ISC 2019 第七届互联网安全大会

人工智能攻防 - 物理对抗自动驾驶目标识别系统

陈恺

中科院信息工程研究所研究员

小鹅助理

扫码添加小鹅助理，与数万科技圈人士

分享重量级活动PPT、干货培训课程、高端会议免费

门票


ISC
陈恺
中科院信息工程研究所 研究员


ISC
人工智能攻防——
物理对抗自动驾驶目标识别系统
陈恺
中科院信息工程研究所 研究员
Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, Kai Chen, "Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors", CCS 2019

汽车安全

Traditional

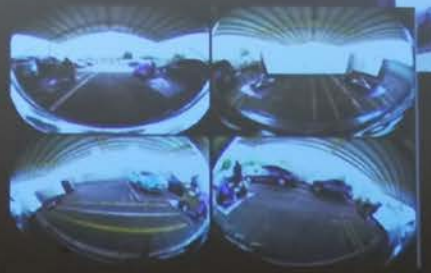
U盘

信号干扰器

Adversarial Attack

CommanderSong

(USENIX Security 2018)



智能视觉感知系统攻击

自动驾驶系统安全

Uber Self-Driving Car That Struck, Killed Pedestrian Wasn't Set to Stop in an Emergency

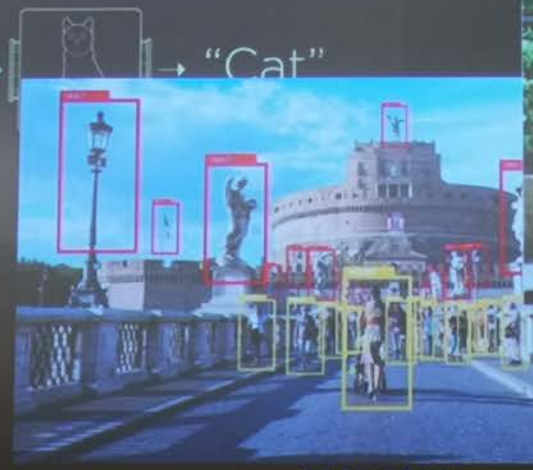
Pedestrian tested positive for methamphetamine and marijuana

Utah driver sues Tesla after crashing in Autopilot mode

Tesla汽车在自动驾驶模式下，将白色大货车误认为是天空，发生交通事故。

目标检测系统

图像识别



目标检测

目标检测（语义分割）

目标检测对抗攻击

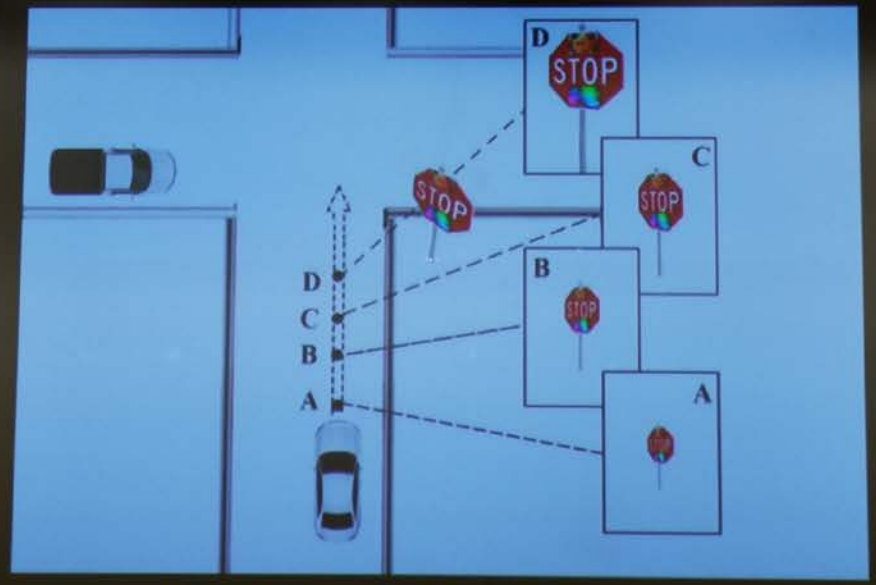
Digital Space

Speed limit 50

Physical World

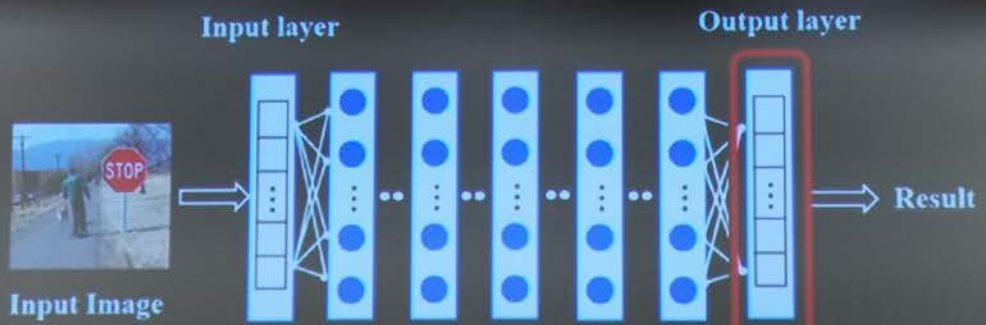
- 距离较短（<=9m）
- 角度受限（<=15°）
- 作用环境受限
- 迁移性差

物理对抗攻击



动态变化：角度、距离、光线

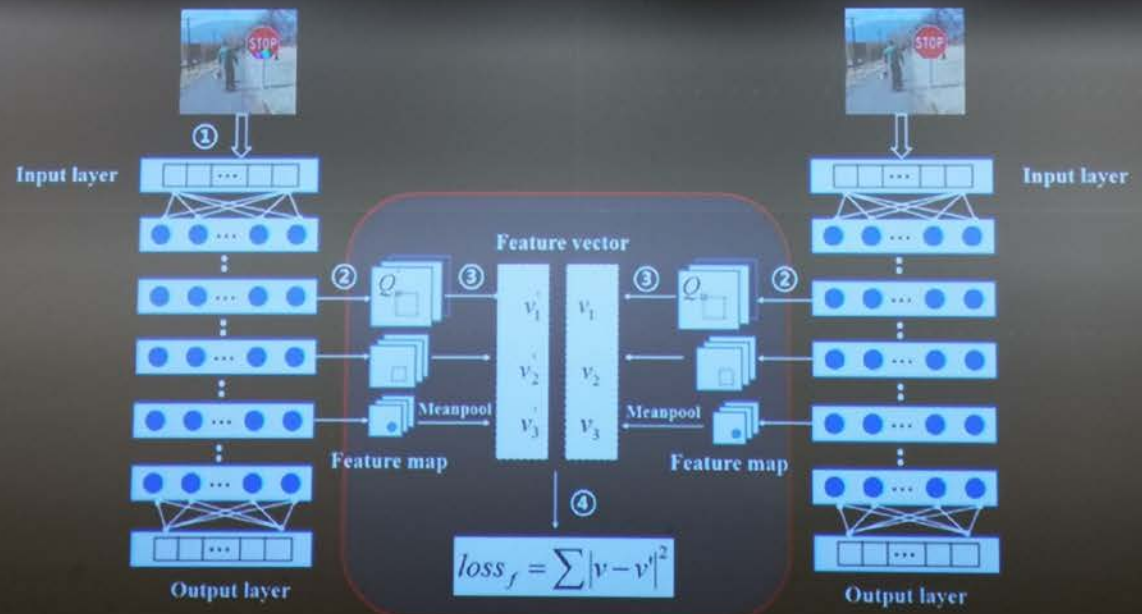
Feature-interference——距离、角度与环境鲁棒性



传统方法

$$\alpha \cdot C_N^{bbox} + \beta \cdot p_N(y^N \mid S)$$

Feature-interference——距离、角度与环境鲁棒性

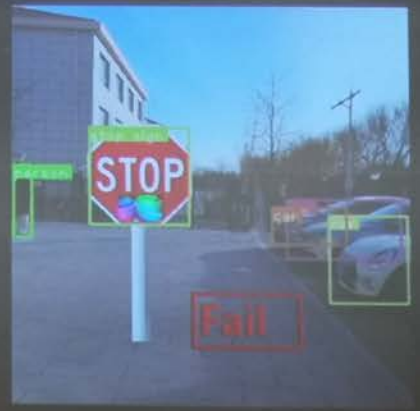


Enhanced Realistic Constraints

----环境鲁棒性

室内

室外（无杆）

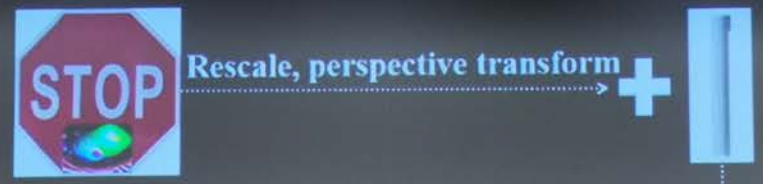


室外（有杆）


Enhanced Realistic Constraints
----环境鲁棒性
'Stop Sign'
Search
Google Image
Stop Sign
(X_center, Y_center)
(width, height)
Angle = arccos(width / height)
Mask Rescale
语义相同背景

Enhanced Realistic Constraints

----环境鲁棒性



Rescale, perspective transform

'Road'

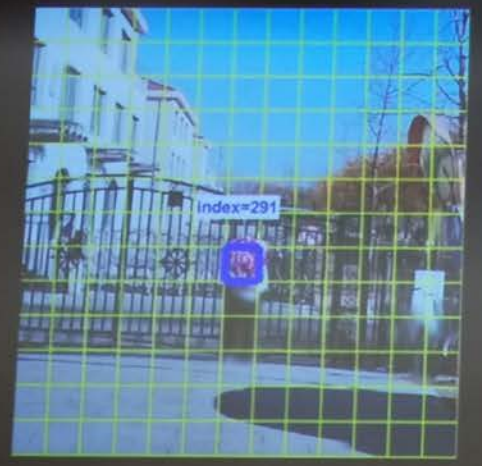
Search

Google Image

Position Region

语义相关背景

Nested AE----距离鲁棒性



Darknet-53

Object Detector

Small → 小目标---远距离

Middle → 中目标---中距离

Large → 大目标---近距离

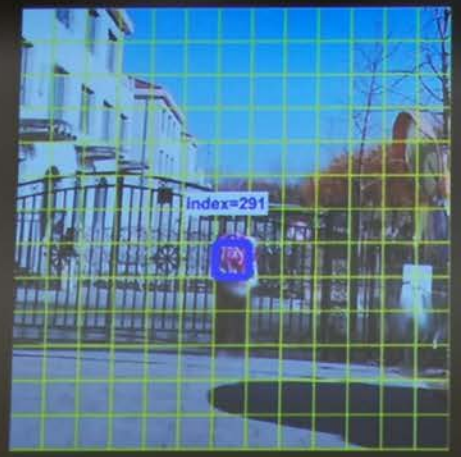
Nested AE----距离鲁棒性

Darknet-53

Object Detector

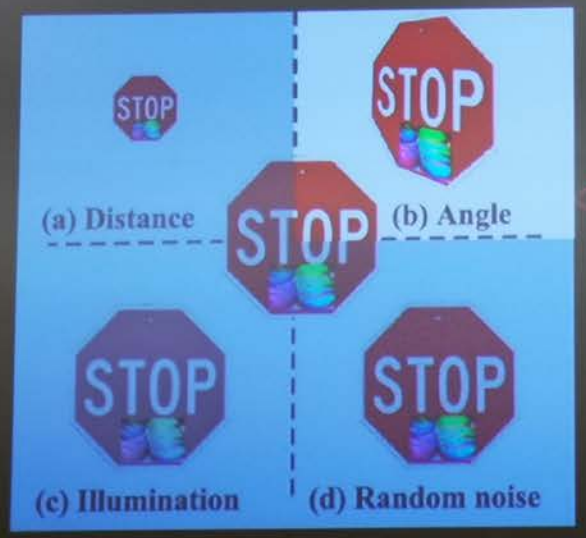
- Small → 小目标---远距离
- Middle → 中目标---中距离
- Large → 大目标---近距离

Nested AE----距离鲁棒性



$$X_{i+1}^{adv} = Clip\begin{cases} X_i + \varepsilon sign(J(X_i)), & S_p \leqslant S_{thres} \\ X_i + \varepsilon M_{center} sign(J(X_i)), & S_p > S_{thres} \end{cases}$$

EOT----角度鲁棒性



角度：Perspective Transformation

Style-customized AEs

----隐匿性

- Pattern

$$loss_{pattern} = \sum |p_j - 1|^2$$

- Color

$$loss_{color} = \sum_{pixel \in X_t} \frac{pixel_R + pixel_G + pixel_B}{pixel_T}$$

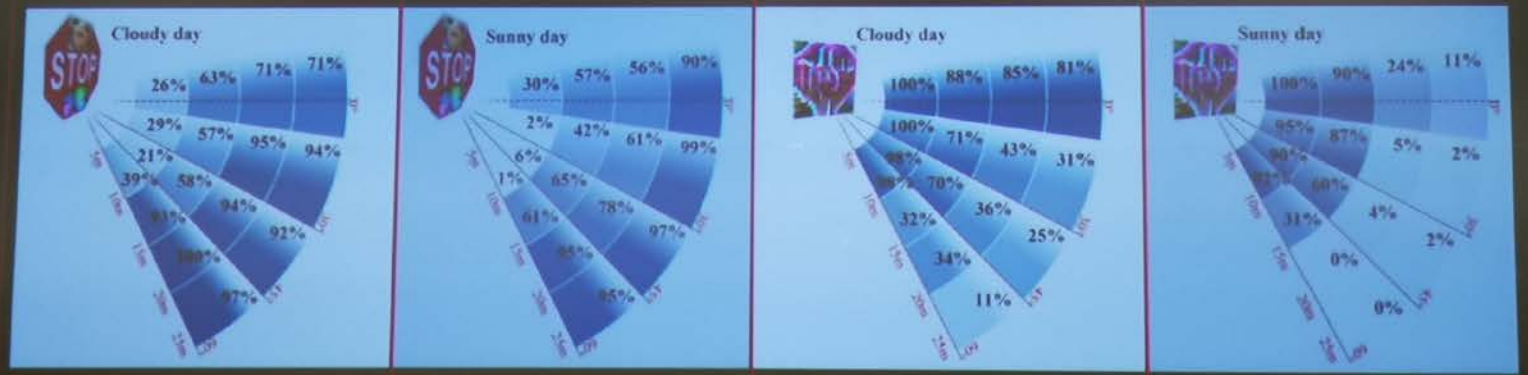
- Shape

$$X^{adv} + P \cdot Mask$$

- Text

Shape+Color

Evaluation----距离、角度与光线



HA AA

Evaluation----性能提升

		Distance	Angle	Perturbation Area	Transferability
HA	Our Method	≤ 25m	≤ 60°	20%~25%	Faster，YOLO，SSD，RFCN，Mask
	ShapeShifter	≤ 40' (12m)	≤ 15°	Full stop except 'STOP'	Uable
	Eykholt's method	≤ 30' (9m)	---	20%~25%	Faster RCNN (18%)
AA	Our Method	≤ 25m	≤ 60°	---	Faster，YOLO，SSD，RFCN，Mask
	ShapeShifter	---	---	---	---
	Eykholt's method	≤ 10' (3m)	---	---	---

作用距离短 ✓ 角度受限 ✓ 环境受限 ✓ 隐匿性 ✓ 迁移性差 ✓

Evaluation----Enhanced Realistic Constraints

Success Rate	Physical		Digital	
	ERG	Without ERG	ERG	Without ERG
YOLO V3	54%	31%	73%	53%
Faster RCNN	67%	43%	74%	63%

满足相关语义环境生成的AEs鲁棒性能够显著提高。

Evaluation----Style-customized AEs

Style-customized AEs		HA				AA
		Pattern	Shape	Text	Color	
YOLO V3	Indoors	92.4%(784)	99.5%(803)	99.4%(716)	97%(986)	88%(1894)
	Outdoors	53%(963)	76.5%(613)	65%(604)	30%(738)	91.7%(1788)
Faster RCNN	Indoors	83.6%(885)	89%(784)	86.2%(802)	28%(696)	---
	Outdoors	78.1%(717)	87.8%(614)	79.6%(766)	31.2%(715)	---

HA和AA不同风格的AEs在多种模型、多种环境（室内、室外)下均可以获得很好的性能（除了HA的Red color hue）。

Evaluation----Transferability of AEs

White-box Model		Black-box Model			
		Faster RCNN /YOLO V3	SSD	RFCN	Mask RCNN
YOLO V3 (HA)	Indoors	21%	71.6%	52.6%	49.7%
	Outdoors	10%	46%	19.2%	9%
Faster RCNN (HA)	Indoors	98.7%	90.7%	91%	85.7%
	Outdoors	76.8%	78%	72%	58%

不同环境，多种模型（One-stage：YOLO V3，SSD；Two-stage：Faster R-CNN，Mask R-CNN,RFCN）之间具有较高迁移性。

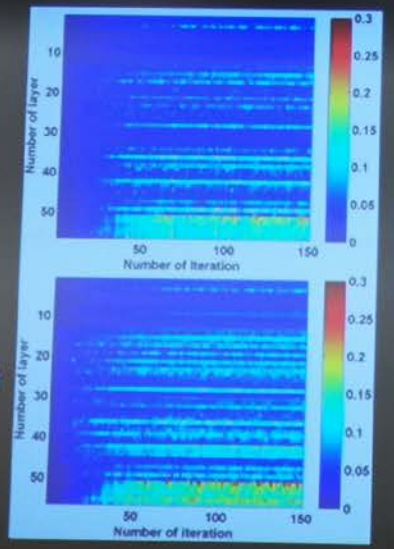
Evaluation----Real-road driving tests

Success Rate	Straight road	Crossroad
HA (6km/h)	75%	64%
AA (6km/h)	63%	81%
HA (30km/h)	72%	60%
AA (30km/h)	76%	78%

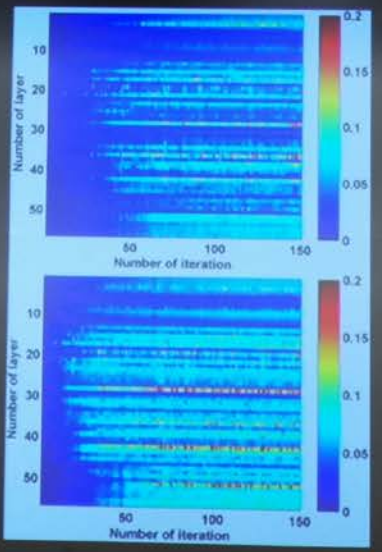
不同车速下，HA和AA在多种路线（直路，弯路）下的性能。

Understanding

(a) with no FIR/ERG



(b) with FIR



(c) with ERG

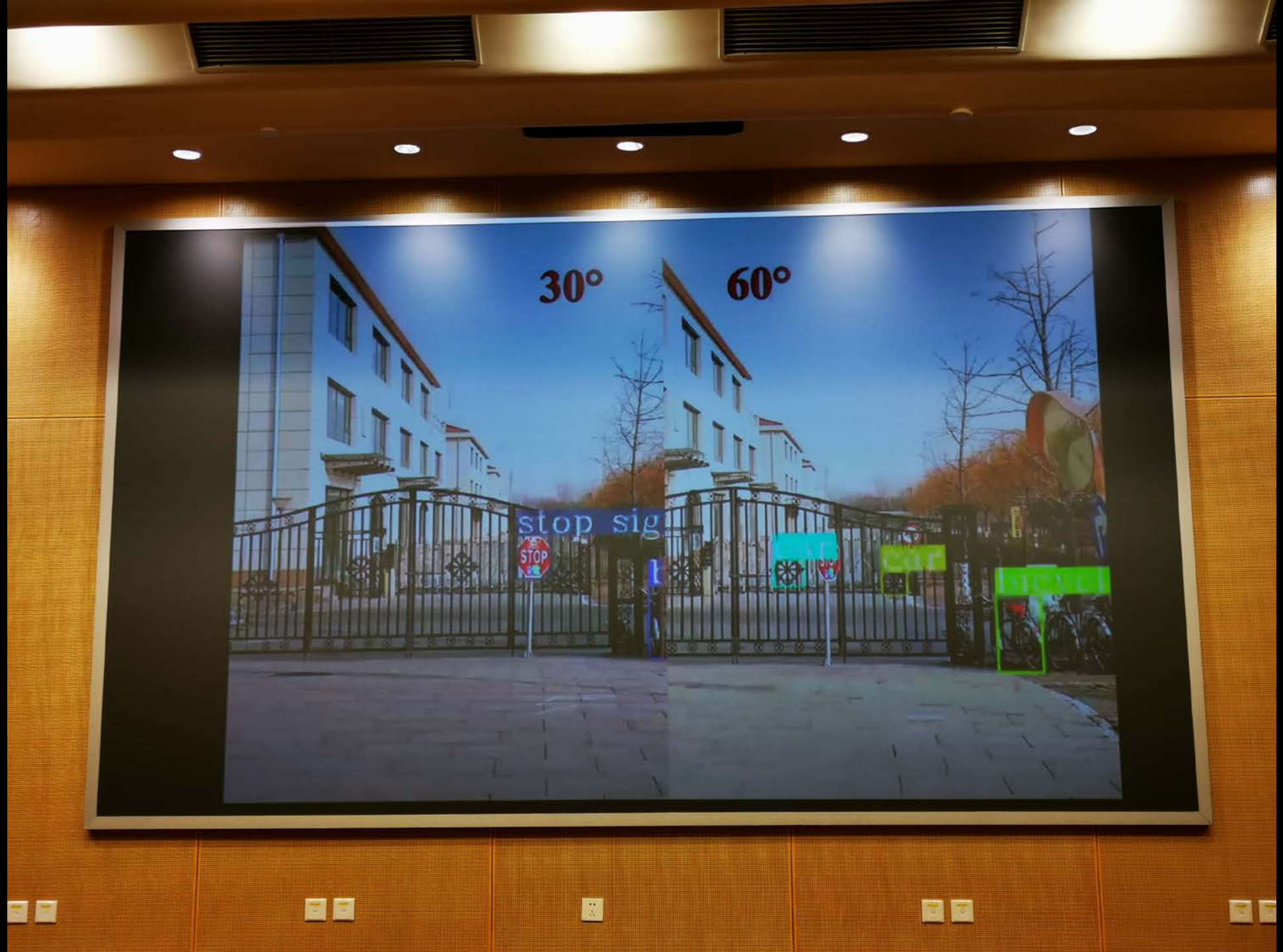
(d) with FIR/ERG

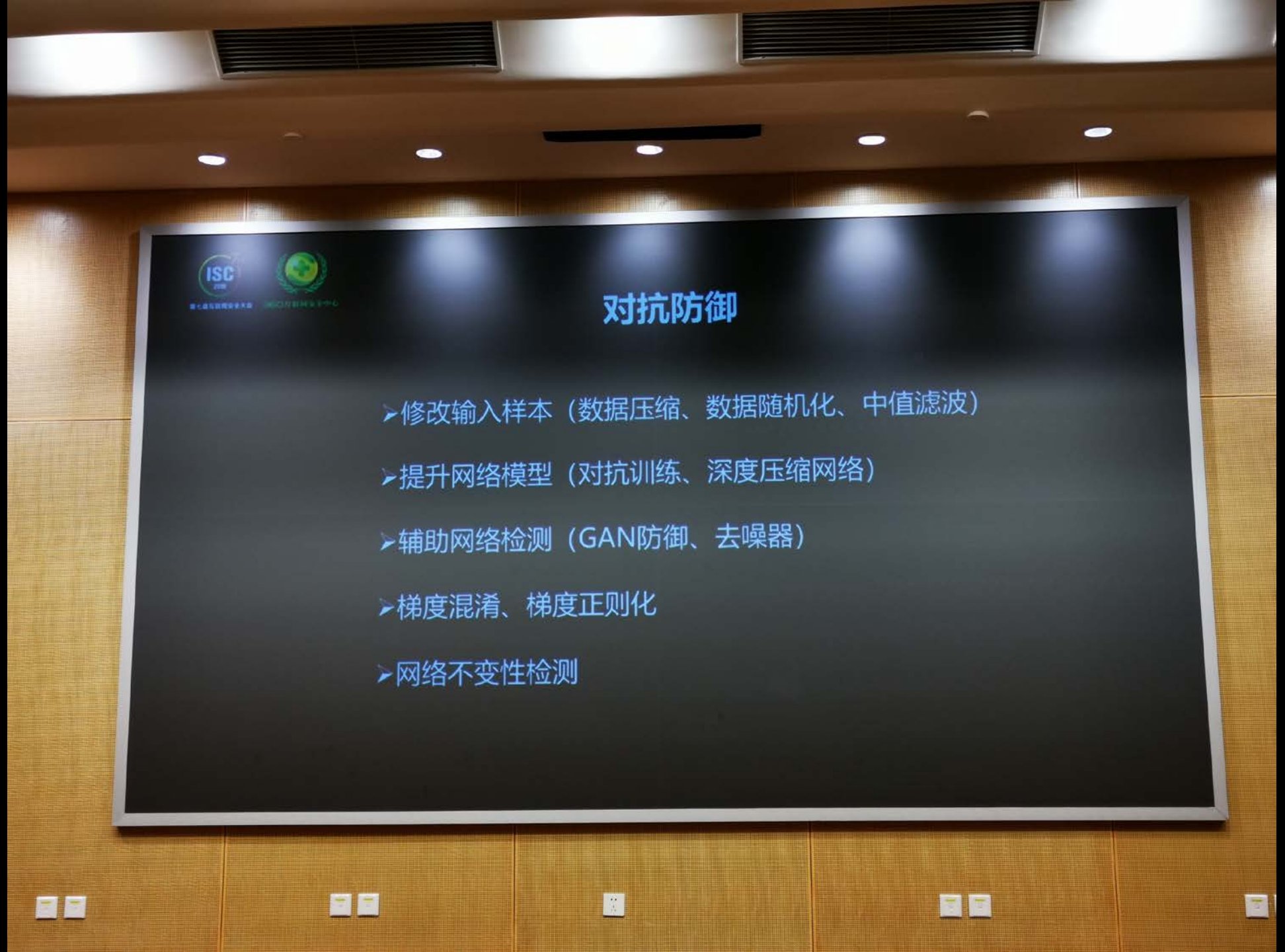
FIR与ERG均能增加隐藏层特征差异（Original VS AE），提高AEs鲁棒性。

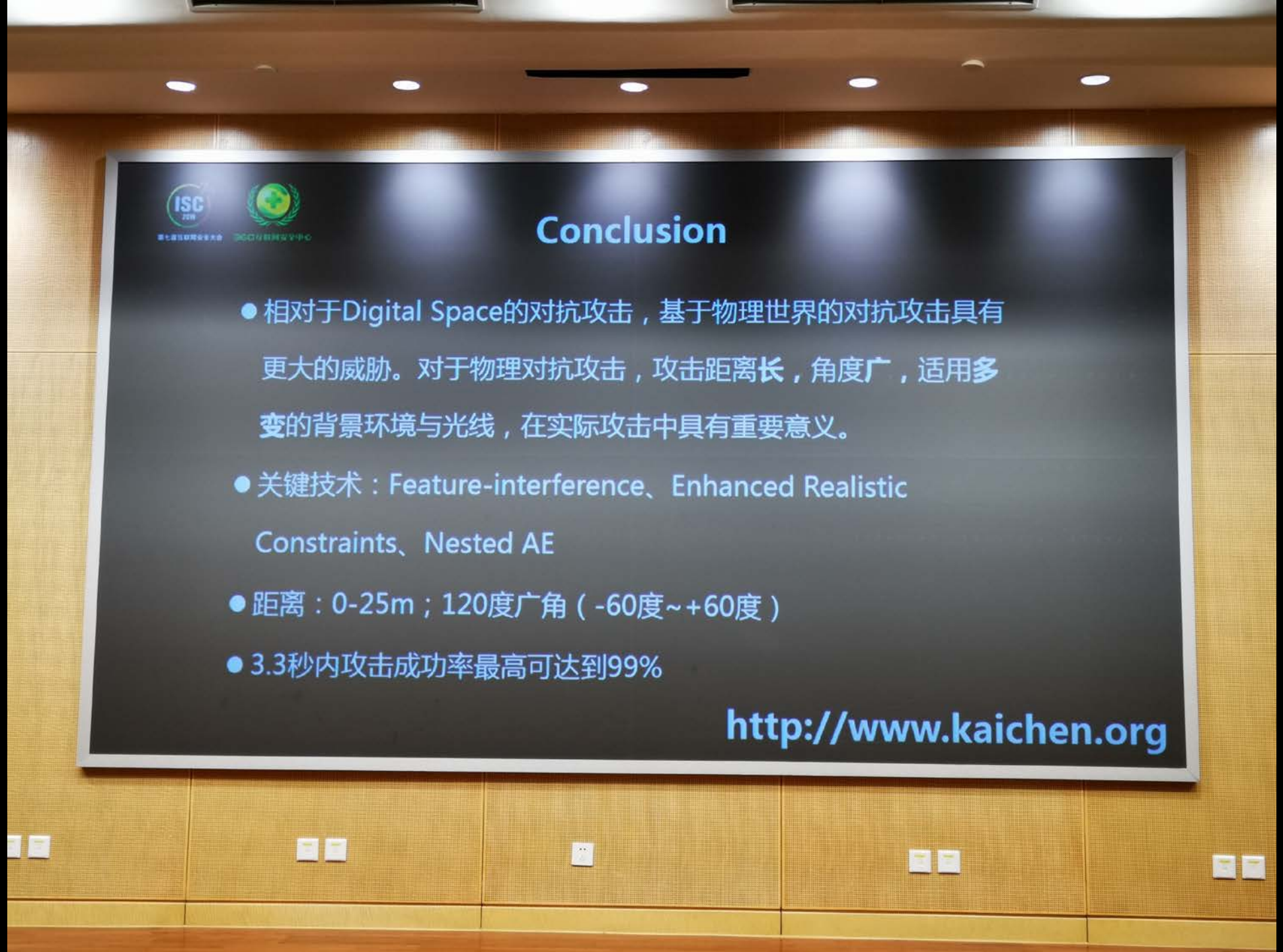

Original stop
Patched stop
stop sign
car
ding attack


Original stop
Patched stop
ccar
car


30°
60°
car


30°
60°
stop sig
STOP
car


ISC
对抗防御
修改输入样本（数据压缩、数据随机化、中值滤波）
提升网络模型（对抗训练、深度压缩网络）
辅助网络检测（GAN防御、去噪器）
梯度混淆、梯度正则化
网络不变性检测


Conclusion
● 相对于Digital Space的对抗攻击，基于物理世界的对抗攻击具有更大的威胁。对于物理对抗攻击，攻击距离长，角度广，适用多变的背景环境与光线，在实际攻击中具有重要意义。
● 关键技术：Feature-interference、Enhanced Realistic Constraints、Nested AE
● 距离：0-25m；120度广角（-60度~+60度）
● 3.3秒内攻击成功率最高可达到99%
http://www.kaichen.org

小鹅助理

谢谢！

扫码添加小鹅助理，与数万科技圈人士

分享重量级活动PPT、干货培训课程、高端会议免费门票