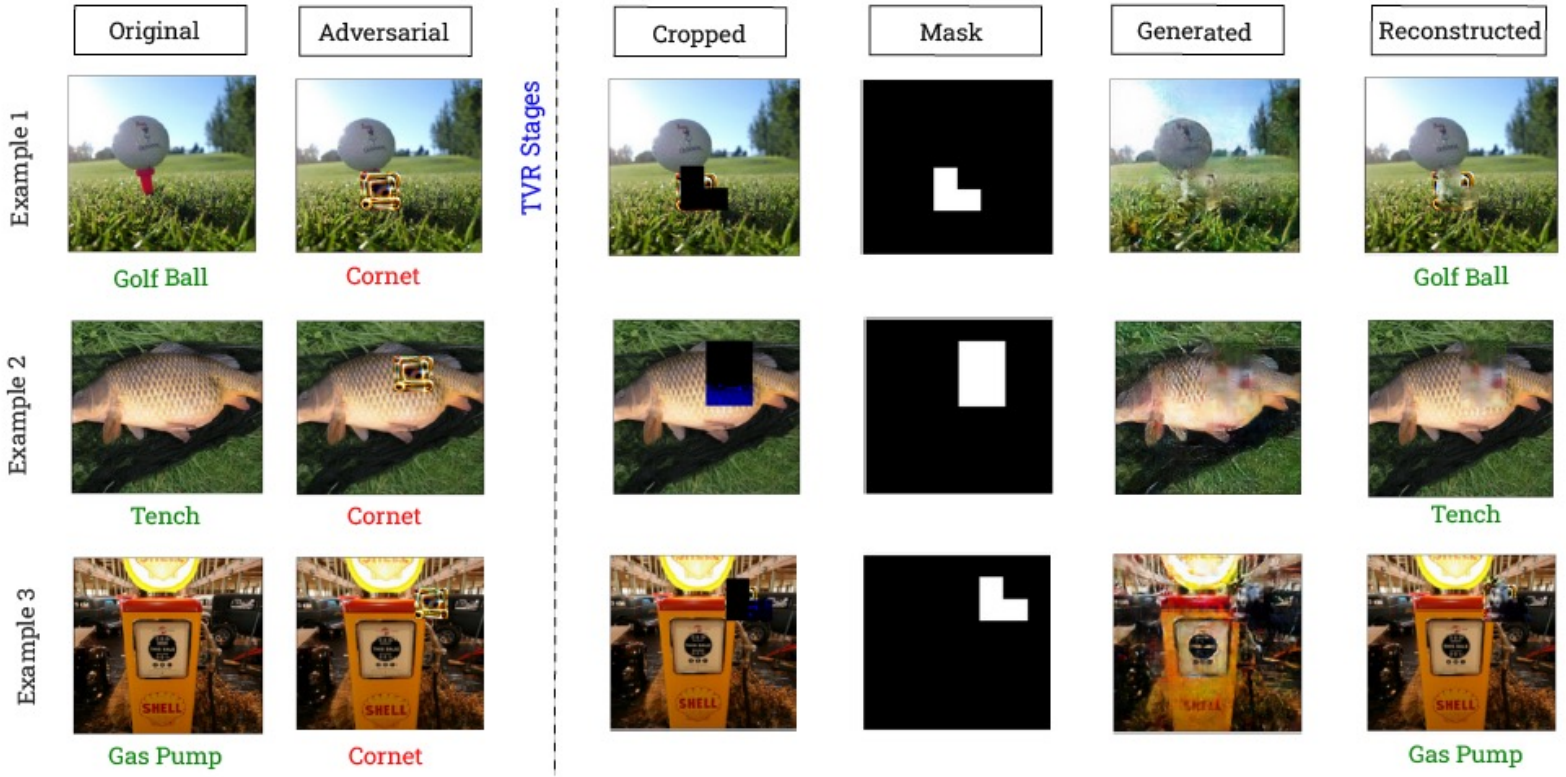


UN+: Unattacked, SD: Soap Dispenser, CN: Cornet, P: Plate, B: Banana, C: Cup, TK: Typewriter, EG: Electric Guitar, HS: Hair Spray, S: Socks, CT: Cellphone

TVR - Visualization

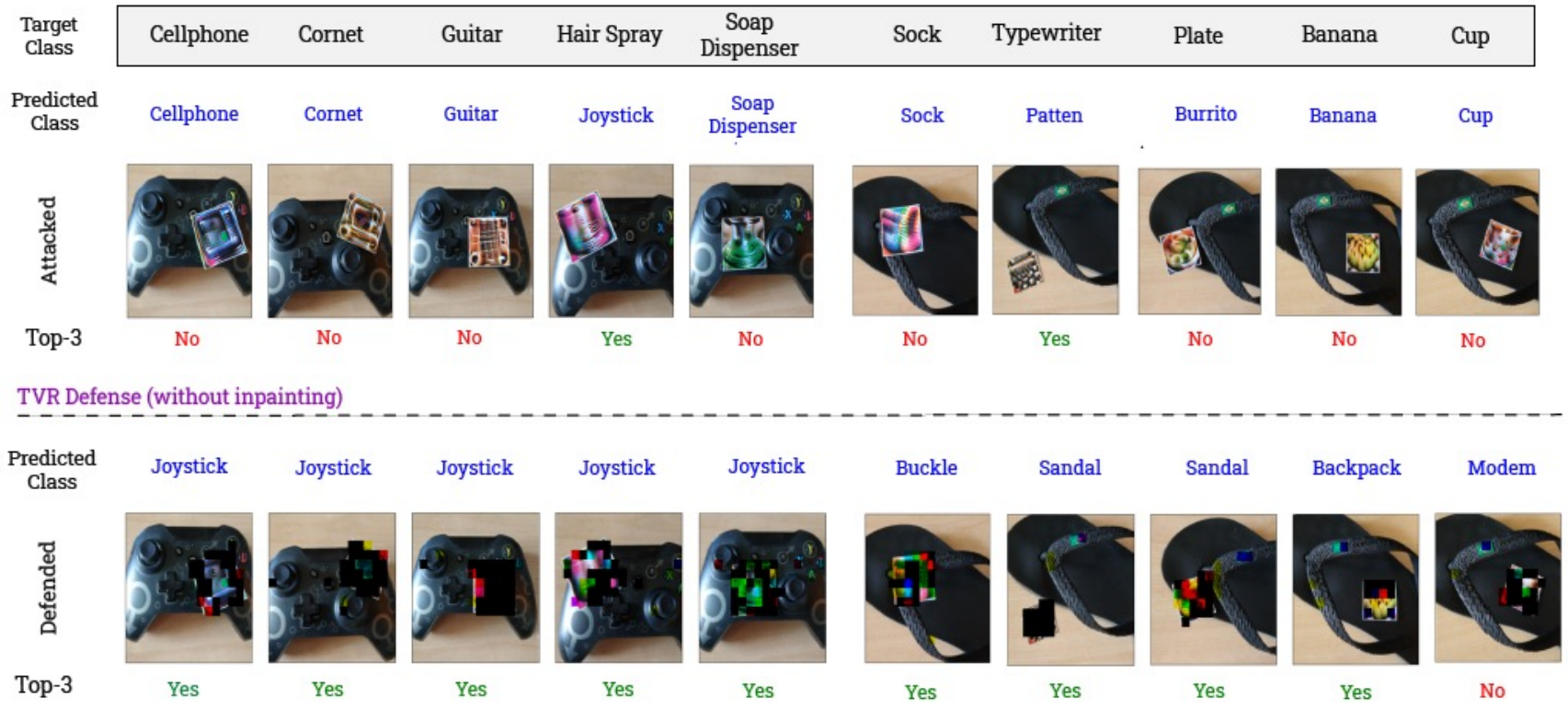


TVR - Demo

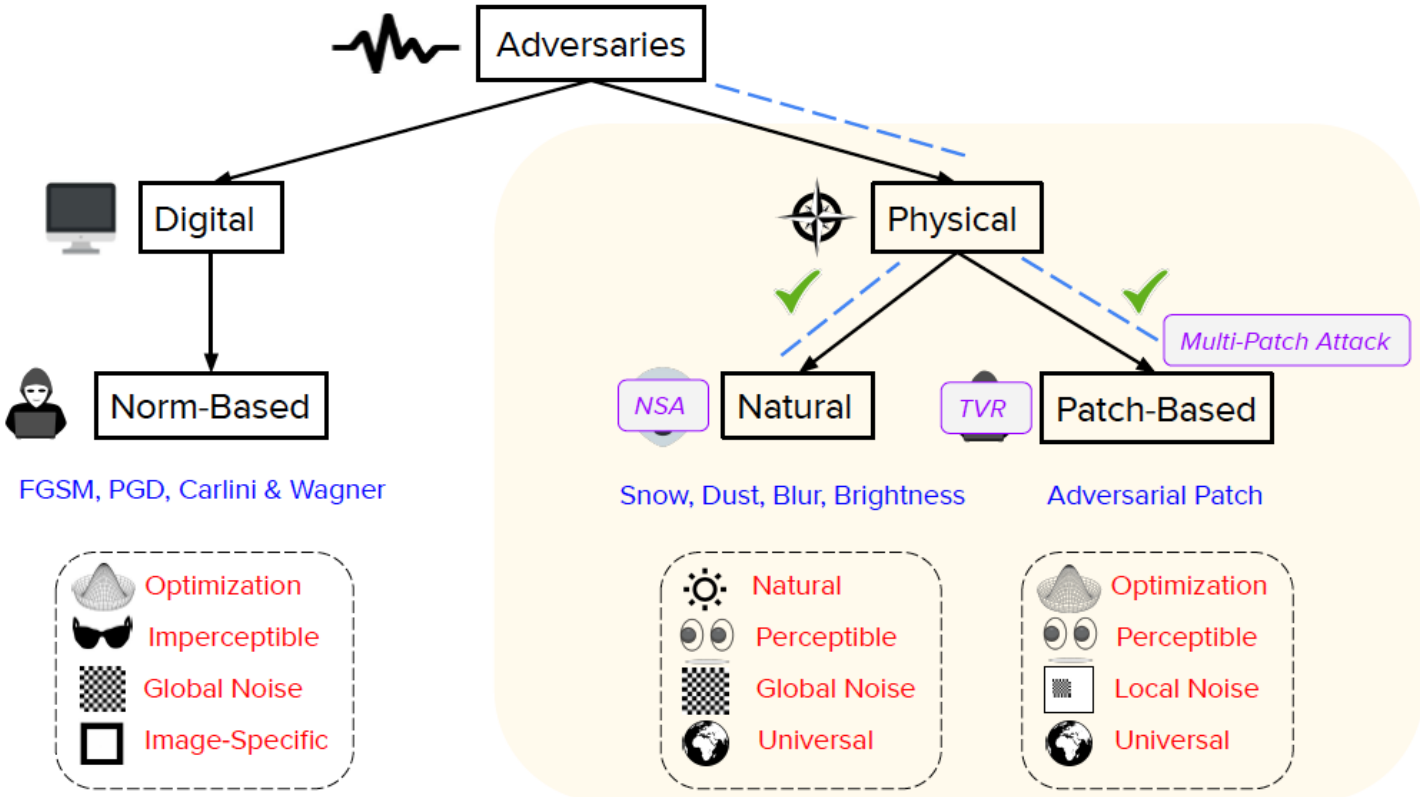


Link to Use the tool: https://t.ly/l1iUA

TVR – Physical Attacks



Summary


Adversaries
Digital
Norm-Based
FGSM, PGD, Carlini & Wagner
Optimization
Imperceptible
Global Noise
Image-Specific
Physical
Multi-Patch Attack
NSA
Natural
TVR
Patch-Based
Snow, Dust, Blur, Brightness
Adversarial Patch
Natural
Perceptible
Global Noise
Universal
Optimization
Perceptible
Local Noise
Universal

Summary

⚠️ We warn against the potential use of Multi-Patch attacks

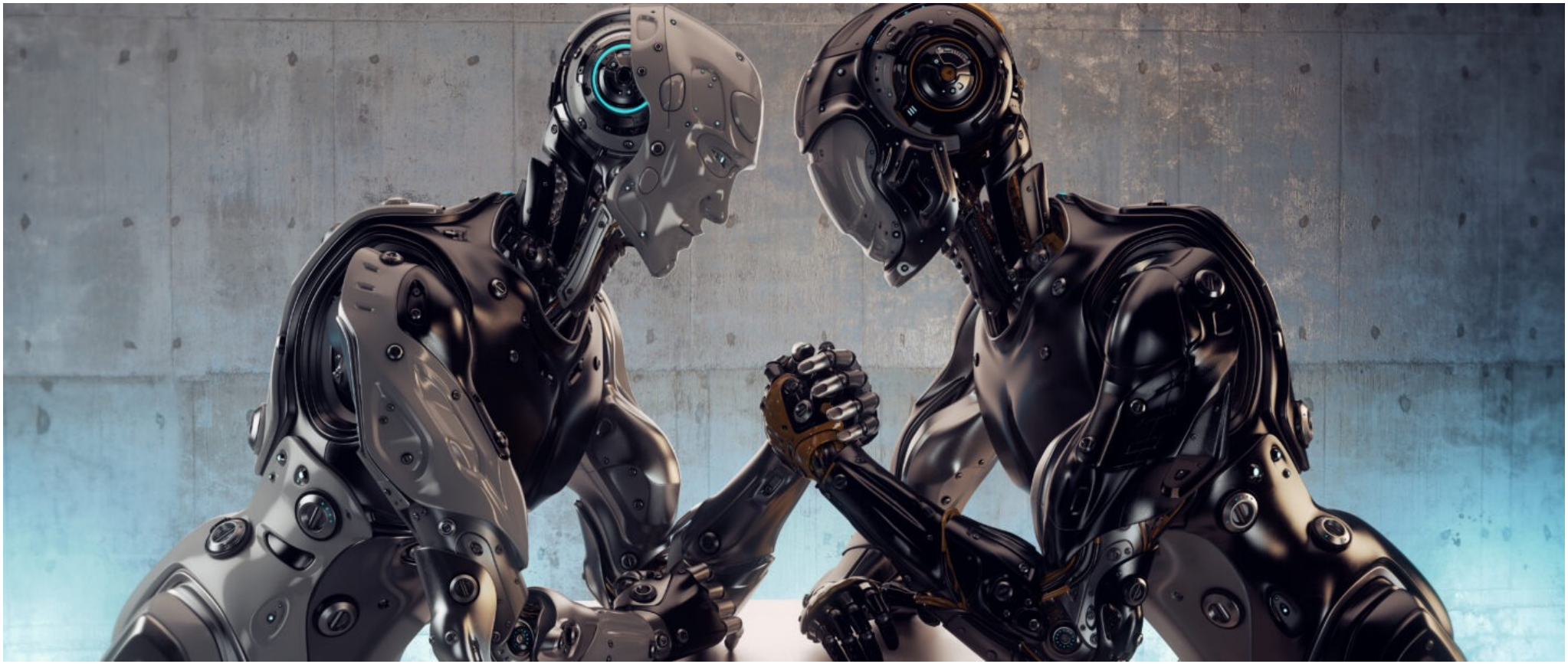
➡️ Can mislead the decisions in safety-critical AI systems.

♾️ Myriad possibilities of Multi-Patch attack variants.

➡️ Challenging to know beforehand to deploy a suitable defense.

🕵️ A new defense proposal always helps a malicious attacker.

➡️ Can upgrade the attack's design to overcome the defense.



There always exists arms race between attackers and defenders

Industrial Partners and Funding Agencies

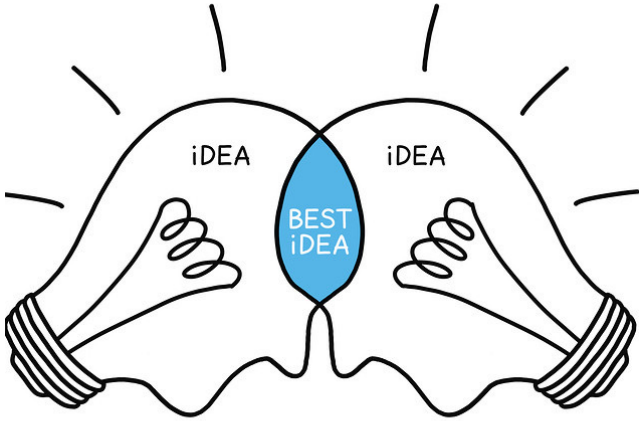
Students

Abhijith Sharma
MSc, UBC

Satyadwoom Kumar
UG, IIT, India

Javier Perez Tobia
UG, UBC

Vatsal Nanda
MITACS Globalink



Collaborate

Thank You !

Hiring MSc/PhDs

http://anarayan.com

Email: apurva.narayan@uwo.ca