SLAP: Improving Physical Adversarial Examples with Short-Lived Adversarial Perturbations

Giulio Lovisotto† (presenting)

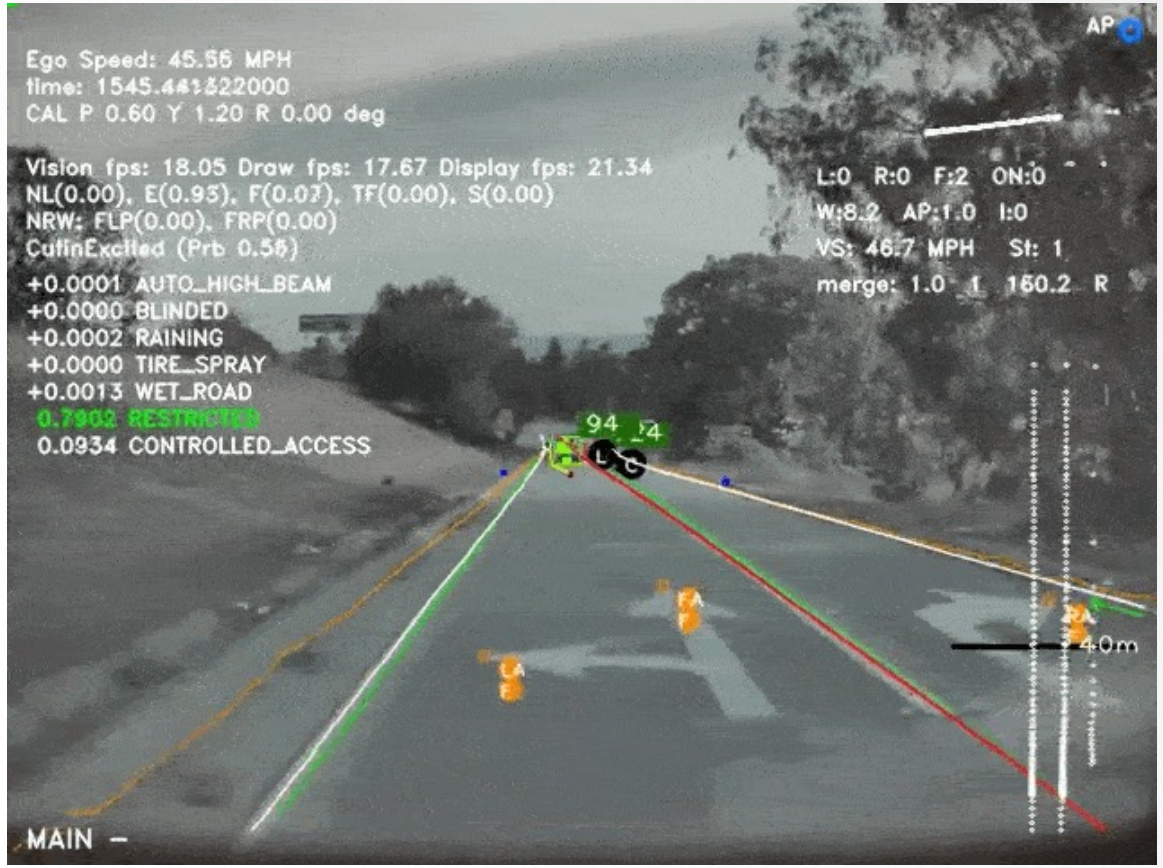
Henry Turner†, Ivo Sluganovic†, Martin Strohmeier*, Ivan Martinovic†

†*University of Oxford*, **armasuisse*

USENIX Security '21

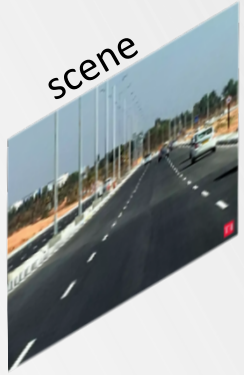
UNIVERSITY OF OXFORD

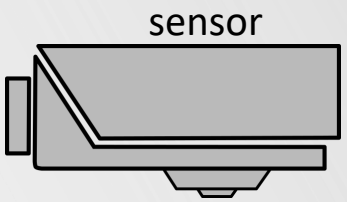
Motivation



Tesla AutoPilot (2019)

- Multiple deep neural networks
- Multi-tasks
- Multiple predictions
- Safety & Security?

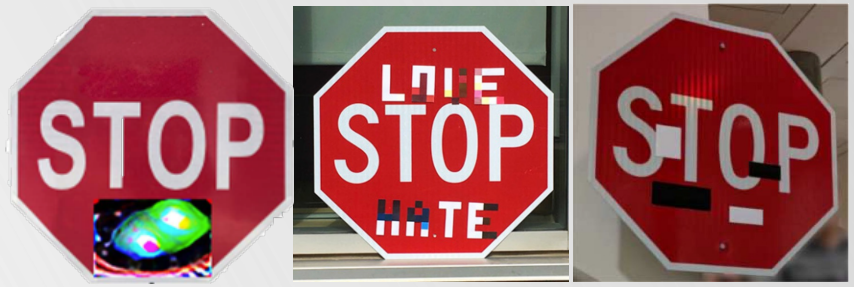




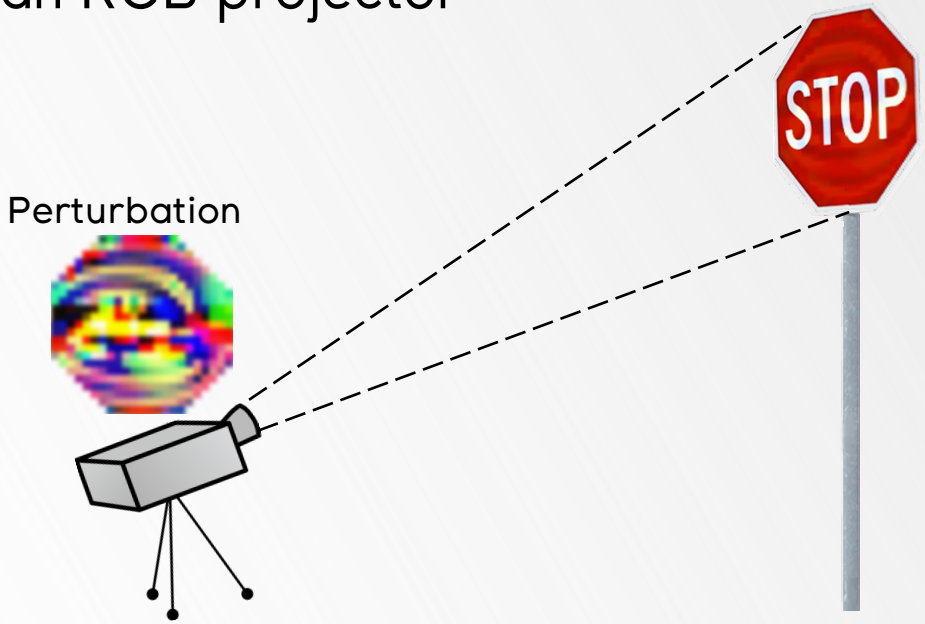
How can we alter a scene so that the camera-perceived inputs deceive a neural network?

Threat Model – Attack Vector

- One attack vector: Adversarial patches
  - Detectable behavior (see SentiNet[3])
  - Static scene modification (conspicuous)
- Any other way?
  - What if we used an RGB projector



Patch-based adversarial examples[1,2]



[1] - Eykholt, K. et al., 2018. Robust physical-world attacks on deep learning visual classification. CVPR.
[2] - Zhao, Y. et al., 2019. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. ACM CCS.
[3] - E. Chou, et al., 2020, SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems, IEEE SPW.

Method – Color Realizability

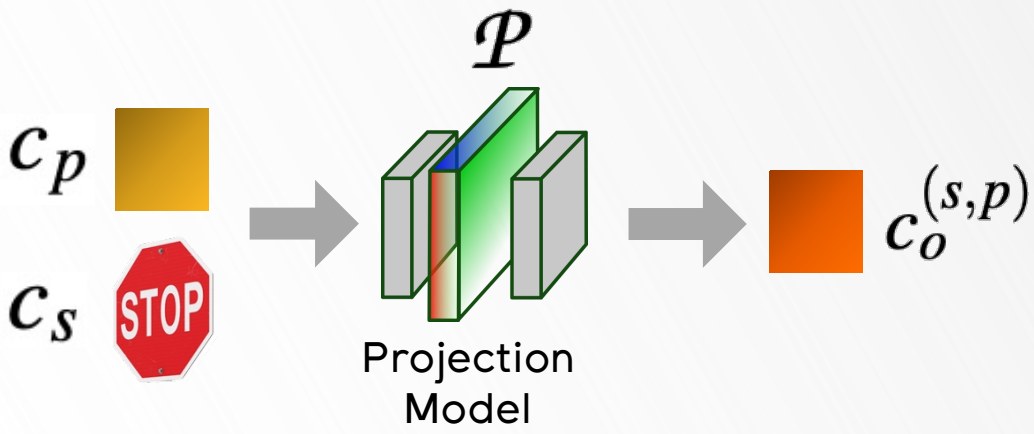
- Learn color space with a neural network
  - overcomes limit of Non-Printability Score[4] used for patches
- Collect (i) surface color, (ii) projected color, (iii) resulting color with camera

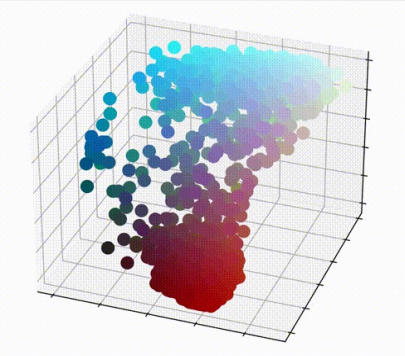
$$Loss_{\mathcal{P}} = \arg\min_{\theta_1} \sum_{\forall c_s, c_p} \left\| \mathcal{P}(\boxed{c_s}, \boxed{c_p}) - \boxed{c_o^{(s,p)}} \right\|_1$$

surface

projected image

resulting color

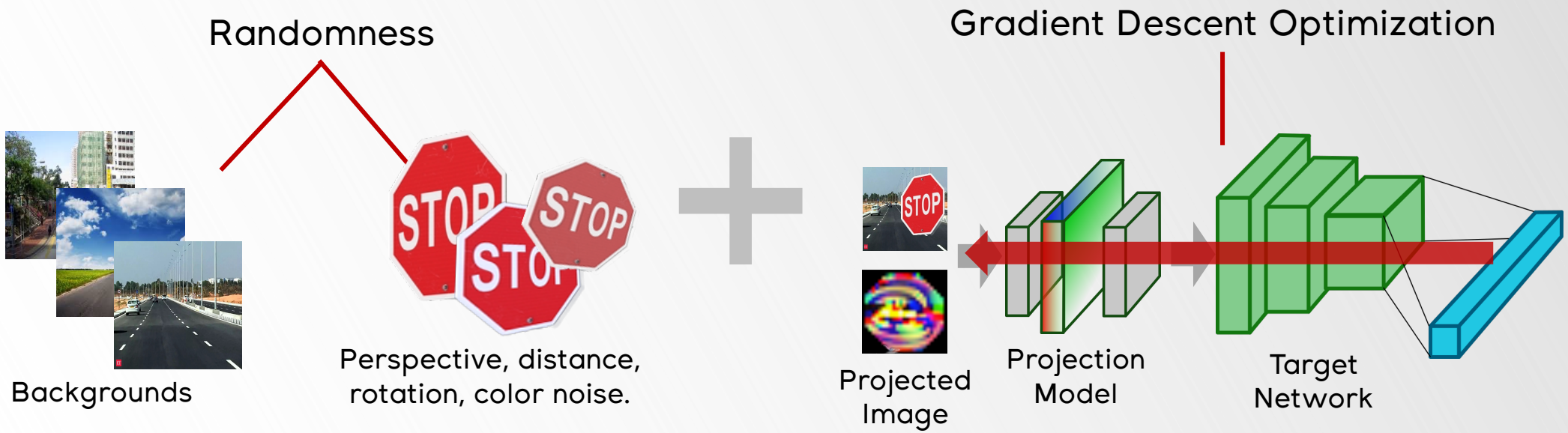




[4] - Sharif, M. et al., 2016. Accessorize to a crime: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. ACM CCS.

Method – AE Generation

- Real-world AE have to be robust to environmental changes
- Use *Expectation-over-transformation* with data augmentation



Real-world Evaluation

Evaluation on 4 target networks: LISACNN[1], GTSRBCNN[1], YOLOv3, MaskRCNN

- Controlled lighting indoor
- Outdoor in moving vehicle settings

YOLOv3

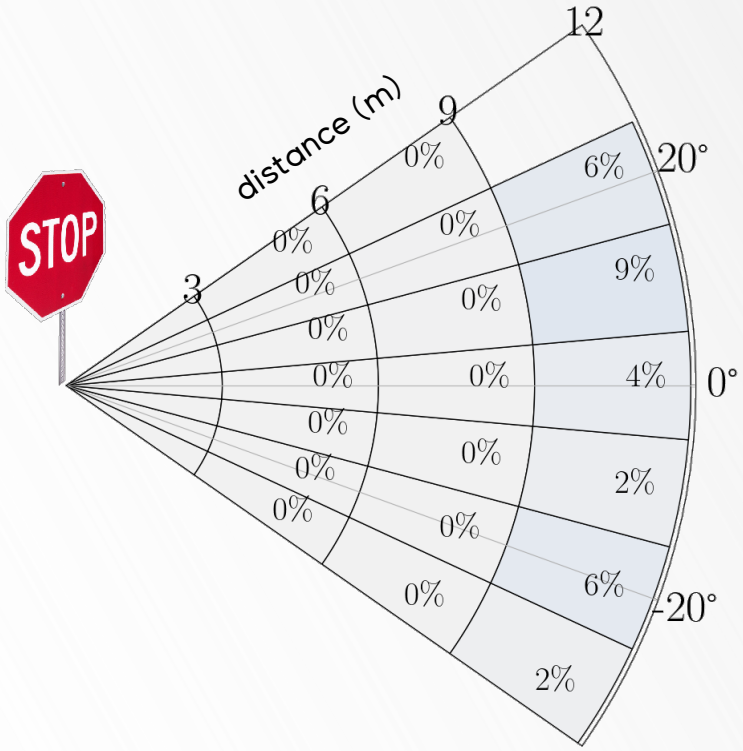
Outdoor - 5.30pm - ~200lx

Results – Attack Success

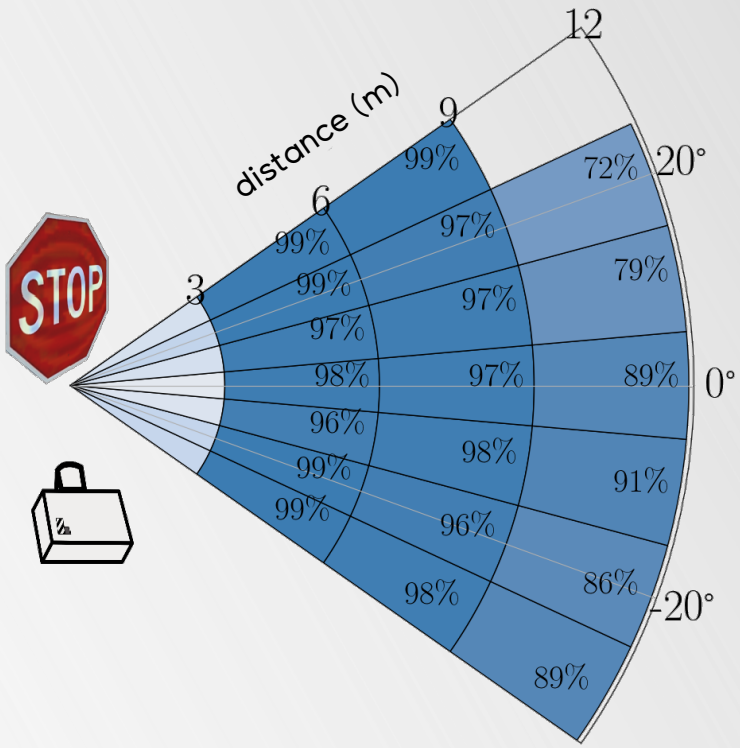
Mis-detection rate (no. of times a stop sign is <u>not</u> detected)

Baseline mis-detection

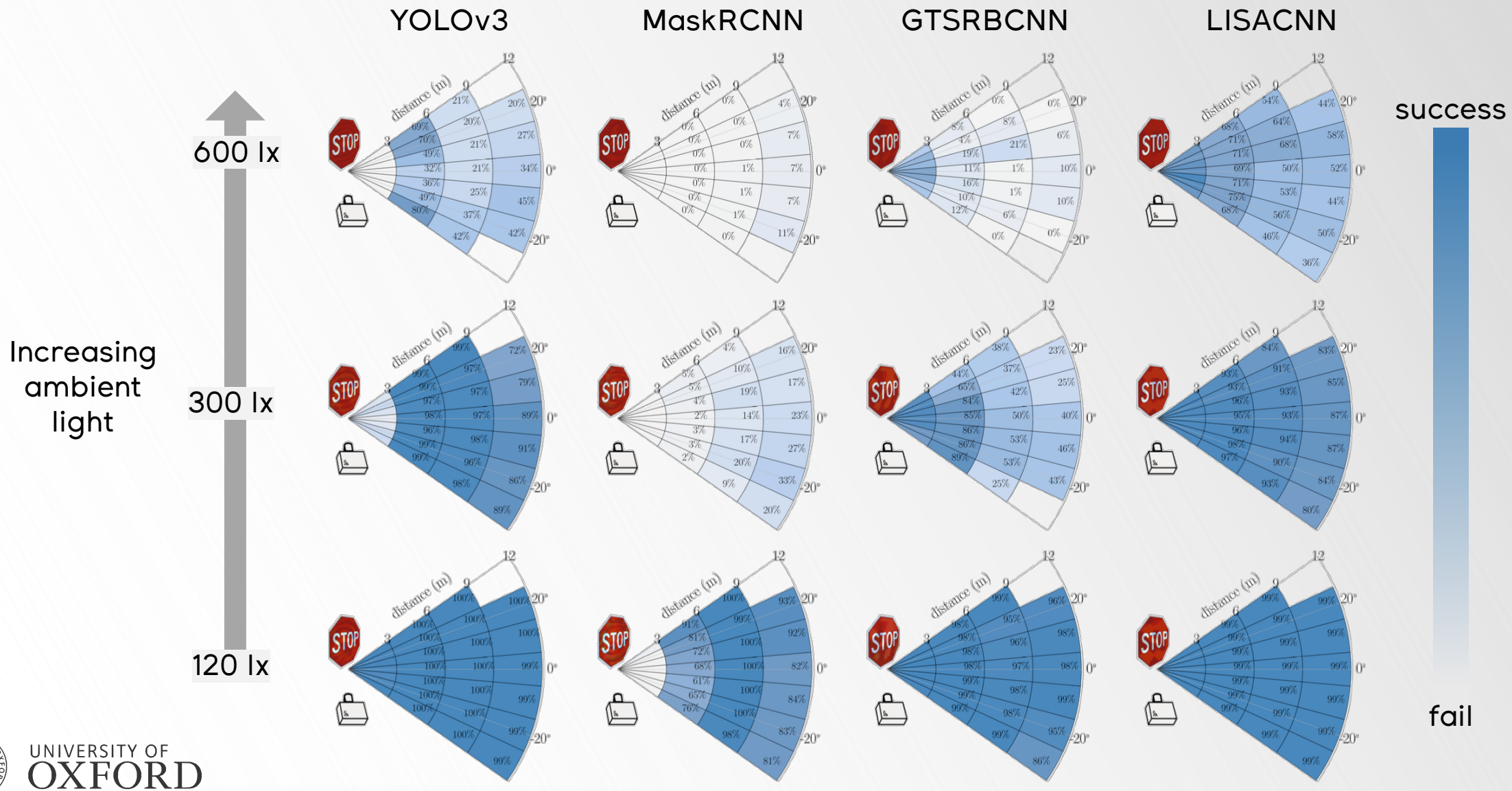
YOLOv3
(120lx)



SLAP



Results – Attack Success

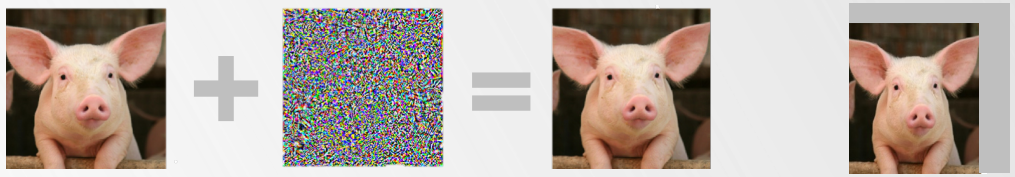


Results - Defenses

- Adversarial learning[5]

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y) + (1-\alpha) J(\boldsymbol{\theta}, \boldsymbol{x} + \epsilon \operatorname{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)))$$

- Input randomization[6]



Resizing + Padding

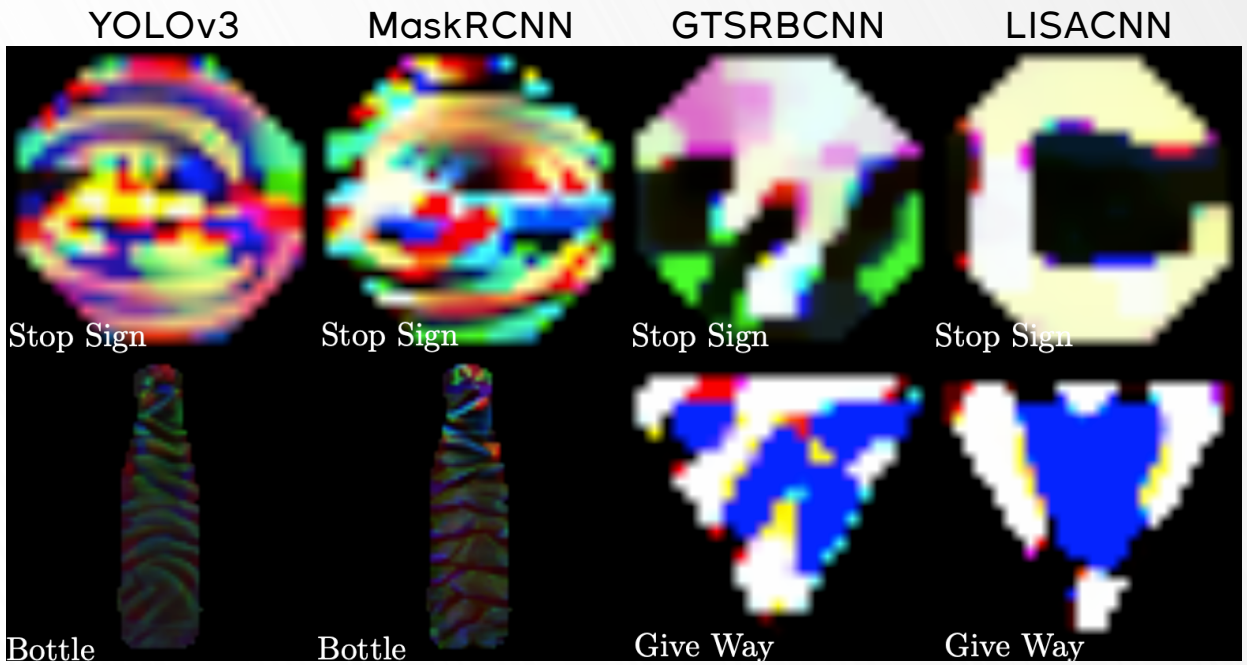
- SentiNet[3] (specifically designed for physical AE)

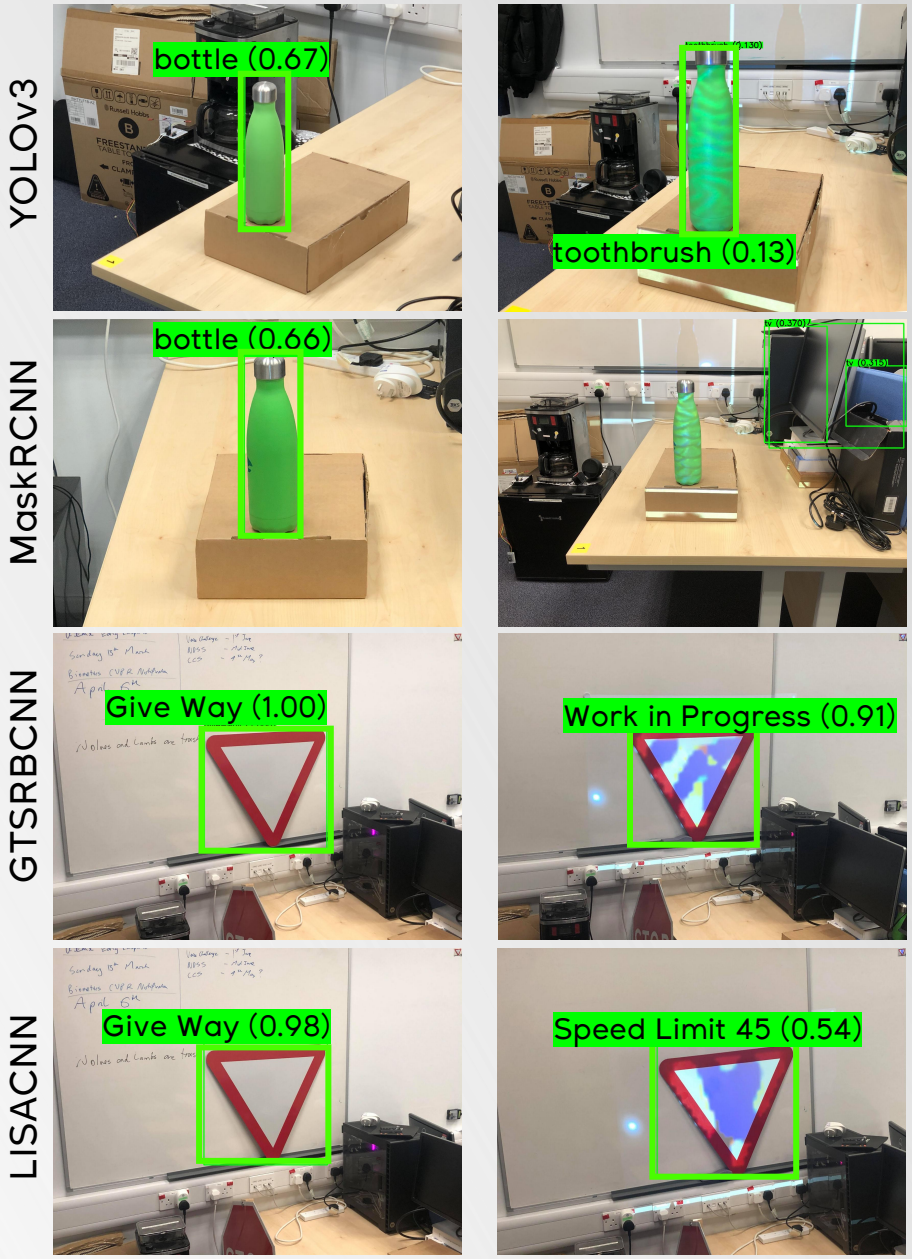
Network	Ambient Light (lx)	Attack Success	Adversarial Learning	Input Randomization	SentiNet Random	SentiNet Checkerboard
Gtsrb-CNN	120	99.96%	20.23% (-79.73%)	99.55% (-0.40%)	93.43%	95.45%
	180	90.53%	23.57% (-66.97%)	90.02% (-0.51%)	93.19%	93.72%
	300	56.51%	48.18% (-8.33%)	86.78% (+30.27%)	96.97%	96.46%
	440	56.34%	40.24% (-16.10%)	82.96% (+26.61%)	95.81%	96.34%
	600	12.79%	10.91% (-1.88%)	51.37% (+38.58%)	95.29%	95.29%
Lisa-CNN	120	100.00%	0.06% (-99.94%)	100.00% (+0.00%)	94.24%	95.29%
	180	99.95%	0.88% (-99.07%)	99.90% (-0.05%)	100.00%	100.00%
	300	99.81%	0.00% (-99.81%)	99.98% (+0.17%)	94.76%	96.86%
	440	98.44%	0.59% (-97.85%)	99.95% (+1.51%)	100.00%	100.00%
	600	69.05%	0.04% (-69.01%)	95.71% (+26.67%)	95.81%	96.86%

[5] - Goodfellow, I.J et al., 2015. Explaining and harnessing adversarial examples. CVPR.
[6] - Xie, C. et al., 2018. Mitigating adversarial effects through randomization. ICLR.
[3] - E. Chou, et al., 2020, SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems, IEEE SPW.

Results - Generalizability

Extension to other models and objects:



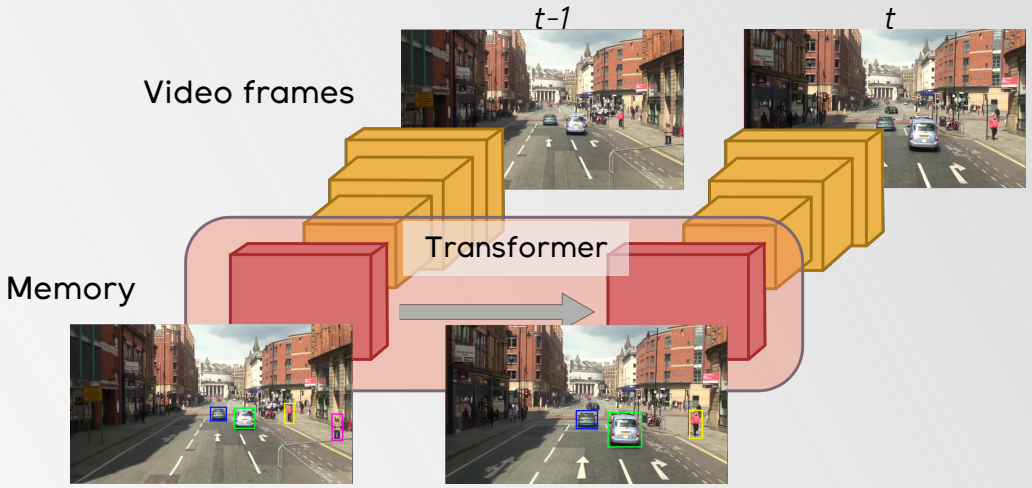


Conclusion

- Demonstrated that SLAP works robustly in real-world for both image classification (traffic signs) and object detection

- No straightforward defense detects SLAP
  - Adversarial learning might end up in arms race

- Introduce new capabilities for AE-based attackers: alter scene content dynamically
  - Highlights the practical need of new type of adversarial robustness for sequential data

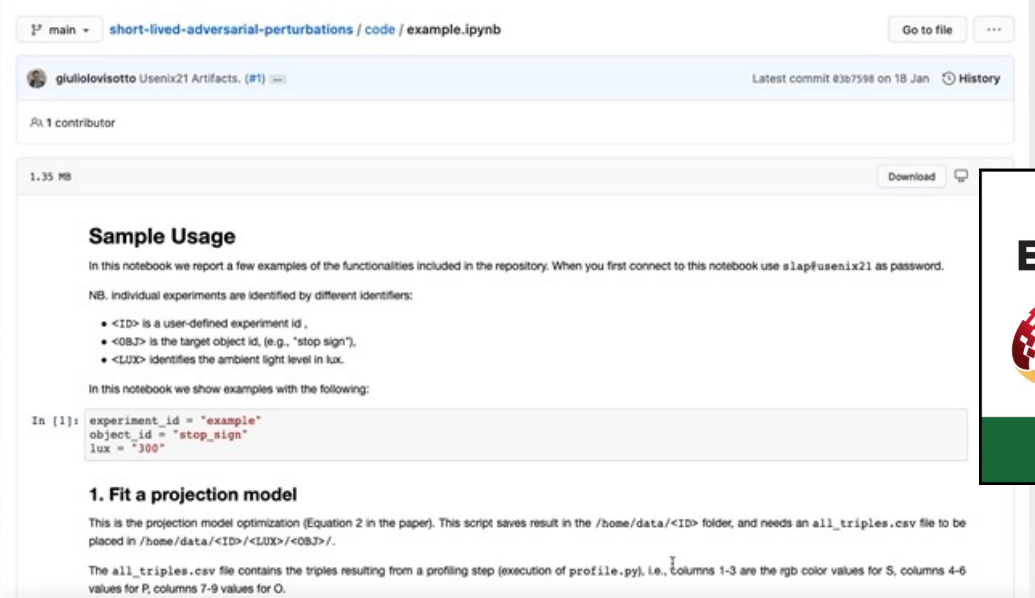


Thanks

Thanks for your attention!

Contact: giulio.lovisotto@cs.ox.ac.uk

SLAP project page https://github.com/ssloxford/short-lived-adversarial-perturbations/



ARTIFACT EVALUATED
usenix ASSOCIATION
PASSED

UNIVERSITY OF OXFORD