Revisiting Physical-World Adversarial Attack on Traffic Sign Recognition: A Commercial Systems Perspective

Ningfei Wang, Shaoyuan Xie, Takami Sato, Yunpeng Luo,

Kaidi Xu*, Qi Alfred Chen

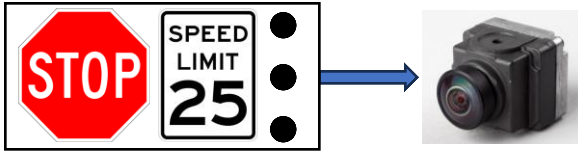
University of California, Irvine and *Drexel University

Traffic Sign Recognition (TSR) Systems

- Traffic Sign Recognition (TSR) system employs camera sensors with Deep Neural Networks (DNNs) to detect road signs

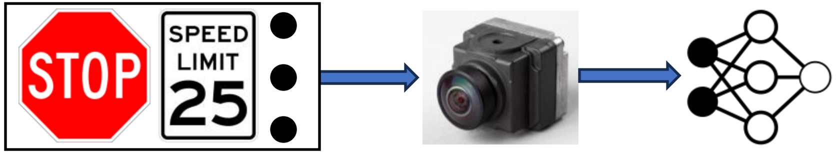
Traffic Sign Recognition (TSR) Systems

- Traffic Sign Recognition (TSR) system employs camera sensors with Deep Neural Networks (DNNs) to detect road signs



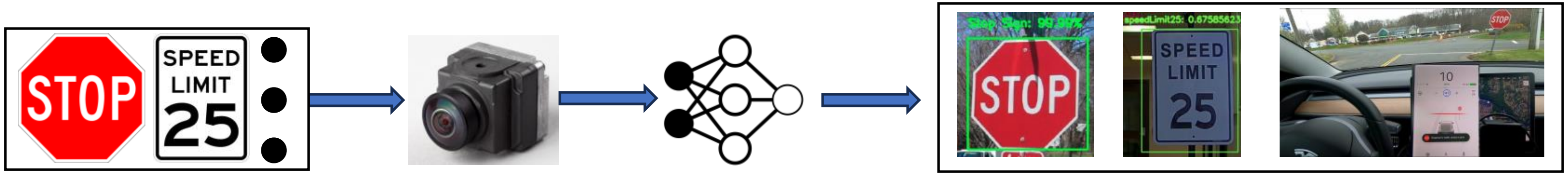
Traffic Sign Recognition (TSR) Systems

- Traffic Sign Recognition (TSR) system employs camera sensors with Deep Neural Networks (DNNs) to detect road signs



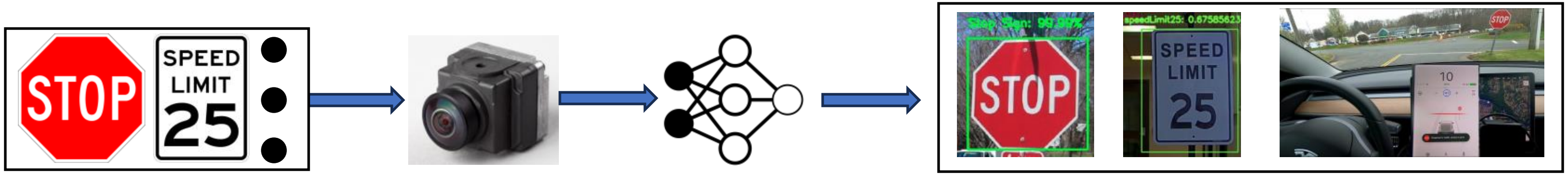
Traffic Sign Recognition (TSR) Systems

- Traffic Sign Recognition (TSR) system employs camera sensors with Deep Neural Networks (DNNs) to detect road signs



Traffic Sign Recognition (TSR) Systems

- Traffic Sign Recognition (TSR) system employs camera sensors with Deep Neural Networks (DNNs) to detect road signs



- Such TSR systems generally exist in top leading car brands in the United States [1]

[1] Leading car brands in the United States in 2023, based on vehicle sales: https://www.statista.com/statistics/264362/leading-car-brands-in-the-us-based-on-vehicle-sales/

Failure of TSR Can Lead to Accidents

Millions of people drive, ride, or walk through stop sign intersections daily.

However, nearly 70,000 accidents occur yearly due to people running stop signs; a third result in injuries.

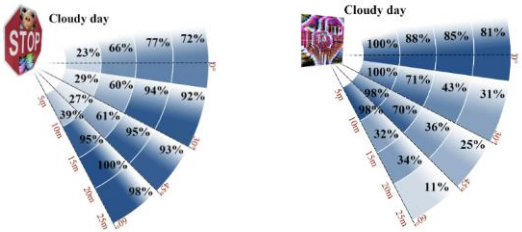
There are many scenarios in which a person may find themselves in a stop sign car accident. For instance, a driver may be hit by someone running a stop sign, or the driver may hit the person running the stop sign. More than two cars may be involved in an intersection with a 3- or 4-way stop. Proving who is at fault can be challenging in stop sign violations that result in an accident. Consulting with a St. Louis car accident lawyer can help you determine liability and pursue fair compensation.

Prior Commercial TSR Security Research

Millions of people drive, ride, or walk through stop sign intersections daily.

However, nearly 70,000 accidents occur yearly due to people running stop signs; a third result in injuries.

There are many scenarios in which a person may find themselves in a stop sign car accident. For instance, a driver may be hit by someone running a stop sign, or the driver may hit the person running the stop sign. More than two cars may be involved in an intersection with a 3- or 4-way stop. Proving who is at fault can be challenging in stop sign violations that result in an accident. Consulting with a St. Louis car accident lawyer can help you determine liability and pursue fair compensation.



Zhao et al.: ACM CCS 2019

You can confuse self-driving cars by altering street signs

It doesn't take much to send autonomous cars crashing into each other.

Jon Fingas
Reporter

Jon Fingas: engadget

● ● ●

Importance of Commercial TSR Security

Limitations:

- Almost all only evaluate attack effects on ***academic TSR models***, leaving the impacts on ***commercial TSR systems*** largely unclear.

Importance of Commercial TSR Security

Limitations:

- Almost all only evaluate attack effects on ***academic TSR models***, leaving the impacts on ***commercial TSR systems*** largely unclear.
- A few recent works tried to understand ***commercial TSR system-level impacts***, but limited to ***one particular vehicle model***, sometimes even an ***unknown one***, making both the ***generalizability*** and ***representativeness*** questionable

Research Question

Research Question:

Can any of the existing physical-world TSR adversarial attacks achieve a general impact on commercial TSR systems today?

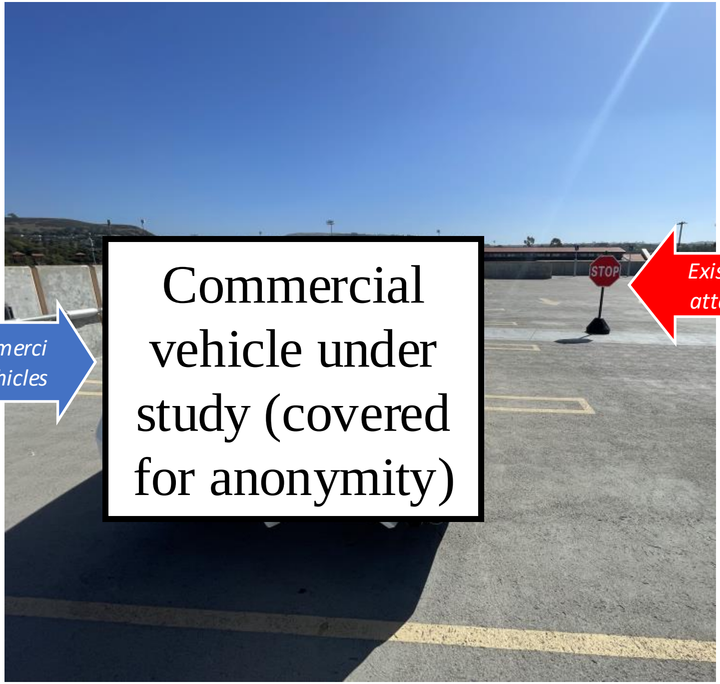
Our Contributions

- The **first large-scale** measurement of **physical-world** adversarial attacks against **commercial TSR systems**
- Discovery and analysis of a **spatial memorization design** that commonly exists in today's commercial TSRs
- Propose new **attack success metric designs** and use this metric to **revisit** the **evaluations**, **designs**, and **capabilities** of existing attacks in this problem space

Measurement Study Setup Overview

4 of these 5 vehicle models are studied (with 1 confusing model for anonymity):

Commercial vehicles



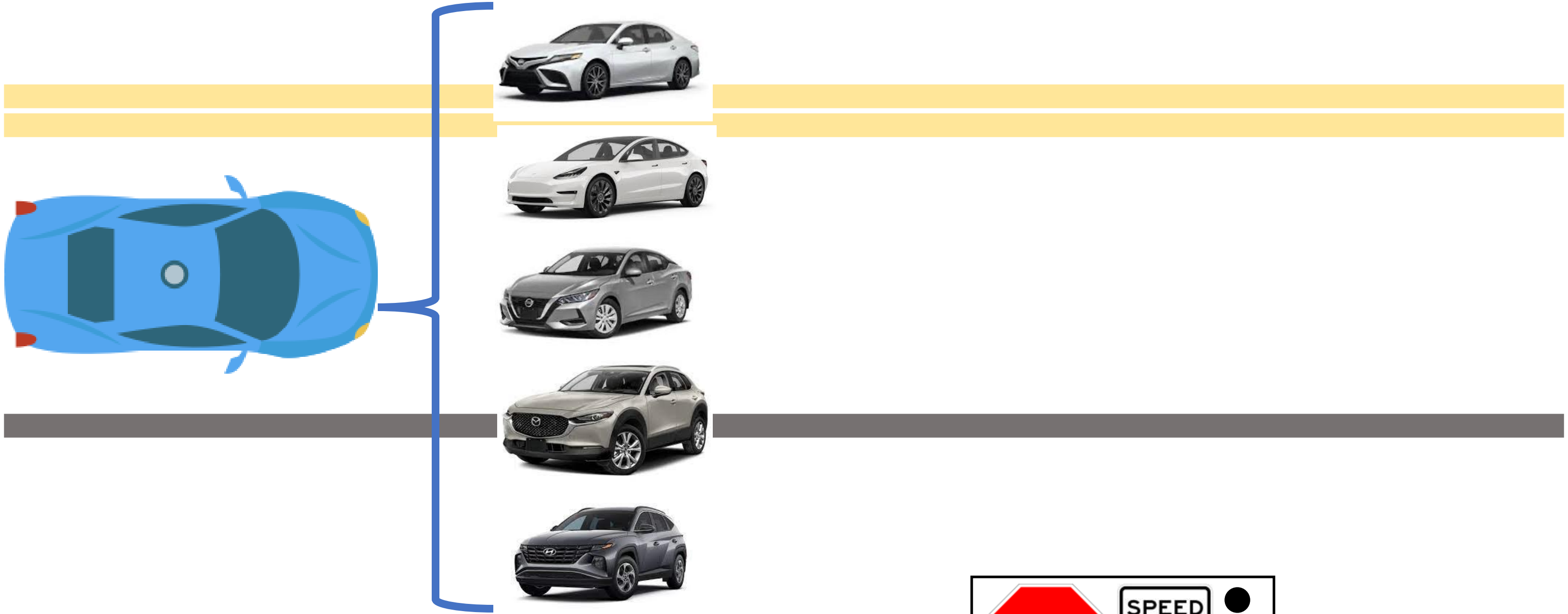
Commercial vehicle under study (covered for anonymity)

Existing attacks

Test Environment Setups

Test Environment Setups

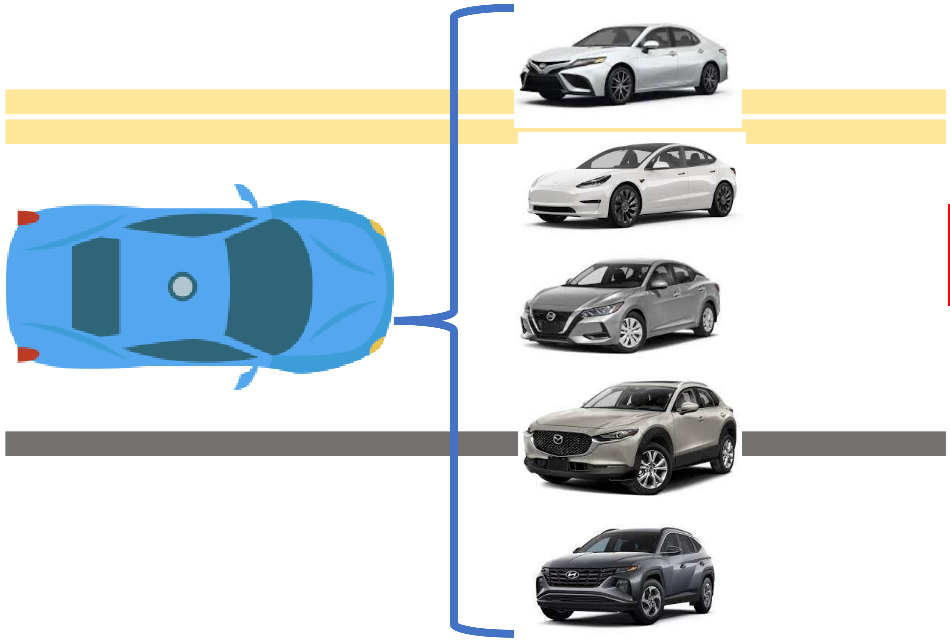
Commercial Systems



4 out of these 5 models are tested by us

- Not to directly reveal the exact model by including 1 confusing vehicle model

Commercial Systems



Top 15 leading car brands in the United States based on vehicle sales in 2023

Car brand	Sales number	TSR
Ford	1,904,038	✓
Toyota	1,888,941	✓
Chevrolet	1,702,700	
Honda	1,156,591	✓
Nissan	834,091	✓
Hyundai	796,506	✓
Kia	782,468	✓
Jeep	641,166	✓
Subaru	632,083	
GMC	563,692	✓
Ram	539,477	✓
Tesla	498,000	✓
Mazda	365,044	✓
BMW	361,654	✓
Volkswagen	329,025	✓

4 out of these 5 models are tested by us

- Not to directly reveal the exact model by including 1 confusing vehicle model

Commercial Systems

Top 15 leading car brands in the United States based on vehicle sales in 2023

Car brand	Sales number	TSR
Ford	1,904,038	✓
Toyota	1,888,941	✓
Chevrolet	1,702,700	
Honda	1,156,591	✓
Nissan	834,091	✓
Hyundai	796,506	✓
Kia	782,468	✓
Jeep	641,166	✓
Subaru	632,083	
GMC	563,692	✓
Ram	539,477	✓
Tesla	498,000	✓
Mazda	365,044	✓
BMW	361,654	✓
Volkswagen	329,025	✓

NOT any of the four tested car models for anonymity

4 out of these 5 models are tested by us

- Not to directly reveal the exact model by including 1 confusing vehicle model

Commercial Systems

TSR functions of the four vehicle models tested in our measurement study

	TSR functionality	
Vehicle model	STOP sign	Speed limit sign
Car 1 (denote as C1)	✓	✗
Car 2 (denote as C2)	✓	✓
Car 3 (denote as C3)	✗	✓
Car 4 (denote as C4)	✗	✓

4 out of these 5 models are tested by us

- Not to directly reveal the exact model by including 1 confusing vehicle model

Selected Attacks

- Focus on the **hiding attack** on measurement study

Selected Attacks

- Focus on the **hiding attack** on measurement study
- **Three prior works** so far that were able to demonstrate **black-box attack transferability** for the hiding attack effect in the physical world

Selected Attacks

- Focus on the **hiding attack** on measurement study
- **Three prior works** so far that were able to demonstrate **black-box attack transferability** for the hiding attack effect in the physical world

Physical Adversarial Examples for Object Detectors

Kevin Eykholt[1], Ivan Evtimov[2], Earlence Fernandes[2], Bo Li[3],
Amir Rahmati[4,6], Florian Tramèr[5], Atul Prakash[1], Tadayoshi Kohno[2], Dawn Song[3]
[1]University of Michigan
[2]University of Washington
[3]University of California, Berkeley
[4]Stony Brook University
[5]Stanford University
[6]Samsung Research America

$\mathbf{RP_2}$: Eykholt et al. WOOT 2017

Session 9B: ML Security III CCS '19, November 11–15, 2019, London, United Kingdom

Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors

Yue Zhao[1,2], Hong Zhu[1,2], Ruigang Liang[1,2], Qintao Shen[1,2], Shengzhi Zhang[3], Kai Chen[1,2*]
[1]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
[3]Department of Computer Science, Metropolitan College, Boston University, USA
{zhaoyue,zhuhong,liangruigang,shenqintao}@iie.ac.cn,shengzhi@bu.edu,chenkai@iie.ac.cn

SIB: Zhao et al. ACM CCS 2019

Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems

Wei Jia, School of Cyber Science and Engineering, Huazhong Univ. of Sci. & Tech., jiaw@hust.edu.cn
Zhaojun Lu*, School of Cyber Science and Engineering, Huazhong Univ. of Sci. & Tech., lzj_cse@hust.edu.cn
Haichun Zhang, Huazhong Univ. of Sci. & Tech., homer@thesimpsons.com
Zhenglin Liu, Huazhong Uni. of Sci. & Tech., liuzhenglin@hust.edu.cn
Jie Wang, Shenzhen Kaiyuan Internet Security Co., Ltdy, wangjie@seczone.cn
Gang Qu, University of Maryland, gangqu@umd.edu

FTE: Jia et al. NDSS 2022

Selected Attacks

- Focus on the **hiding attack** on measurement study
- **Three prior works** so far that were able to demonstrate **black-box attack transferability** for the hiding attack effect in the physical world
  - Highest potential to successfully attack **commercial systems**

Physical Adversarial Examples for Object Detectors

Kevin Eykholt[1], Ivan Evtimov[2], Earlence Fernandes[2], Bo Li[3],
Amir Rahmati[4,6], Florian Tramèr[5], Atul Prakash[1], Tadayoshi Kohno[2], Dawn Song[3]
[1]University of Michigan
[2]University of Washington
[3]University of California, Berkeley
[4]Stony Brook University
[5]Stanford University
[6]Samsung Research America

$\mathbf{RP_2}$: Eykholt et al. WOOT 2017

Session 9B: ML Security III
CCS '19, November 11–15, 2019, London, United Kingdom

Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors

Yue Zhao[1,2], Hong Zhu[1,2], Ruigang Liang[1,2], Qintao Shen[1,2], Shengzhi Zhang[3], Kai Chen[1,2*]
[1]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
[3]Department of Computer Science, Metropolitan College, Boston University, USA
{zhaoyue,zhuhong,liangruigang,shenqintao}@iie.ac.cn,shengzhi@bu.edu,chenkai@iie.ac.cn

SIB: Zhao et al. ACM CCS 2019

Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems

Wei Jia
School of Cyber Science and Engineering
Huazhong Univ. of Sci. & Tech.
jiaw@hust.edu.cn

Zhaojun Lu*
School of Cyber Science and Engineering
Huazhong Univ. of Sci. & Tech.
lzj_cse@hust.edu.cn

Haichun Zhang
Huazhong Univ. of Sci. & Tech.
homer@thesimpsons.com

Zhenglin Liu
Huazhong Uni. of Sci. & Tech.
liuzhenglin@hust.edu.cn

Jie Wang
Shenzhen Kaiyuan Internet Security Co., Ltdy
wangjie@seczone.cn

Gang Qu
University of Maryland
gangqu@umd.edu

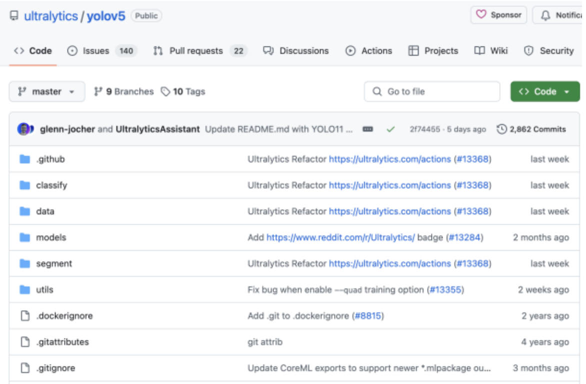
FTE: Jia et al. NDSS 2022

Surrogate Model

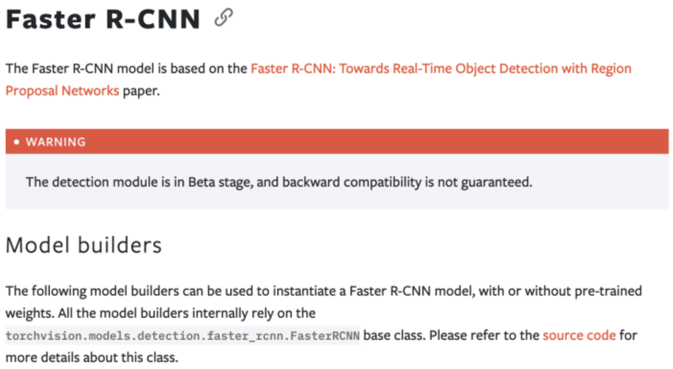
- Cover both **one-stage** and **two-stage** object detectors

Surrogate Model

- Cover both **one-stage** and **two-stage** object detectors



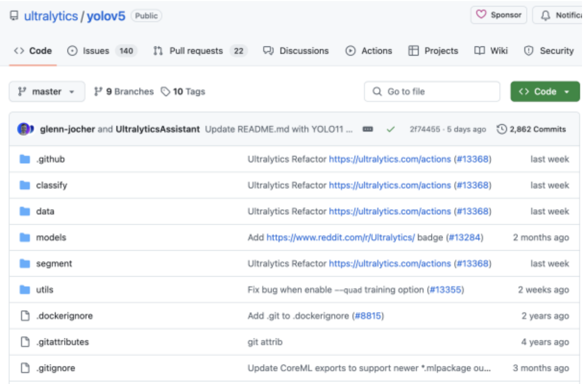
YOLO v5 (Y5)



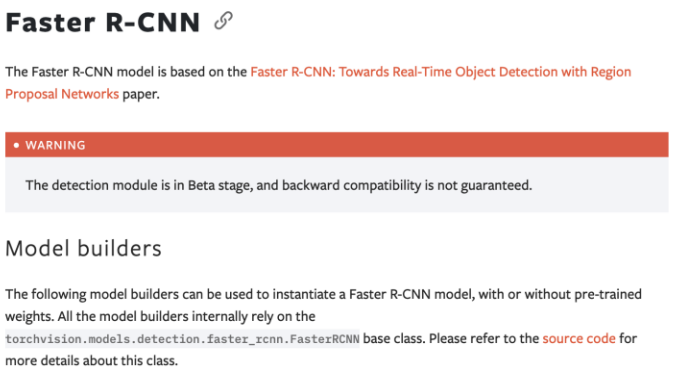
Faster-RCNN (FR)

Surrogate Model

- Cover both **one-stage** and **two-stage** object detectors
  - Generally used as surrogate model in the prior security research on TSR



YOLO v5 (Y5)



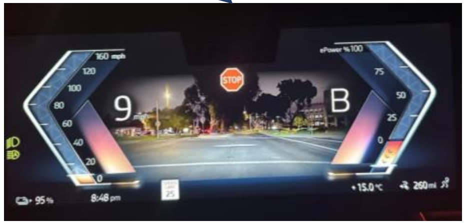
Faster-RCNN (FR)

Generated Attack Visualization

Real-World

TSR System-Level Attack Success Metric

TSR System-Level Attack Success Metric



NOT any of the four tested car models for anonymity

TSR System-Level Attack Success Metric

If the TSR system is able to correctly display the sign, the attack fails; otherwise, the attack succeed. Repeat *N* times.

NOT any of the four tested car models for anonymity

Overall Testing Results

	Original paper transferability	Surrogate model	C1	C2		C3	C4	Ave.
			STOP	STOP	Speed limit	Speed limit	Speed limit	
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	89.8%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

Certain Commercial TSRs are More Vulnerable

	Original paper transferability	Surrogate model	C1	C2		C3	C4	Ave.
			STOP	STOP	Speed limit	Speed limit	Speed limit	
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	$\boxed{\textbf{100\% (3/3)}}$	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	$\boxed{\textbf{100\% (3/3)}}$	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	89.8%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

Certain Commercial TSRs are More Vulnerable

	Original paper transferability	Surrogate model	C1	C2		C3	C4	Ave.
			STOP	STOP	Speed limit	Speed limit	Speed limit	
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	89.8%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

Certain Commercial TSRs are More Vulnerable

	Original paper transferability	Surrogate model	C1 STOP	C2 STOP	C2 Speed limit	C3 Speed limit	C4 Speed limit	Ave.
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	89.8%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

<u>Observation #1</u>: For certain commercial TSR systems, although from top brands in the US, their TSR functionality can **actually be much more vulnerable than academic TSR models** under black-box transfer attacks.

Attack Lacks Generalization across Commercial TSRs

	Original paper transferability	Surrogate model	C1 STOP	C2 STOP	C2 Speed limit	C3 Speed limit	C4 Speed limit	Ave.
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	89.8%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

Attack Lacks Generalization across Commercial TSRs

	Original paper transferability	Surrogate model	C1	C2		C3	C4	Ave.
			STOP	STOP	Speed limit	Speed limit	Speed limit	
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	89.8%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	$\boxed{51.6\%}$		0%	**33.3%**	0%	0%	0%	$\boxed{\mathbf{6.67\%}}$

Attack Lacks Generalization across Commercial TSRs

	Original paper transferability	Surrogate model	C1 STOP	C2 STOP	C2 Speed limit	C3 Speed limit	C4 Speed limit	Ave.
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	89.8%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

Attack Lacks Generalization across Commercial TSRs

	Original paper transferability	Surrogate model	C1	C2		C3	C4	Ave.
			STOP	STOP	Speed limit	Speed limit	Speed limit	
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	89.8%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

<ins>Observation #1 (cont'd)</ins>: Such black-box commercial system attack capability is currently not generalizable over different representative commercial system models and sign types.

Attack Lacks Generalization across Commercial TSRs

	Original paper transferability	Surrogate model	C1	C2		C3	C4	Ave.
			STOP	STOP	Speed limit	Speed limit	Speed limit	
Benign traffic sign			**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100% (3/3)**	**100%**
RP_2	18.9%	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
SIB	46.1%	Y5	0% (0/3)	**100% (3/3)**	0% (0/3)	0% (0/3)	0% (0/3)	**20%**
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
FTE	$\boxed{89.8\%}$	Y5	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
		FR	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

Attack Lacks Generalization across Commercial TSRs

3) ***Successful attacks against YOLO v5 based object detector and TSR system in a 2021 model vehicle:*** To the best of our knowledge, this is the first set of adversarial attacks against YOLO v5 based object detectors in the physical domain. We successfully launch four attack vectors, especially NTA and TA, that are life-threatening in the real world. Our physical AEs also exhibit satisfactory transferability when attacking a production-grade TSR system of a brand-new 2021 model vehicle.

[Jia et al. NDSS 2022: Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems]

	Original pa[…] transferabil[…]						C4 Speed limit	Ave.
Benign traffic sign							**100% (3/3)**	**100%**
RP_2	18.9%						0% (0/3) 0% (0/3)	0% **20%**
SIB	46.1%						0% (0/3) 0% (0/3)	**20%** 0%
FTE	89.8%	Y5 FR	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% 0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

Attack Lacks Generalization across Commercial TSRs

> *3) **Successful attacks against YOLO v5 based object detector and TSR system in a 2021 model vehicle**:* To the best of our knowledge, this is the first set of adversarial attacks against YOLO v5 based object detectors in the physical domain. We successfully launch four attack vectors, especially NTA and TA, that are life-threatening in the real world. Our physical AEs also exhibit satisfactory transferability when attacking a production-grade TSR system of a brand-new 2021 model vehicle.
>
> *[Jia et al. NDSS 2022: Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems]*

	Original pa transferabil						C4 Speed limit	Ave.
Benign traffic sign							**100% (3/3)**	**100%**
RP$_2$	18.9%						0% (0/3) 0% (0/3)	0% **20%**
SIB	46.1%						0% (0/3) 0% (0/3)	**20%** 0%
FTE	89.8%	Y5 FR	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% 0%
Ave. over all attacks	51.6%		0%	**33.3%**	0%	0%	0%	**6.67%**

<u>Observation #1 (cont'd)</u>: This further reveals the lack of generalizability of the reported commercial TSR system attack success in the original FTE paper, which cannot be revealed without the large-scale commercial system testing efforts in this paper.

Discrepancy in Commercial and Academic TSR

	Original paper transferability	Surrogate model
Benign traffic sign		
RP_2	18.9%	Y5 FR
SIB	46.1%	Y5 FR
FTE	89.8%	Y5 FR
Ave. over all attacks	$\boxed{51.6\%}$	

it	C3 Speed limit	C4 Speed limit	Ave.
3)	**100% (3/3)**	**100% (3/3)**	**100%**
	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% **20%**
	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	**20%** 0%
	0% (0/3) 0% (0/3)	0% (0/3) 0% (0/3)	0% 0%
	0%	0%	$\boxed{\textbf{6.67\%}}$

Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- Observation #2: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.

Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- Observation #2: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.

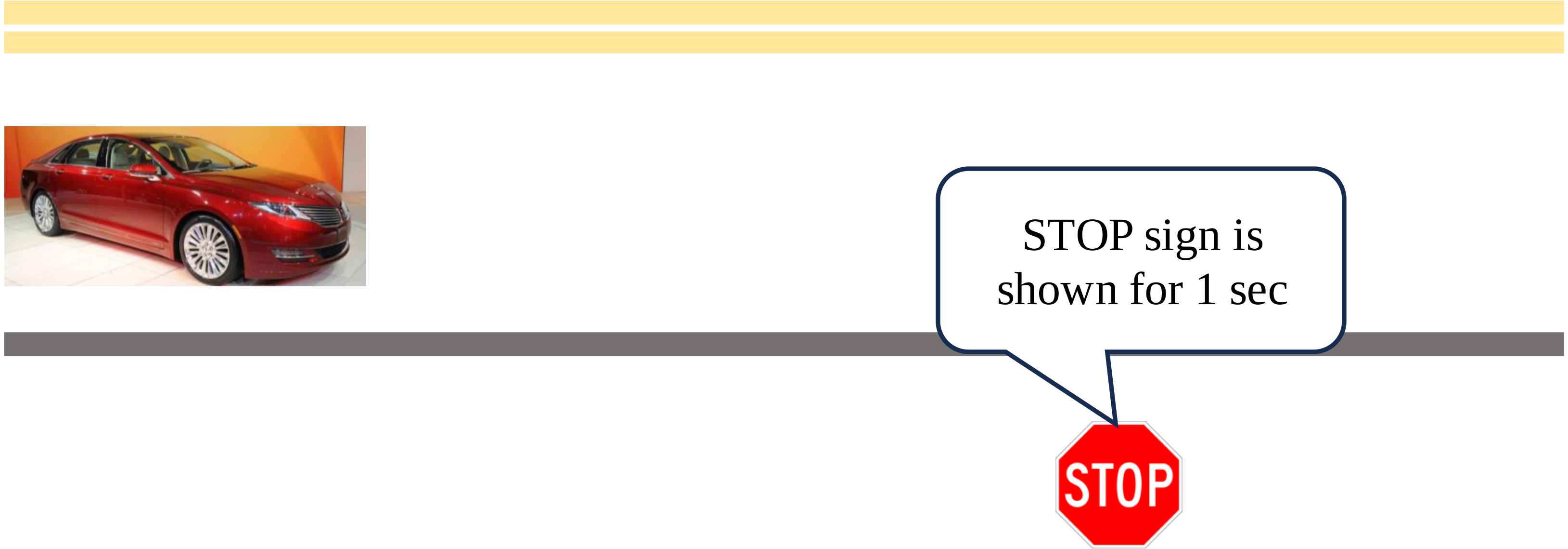
Observation #2 (cont'd): **Spatial memorization design** exhibits an effect that once a sign is detected, both the **detected sign type** and the **detected location** are **persistently memorized** until **the sign's reaction task is finished**

Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- Observation #2: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.

Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- Observation #2: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.



Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- Observation #2: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.

Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- <u>Observation #2</u>: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.

Hide the STOP sign

Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- Observation #2: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.

Wait for 60 sec

Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- Observation #2: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.

Finding: Unexpected Spatial Memorization Design in Commercial TSR Systems

- Observation #2: One major factor might be an unexpected **spatial memorization** design that commonly exists in commercial TSRs.

Observation #2 (cont'd): **Spatial memorization design** exhibits an effect that once a sign is detected, both the **detected sign type** and the **detected location** are **persistently memorized** until **the sign's reaction task is finished**

Limitation of Existing Model-Level Attack Success Metrics

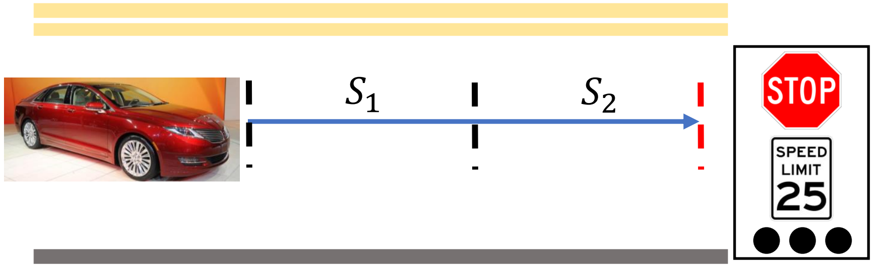
- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level

Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level

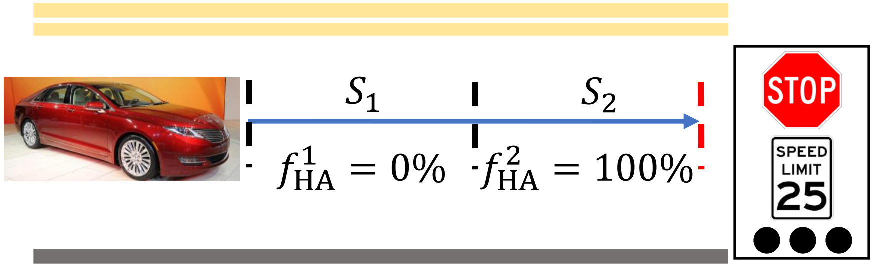
Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level



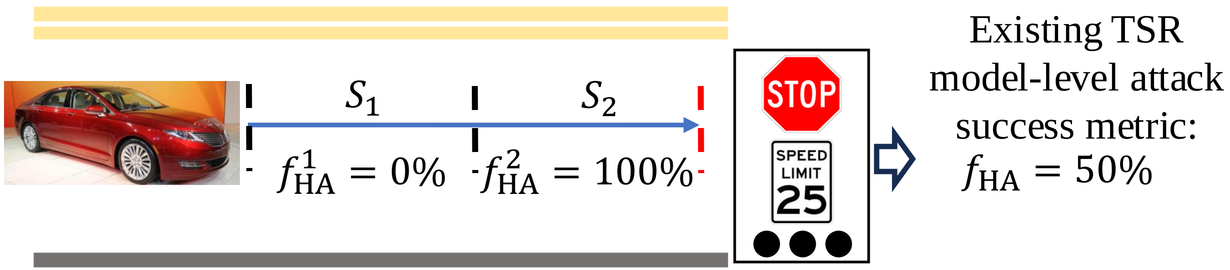
Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level



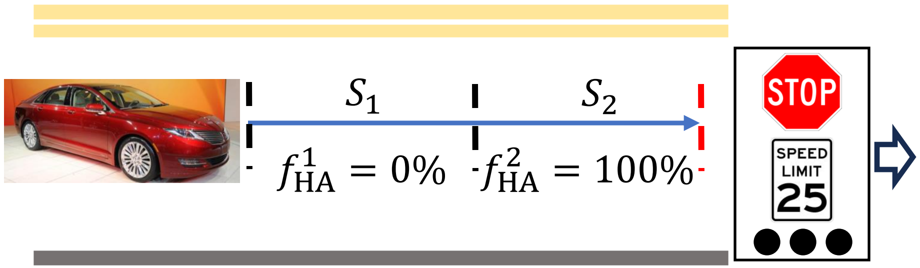
Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level



Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level

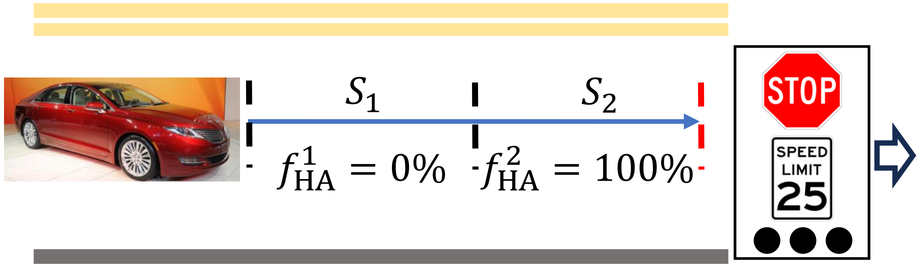


Existing TSR model-level attack success metric: $f_{\mathrm{HA}} = 50\%$

TSR system-level attack success: 0%

Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level



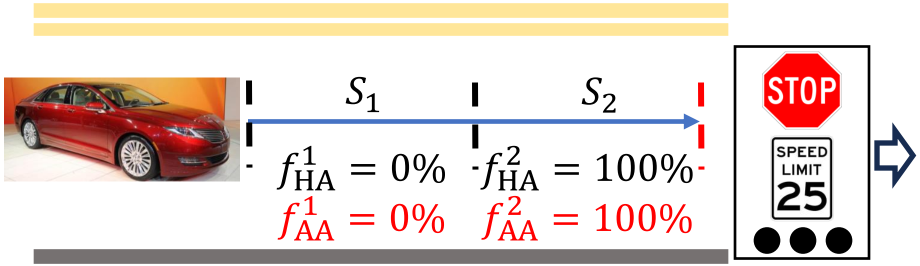
Existing TSR model-level attack success metric: $f_{\mathrm{HA}} = 50\%$

TSR system-level attack success: 0%

Due to the spatial memorization

Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level

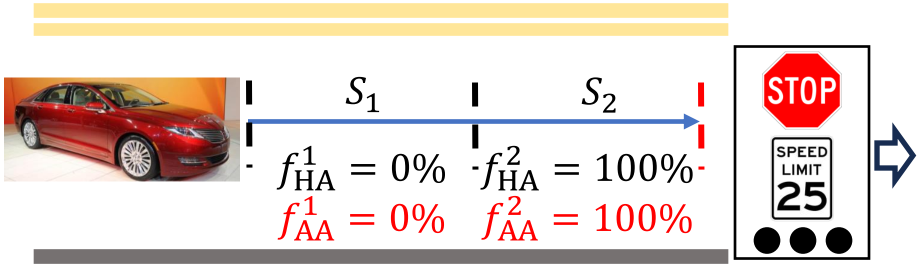


Existing TSR model-level attack success metric: $f_{\mathrm{HA}} = 50\%$

TSR system-level attack success: 0%

Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level

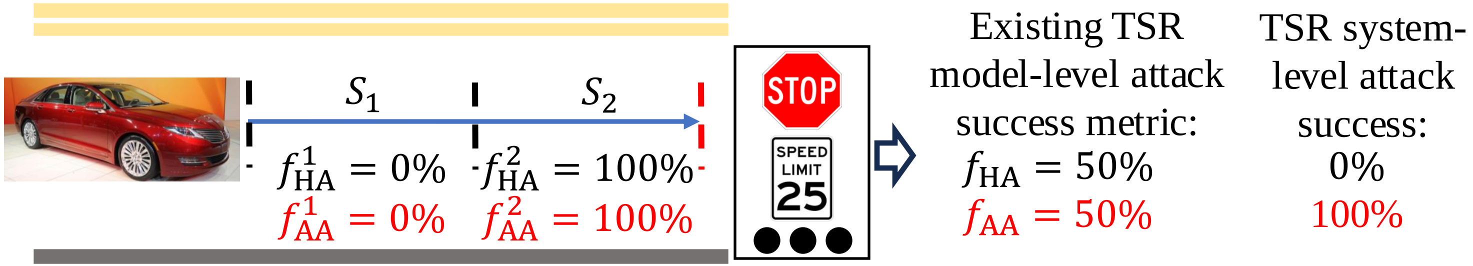


Existing TSR model-level attack success metric:
$f_{\mathrm{HA}} = 50\%$
$f_{\mathrm{AA}} = 50\%$

TSR system-level attack success:
0%

Limitation of Existing Model-Level Attack Success Metrics

- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level

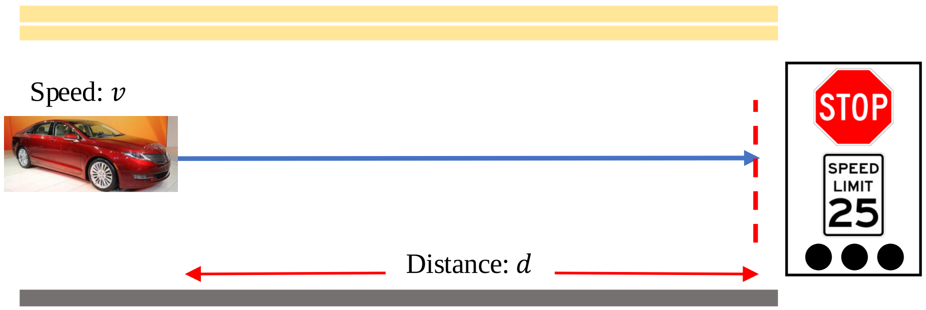


Limitation of Existing Model-Level Attack Success Metrics

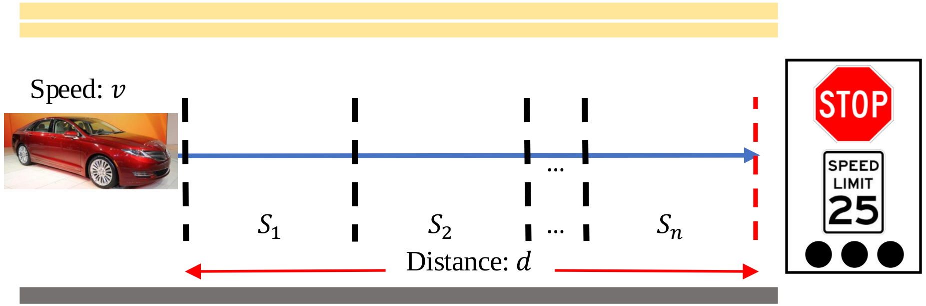
- The spatial memorization design can significantly impact the success of existing adversarial attacks at the TSR system level

Given that such an unexpected **spatial memorization design** can create such a significant **discrepancy** between the **TSR model-level** attack effect and that at the **TSR system level**, we further design **new attack success metrics** that can mathematically model its impact on the **TSR system-level** attack success for both **hiding and appearing attacks**

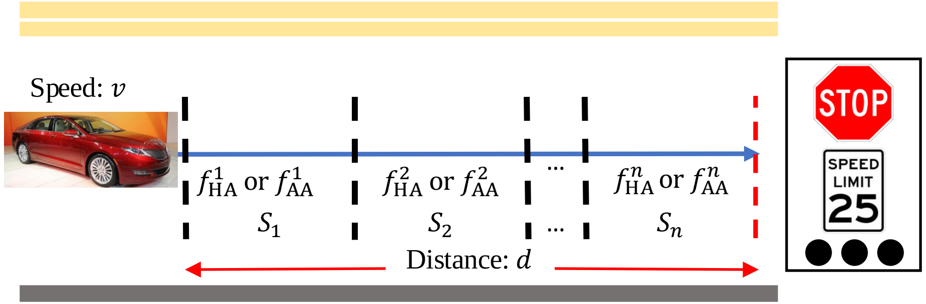
New Attack Success Metric Design that Can Mathematically Model Spatial Memorization



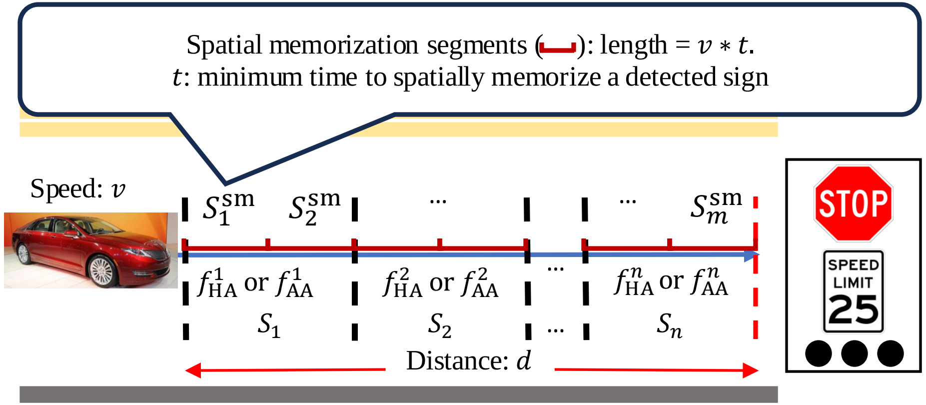
New Attack Success Metric Design that Can Mathematically Model Spatial Memorization



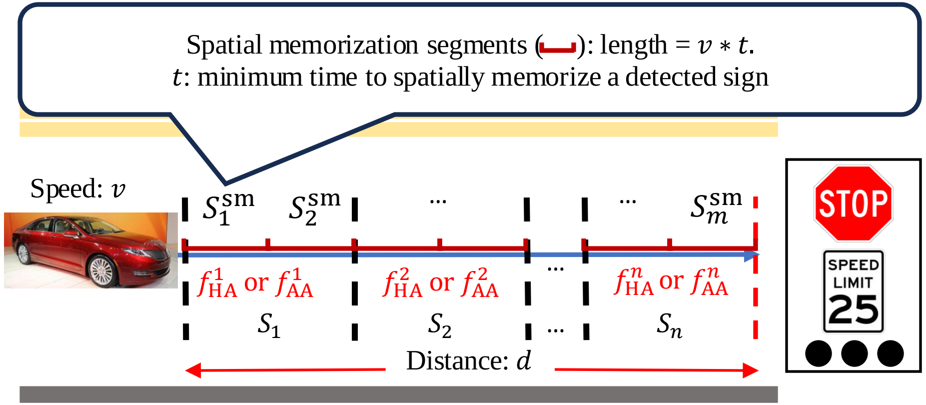
New Attack Success Metric Design that Can Mathematically Model Spatial Memorization



New Attack Success Metric Design that Can Mathematically Model Spatial Memorization



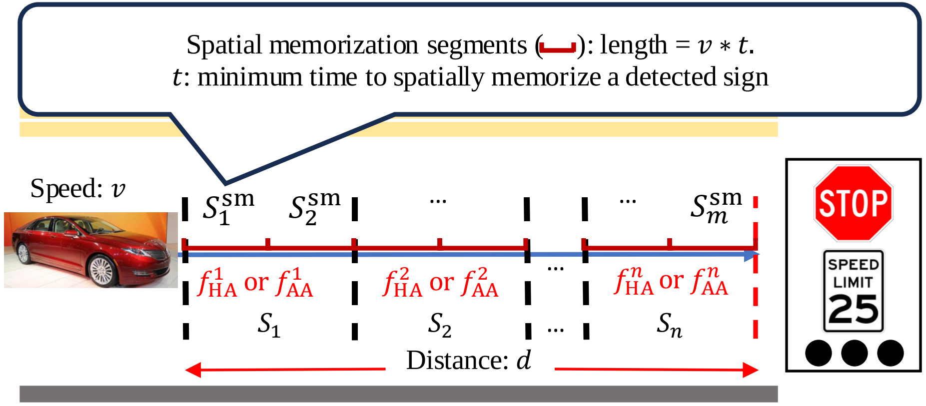
New Attack Success Metric Design that Can Mathematically Model Spatial Memorization



$$\mathrm{SysHA} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{m}{n}} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{d}{nvt}}$$

- *Hiding attack*: The attack has to be **continuously successful** at **all** possible detection moments that can trigger such memorization **before the vehicle passes the sign**.

New Attack Success Metric Design that Can Mathematically Model Spatial Memorization

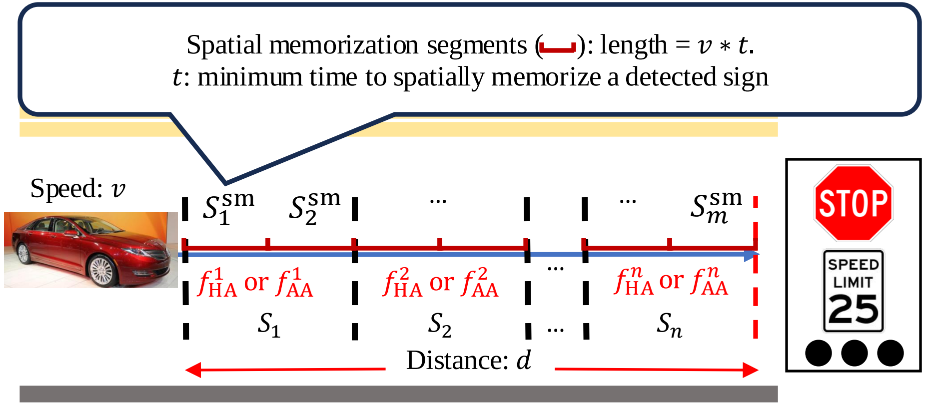


$$\mathrm{SysHA} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^i)^{\frac{m}{n}} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^i)^{\frac{d}{nvt}}$$

$$\mathrm{SysAA} = 1 - \prod_{i=1}^{n} (1 - f_{\mathrm{AA}}^i)^{\frac{d}{nvt}}$$

- ***Appearing attack***: As long as attack can succeed in **any** of detection moments, the TSR system-level attack effect can be achieved.

New Attack Success Metric Design that Can Mathematically Model Spatial Memorization



$$\text{SysHA} = \prod_{i=1}^{n} (f_{\text{HA}}^i)^{\frac{m}{n}} = \prod_{i=1}^{n} (f_{\text{HA}}^i)^{\frac{d}{nvt}}$$

$$\text{SysAA} = 1 - \prod_{i=1}^{n} (1 - f_{\text{AA}}^i)^{\frac{d}{nvt}}$$

- *Appearing attack*: As long as attack can succeed in **any** of detection moments, the TSR system-level attack effect can be achieved.

Observation #3: Due to spatial memorization, hiding attacks are theoretically harder (if not equally hard) than appearing attacks in achieving TSR system-level attack success.

New Attack Success Metric Design that Can Mathematically Model Spatial Memorization

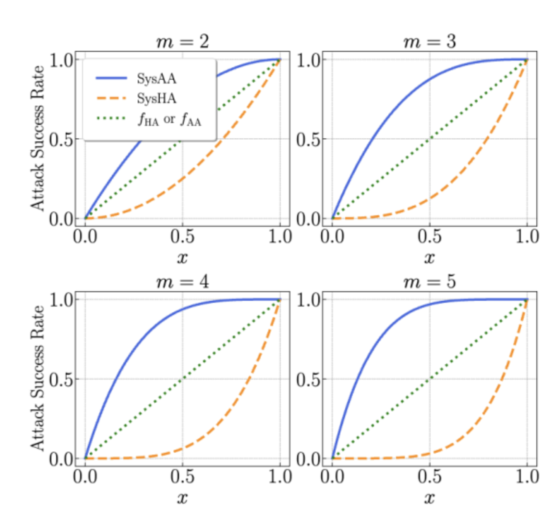
Theoretical Analysis

$$\mathrm{SysAA} = \sum_{A\in(\mathcal{P}(S)\setminus\emptyset)} \left(\prod_{S_i\in A} (f^i_{\mathrm{AA}})^{\frac{d}{nvt}} \prod_{S_j\in(S\setminus A)} (1-f^j_{\mathrm{AA}})^{\frac{d}{nvt}} \right) \tag{3}$$

When $f^i_{\mathrm{HA}} = f^i_{\mathrm{AA}}$, $\mathrm{SysHA} = \prod_{i=1}^{n}(f^i_{\mathrm{HA}})^{\frac{d}{nvt}} = \prod_{i=1}^{n}(f^i_{\mathrm{AA}})^{\frac{d}{nvt}}$, which is actually one instance of $A \in (\mathcal{P}(S)\setminus A)$, i.e., $A = S$. Thus, we can calculate $\mathrm{SysAA} - \mathrm{SysHA}$:

$$\mathrm{SysAA} - \mathrm{SysHA} = \sum_{A\in\mathcal{P}(S)\setminus\{\emptyset,S\}} \left(\prod_{S_i\in A} (f^i_{\mathrm{AA}})^{\frac{d}{nvt}} \prod_{S_j\in(S\setminus A)} (1-f^j_{\mathrm{AA}})^{\frac{d}{nvt}} \right) \tag{4}$$

For $\forall A \in \mathcal{P}(S)\setminus\{\emptyset,S\}$, $f^i_{\mathrm{AA}} \geqslant 0$ and $(1-f^j_{\mathrm{AA}}) \geqslant 0$, where $S_i \in A$ and $S_j \in (S\setminus A)$, we can have $\prod(f^i_{\mathrm{AA}})^{\frac{d}{nvt}}\prod(1-f^j_{\mathrm{AA}})^{\frac{d}{nvt}} \geqslant 0$, thus $\mathrm{SysAA} - \mathrm{SysHA} \geqslant 0$, and consequently, $\mathrm{SysAA} \geqslant \mathrm{SysHA}$. □



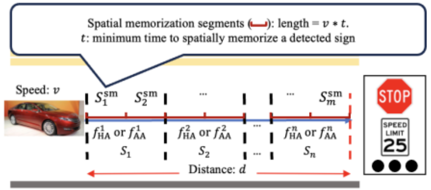
Numerical Analysis

Observation #3: Due to spatial memorization, hiding attacks are theoretically harder (if not equally hard) than appearing attacks in achieving TSR system-level attack success.

Observations for Revisiting Existing Research

New Metric Design:
Surrogate TSR System-Level
Attack Success Metrics

Revisiting Evaluations, Designs, and Attack Capabilities of Prior Works in this Problem Space

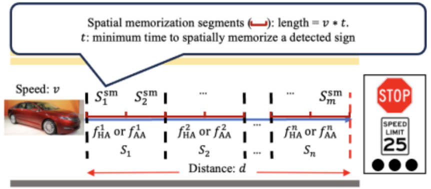


$$\mathrm{SysHA} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{m}{n}} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{d}{nvt}}$$

$$\mathrm{SysAA} = 1 - \prod_{i=1}^{n} (1 - f_{\mathrm{AA}}^{i})^{\frac{d}{nvt}}$$

Observations for Revisiting Existing Research

New Metric Design: Surrogate TSR System-Level Attack Success Metrics



$$\mathrm{SysHA} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{m}{n}} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{d}{nvt}}$$

$$\mathrm{SysAA} = 1 - \prod_{i=1}^{n} (1 - f_{\mathrm{AA}}^{i})^{\frac{d}{nvt}}$$

Revisiting Evaluations, Designs, and Attack Capabilities of Prior Works in this Problem Space

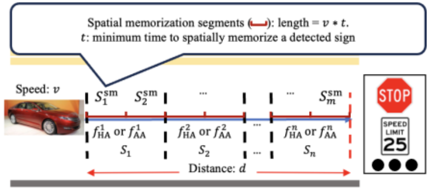
White-Box Attack

Prior works **may not be effective** at TSR system level. (Drop from ~56% to ~7%)

	f_{HA}							
	Distance ranges (meters)						Ave.	SysHA
	0-5	5-10	10-15	15-20	20-25	25-30		
RP_2	41.8%	10.0%	23.8%	65.4%	99.9%	100%	56.8%	6.6%
SIB	84.6%	56.6%	82.0%	99.2%	100%	100%	87.1%	45.1%
FTE	88.9%	57.1%	13.6%	3.1%	47.8%	74.5%	47.5%	5.2%

Observations for Revisiting Existing Research

New Metric Design: Surrogate TSR System-Level Attack Success Metrics



$$\mathrm{SysHA} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{m}{n}} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{d}{nvt}}$$

$$\mathrm{SysAA} = 1 - \prod_{i=1}^{n} (1 - f_{\mathrm{AA}}^{i})^{\frac{d}{nvt}}$$

Revisiting Evaluations, Designs, and Attack Capabilities of Prior Works in this Problem Space

White-Box Attack

Prior works **may not be effective** at TSR system level. (Drop from ~56% to ~7%)

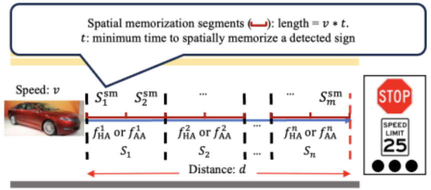
Black-Box Transfer Attack

Attack success of prior works at TSR system level can **be much lower than expected** (~13%) for hiding attack.

	Original paper transferability	Transfer attack success rates (averaged over a set of six transfer target models (§IV-B)							
		f_{HA}							SysHA
		0-5m	5-10m	10-15m	15-20m	20-25m	25-30m	Ave.	
RP_2	18.9%	36.4%	32.0%	29.6%	46.0%	61.3%	50.0%	42.6%	14.5%
SIB	46.1%	20.7%	26.5%	37.2%	42.6%	54.9%	51.2%	38.9%	12.4%
FTE	89.8%	29.2%	36.4%	29.3%	34.0%	45.5%	40.1%	35.7%	11.0%
Ave.	51.6%	28.8%	31.6%	32.0%	40.9%	53.9%	47.1%	39.1%	12.6%

Observations for Revisiting Existing Research

New Metric Design: Surrogate TSR System-Level Attack Success Metrics



$$\mathrm{SysHA} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{m}{n}} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{d}{nvt}}$$

$$\mathrm{SysAA} = 1 - \prod_{i=1}^{n} (1 - f_{\mathrm{AA}}^{i})^{\frac{d}{nvt}}$$

Revisiting Evaluations, Designs, and Attack Capabilities of Prior Works in this Problem Space

White-Box Attack

Prior works **may not be effective** at TSR system level. (Drop from ~56% to ~7%)

Black-Box Transfer Attack

Attack success of prior works at TSR system level can **be much lower than expected** (~13%) for hiding attack.

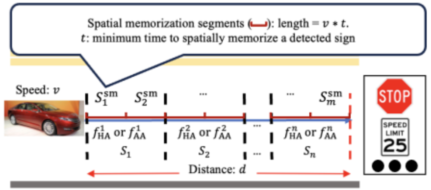
Revisiting Existing Attack Success Metrics

Using the **hiding and appearing attacks** proposed from the same prior work, the **hiding** one can be **much harder**. However, if using the **existing metrics**, such **relative attack hardness** can be the **completely opposite**

	Hiding attack		Appearing attack	
SIB [5]	f_{HA}	SysHA	f_{AA}	SysAA
White-box attack	87.1%	45.1%	29.1%	87.6%
Black-box transfer attacks	38.9%	12.4%	31.7%	64.2%

Observations for Revisiting Existing Research

New Metric Design: Surrogate TSR System-Level Attack Success Metrics



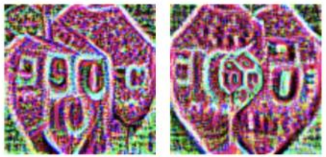
$$\mathrm{SysHA} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{m}{n}} = \prod_{i=1}^{n} (f_{\mathrm{HA}}^{i})^{\frac{d}{nvt}}$$

$$\mathrm{SysAA} = 1 - \prod_{i=1}^{n} (1 - f_{\mathrm{AA}}^{i})^{\frac{d}{nvt}}$$

Revisiting Evaluations, Designs, and Attack Capabilities of Prior Works in this Problem Space

White-Box Attack | Black-Box Transfer Attack

RP_2 [9]	f_{AA}						SysAA
	0-5m	5-10m	10-15m	15-20m	20-25m	Ave.	
w/o NAE	86.8%	100%	64.7%	66.9%	19.5%	67.6%	98.2%
w/ NAE	100%	100%	100%	88.3%	25.8%	82.8%	100%



RP_2 w/o Nested AE | RP_2 w/ Nested AE

Judgement of the Value of New Attack Designs

The benefits of certain attack designs can be **seemingly high** (e.g., >20% attack success rate increase) using **prior TSR model-level success metrics**, but **nearly negligible** (e.g., only 1% increase) at the **TSR system level**

Conclusion

- *First large-scale measurement of physical-world adversarial attacks against commercial TSR:*
  - Uncover a total of **7 novel observations**
- *Discovery and analysis of spatial memorization:*
  - Discover a **spatial memorization design** that commonly exists in today's commercial TSRs
  - Create a **discrepancy** between TSR model-level attack effect and that at TSR system level.
- *New attack success metric designs:*
  - Mathematically model the impact of this design on the TSR system-level attack success
  - **Revisit** the evaluations, designs, and capabilities of existing attacks in this problem space

Conclusion

- ***First large-scale measurement of physical-world adversarial attacks against commercial TSR:***
  - Uncover a total of **7 novel observations**
- ***Discovery and analysis of spatial memorization:***
  - Discover a **spatial memorization design** that commonly exists in today's commercial TSRs
  - Create a **discrepancy** between TSR model-level attack effect and that at TSR system level.
- ***New attack success metric designs:***
  - Mathematically model the impact of this design on the TSR system-level attack success
  - **Revisit** the evaluations, designs, and capabilities of existing attacks in this problem space
- ***Performed Responsible Vulnerability Disclosure:***
  - **Informed** AD companies under our measurements and provided **anonymity** to protect the affected vehicle manufacturer

Thank you!

Revisiting Physical-World Adversarial Attack on Traffic Sign Recognition: A Commercial Systems Perspective

Ningfei Wang, Shaoyuan Xie, Takami Sato, Yunpeng Luo,
Kaidi Xu*, Qi Alfred Chen
University of California, Irvine and *Drexel University
ningfei.wang@uci.edu, *kx46@drexel.edu*, *and* *alfchen@uci.edu*

Scan to visit our project website