THNS 2024, November 5-7, 2024

AI & Road Flow

Trusted Perception Method for Traffic Signs That Are Physically Attacked

Shize Huang[1], Qunyao Tan[1], Zhaoxin Zhang*[1], Qianhui Fan[1], Yi Zhang[1], Xingying Li[1].

1 Key Laboratory of Rail Infrastructure Durability and System Safety, Tongji University, Shanghai, China, 201804.

Speaker: Qunyao Tan

Tuesday 5/11/2024

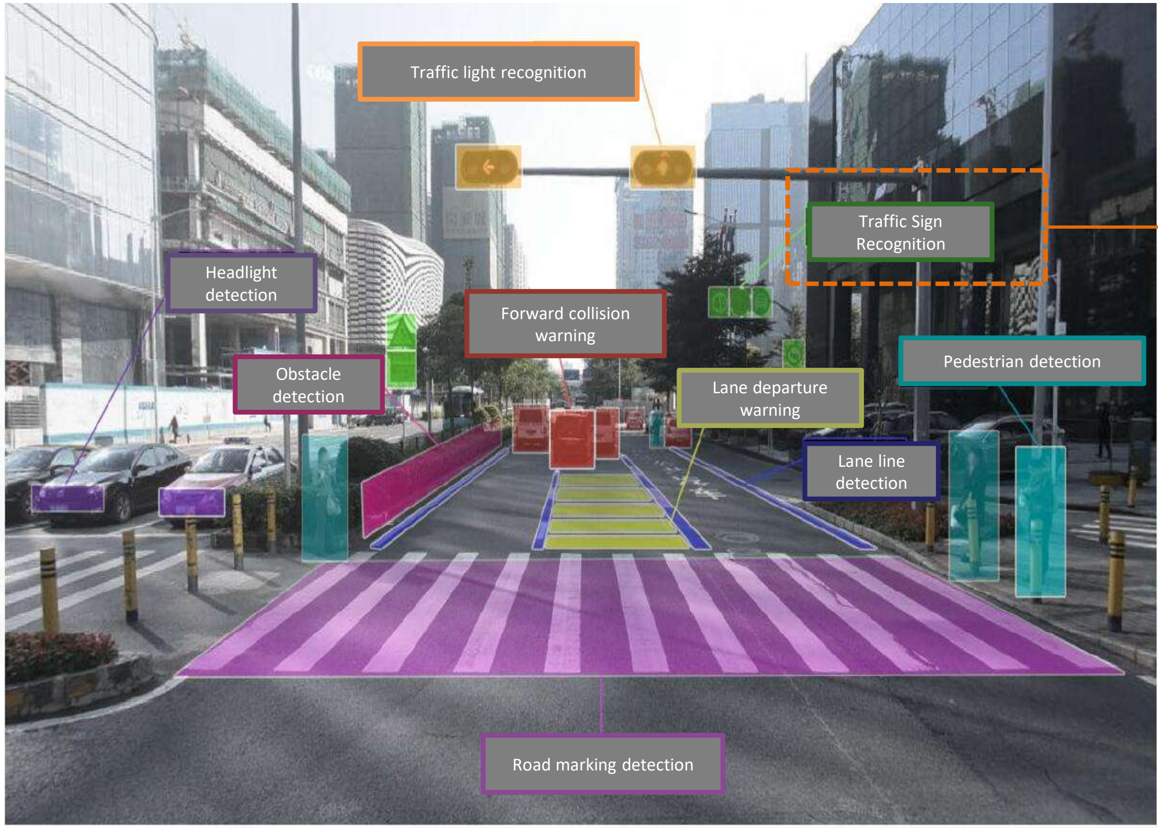
同濟大學
TONGJI UNIVERSITY

目录
CONTENTS

Background

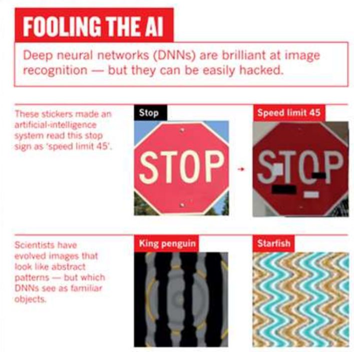
Background

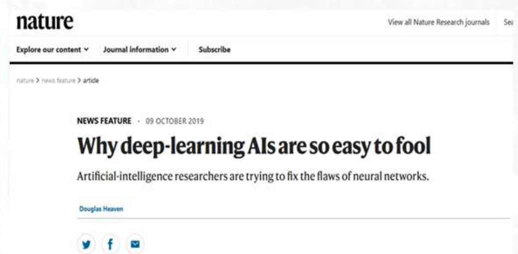
- **Applications and Challenges of Deep Learning in Intelligent Transportation Systems.**



Deep Learning in Traffic Perception

Vulnerabilities of Current Systems





[1] Heaven D. Deep Trouble for Deep Learning[J]. Nature, ***574****, 163-166 (2019)*

Challenges for trusted perception method

Background

Traffic Sign Recognition

Prohibition signs

Warning signs

Directive signs

Autonomous vehicles

Autonomous trams

Background

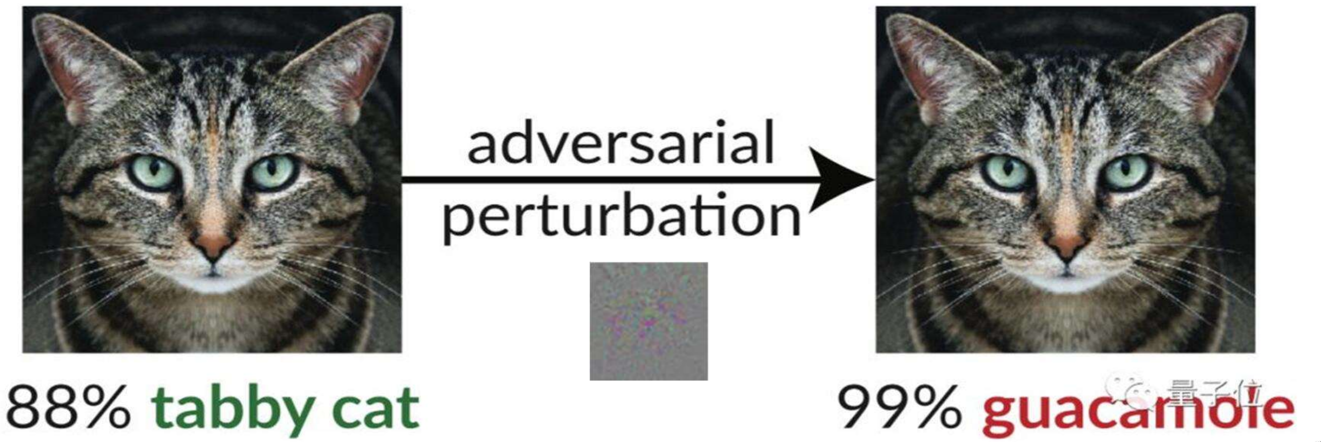
- **Adversarial attack : Introduction of Adversarial Examples**

Concept introduced by Christian Szegedy et al. [2]

Adding subtle perturbations to input data

$$\min \| x_{adv} - x \|, \text{s.t.} F(x_{adv}) \neq y$$

Model outputs incorrect results with high confidence



An example of an adversarial attack

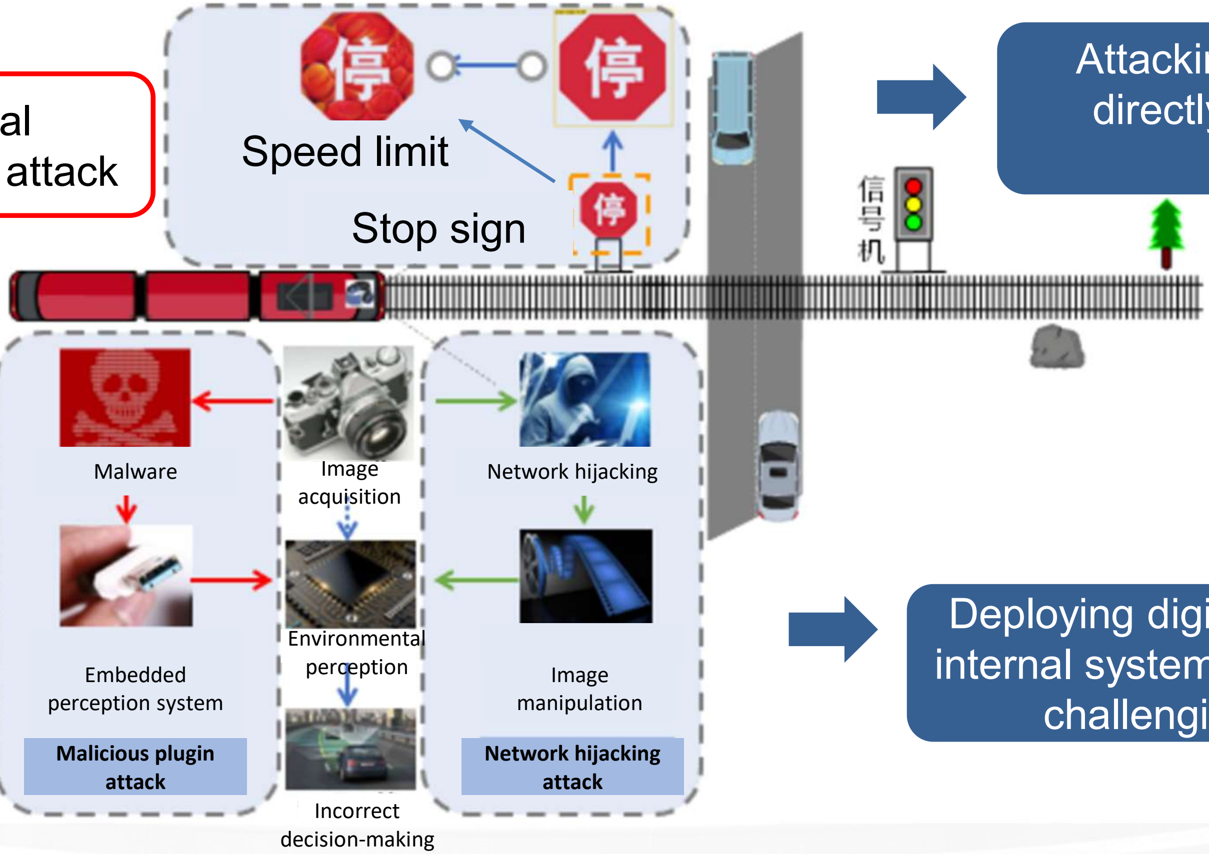
[2] Szegedy C, Zaremba W, Sutskever I, Bruna, J, Erhan, D, Goodfellow, I, Fergus R. Intriguing properties of neural networks. Computer Science. 2013:1-10.

Background

- **Adversarial attack on traffic sign recognition systems**

Physical adversarial attack

Digital adversarial attack



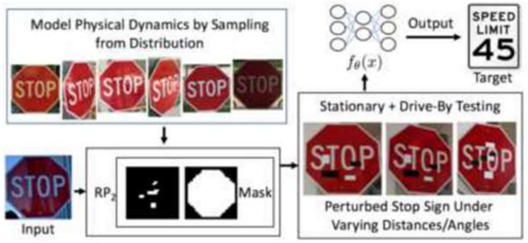
Attacking roadside signs directly without system intrusion.

Deploying digital attacks requires internal system intrusion, making it challenging to achieve.

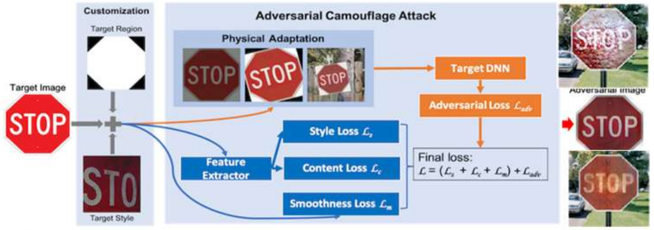
- **Physical adversarial attack is of more practical relevance.**

Background

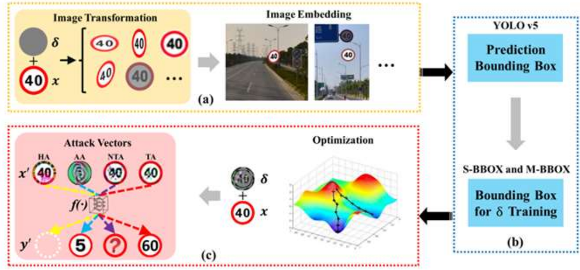
- **Different forms of physical adversarial examples**



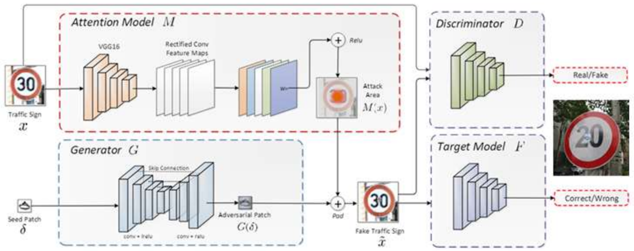
Robust Physical Perturbations (RP2) [3]



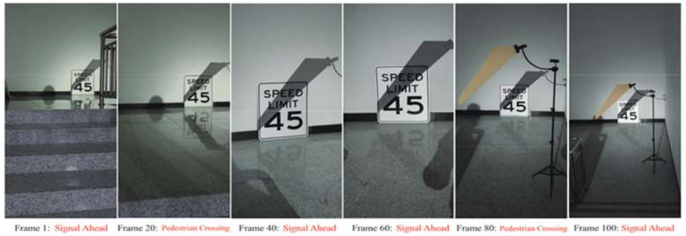
Adversarial Camouflage (AdvCam) [4]



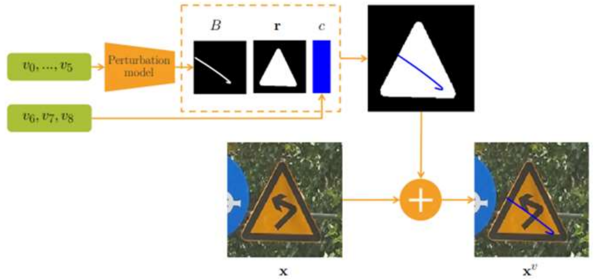
4A physical adversarial attack[5]



Perceptual-sensitive generative adversarial network (PS-GAN) [6]



Stealthy and Effective Physical-world Adversarial Attack（ShadowAttack） [7]



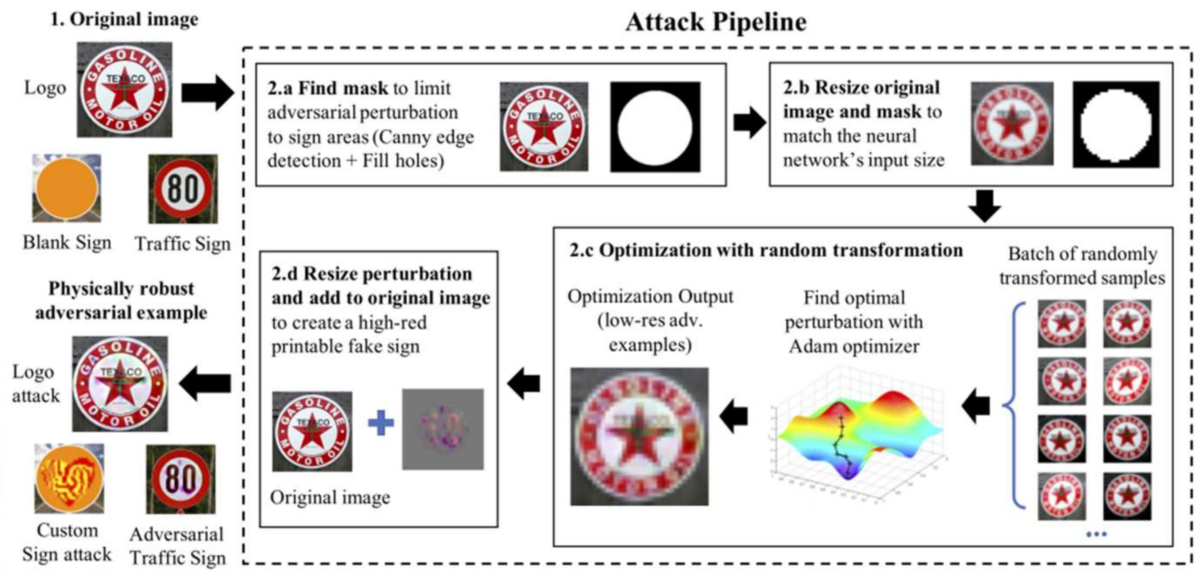
Adversarial Scratches[8]

- **The diversity of these physical adversarial samples poses challenges for reliable detection methods.**

Methods

Adversarial sign generation

- **Physical adversarial traffic sign generation method--- DARTS [9]**



Ground label: Speed limit 30
Recognition result: Speed limit 80

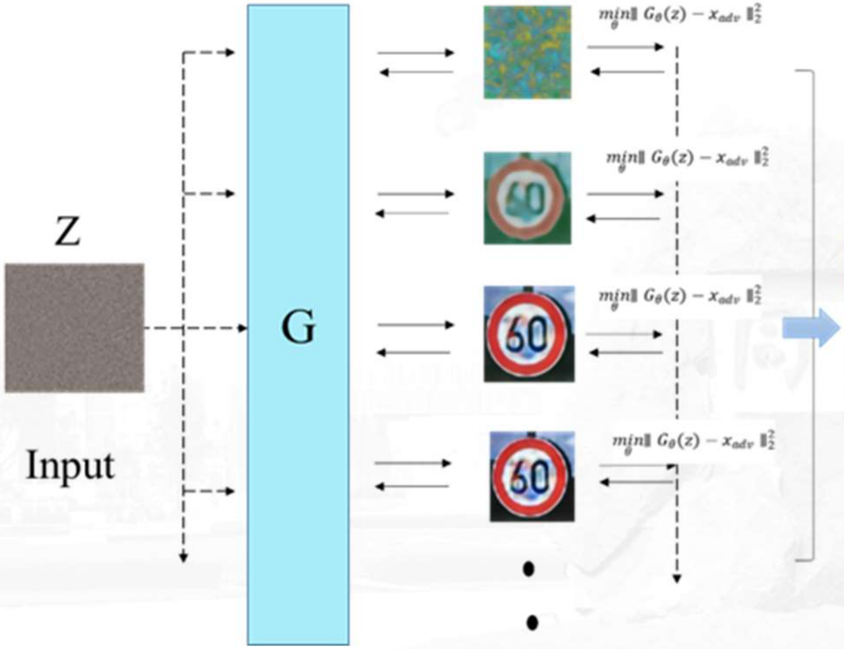
Ground label: Speed limit 60
Recognition result: Speed limit 20

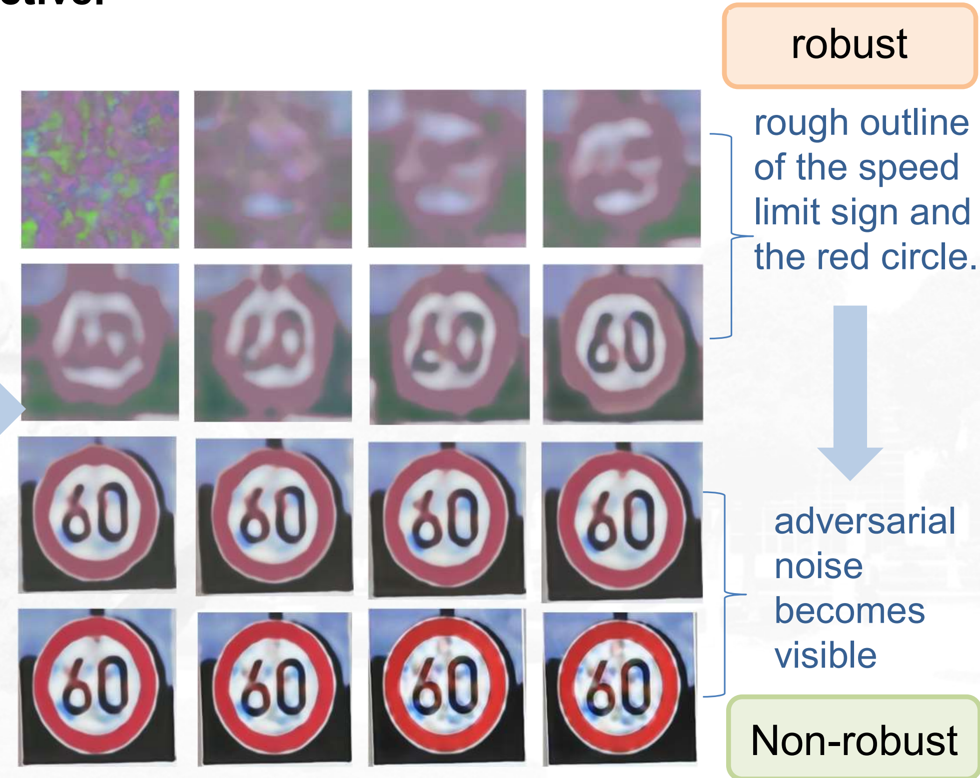
Ground label: No pass
Recognition result: Speed limit 20

Ground label: No Entry
Recognition result: Speed limit 60

Physical adversarial traffic signs robust to different distances and angles in the real world

- **Our defense pipeline is motivated by the insight to take unsupervised image reconstruction from a robust/non-robust feature learning perspective.**

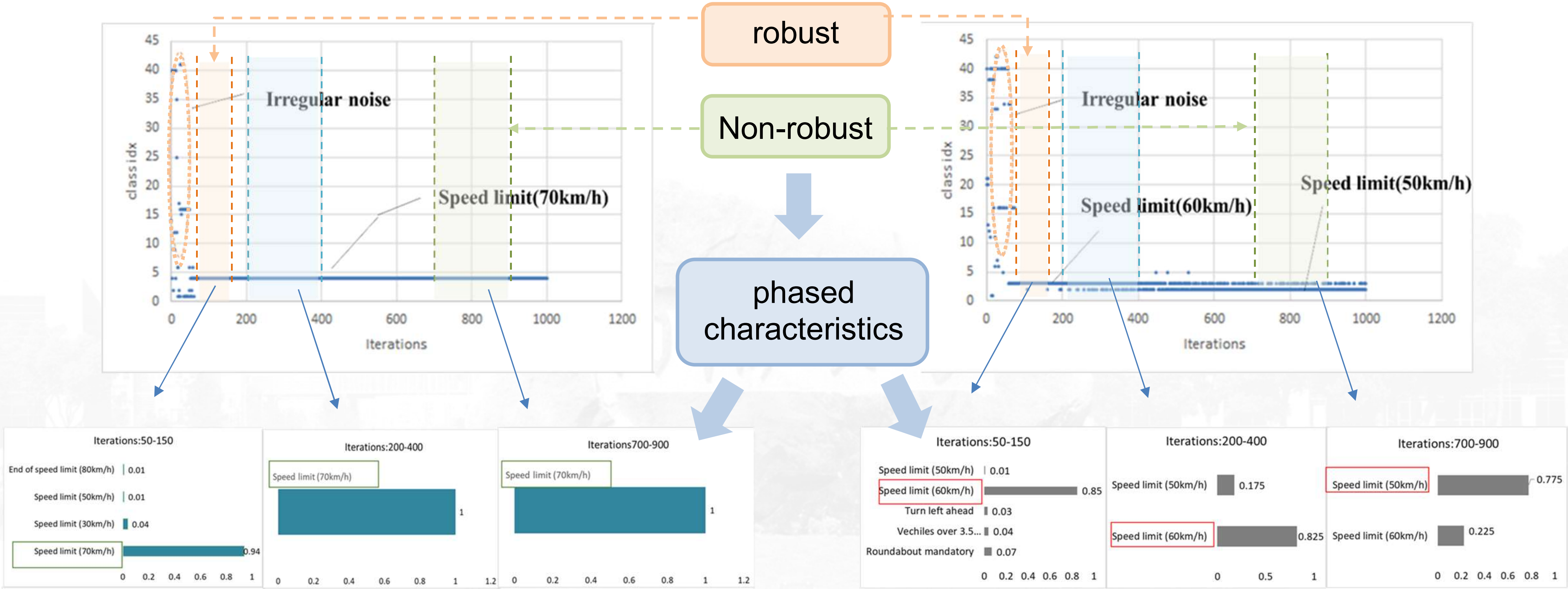




- ➢ Process of reconstructing with Deep Image Prior(DIP)[10], G is a generator network based on U-Net structure.
- ➢ Reconstruction process images of a physical adversarial traffic sign misclassified as ‘speed limit 50’.

Phased classification results

- **Class distribution of reconstructed images**

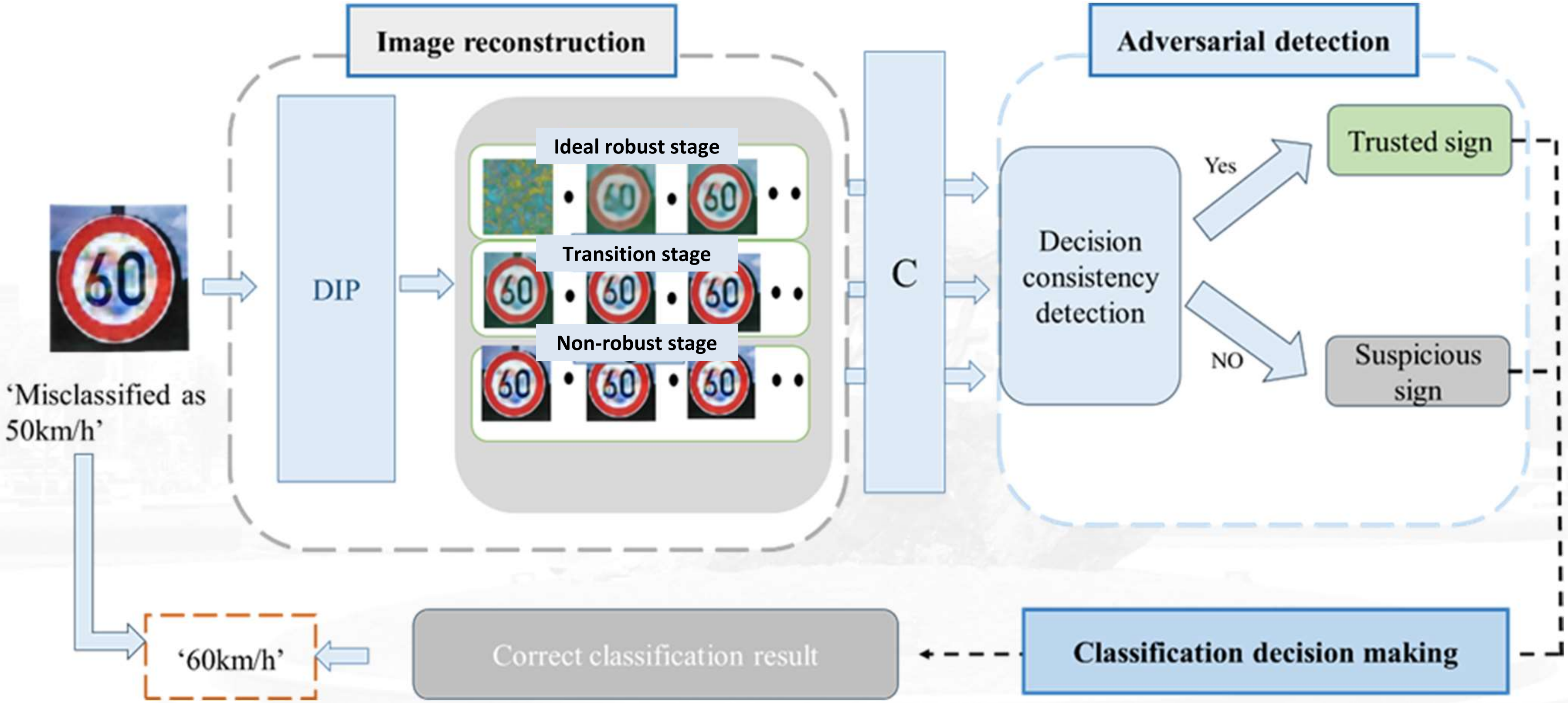


- Class distribution of the clean traffic sign 'speed limit (70km/h)' during reconstruction.
- Class distribution of adversarial traffic sign ' Speed limit(60km/h)' misclassified as 'Speed limit(50km/h)' during reconstruction.

- **These figures indicate the significant difference of class distribution of classifier between clean and physical adversarial traffic signs during the process of image reconstruction.**

Defense pipeline based on DIP

- Our defense pipeline based on deep image prior method, C is the victim classifier trained on the GTSRB dataset， achieving the best accuracy of 98.70% on the test set.



- **Advantages**

No need for training data

No prior information about attack form

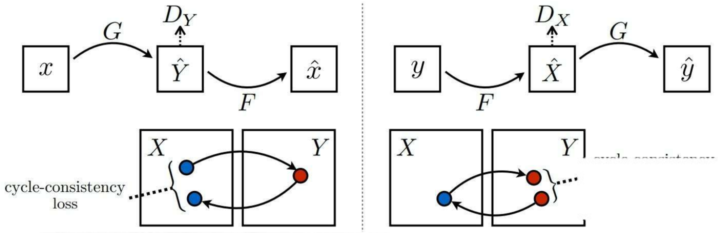
Not limited to specific forms of attacks

Discussion

Defense Results

- **Success rate of correctly classify traffic signs under different defense methods. our defense approach demonstrated better performance against physical adversarial traffic signs.**

CycleGAN [11] :



Defense method /input images	**Physical adversarial traffic signs**	**Clean traffic signs**
Jpeg	0	1.0
CycleGAN	0.21	0.89
Median filter	0.40	1.0
Bilateral filter	0.60	1.0
Our dip-based method	**0.84**	0.97

a_3/a_2	**150**	**175**	**200**	**225**	**250**
400	0.82	0.84	0.84	0.82	0.76
500	0.79	0.81	0.81	0.78	0.72
600	0.78	0.79	0.80	0.77	0.73

Effect of stage division parameters selection on defense success rate.

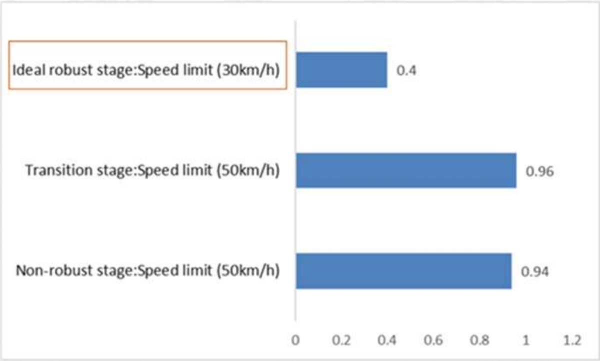
Generality test of our approach

- **Attempts at other types of physical adversarial traffic signs.**

Class before defend: 'Yield'
Defend result: Suspicious sign

Class before defend: 'speed limit 120'
Defend result: Suspicious sign

Class before defend: 'Ahead only'
Defend result: Suspicious sign



Defense on adversarial signs generated based on out-of-distribution attacks and shadow.

Conclusion

Our works

Defense method against physical adversarial traffic signs

- ✓ Based on the inherent priors of traffic signs, we propose an effective defense method for classifiers against physical adversarial traffic signs. This approach is easily deployable and serves to address the existing research gap in physical adversarial defense methods.

Unsupervised defense strategy based on image reconstruction

- ✓ By leveraging the decision consistency of the classifier across different reconstruction stages, our method operates without the need for training data and advanced training.

Conduct extensive testing to assess the generalization capability

- ✓ We conduct extensive testing to assess the generalization capability of our method in handling various types of physical adversarial traffic signs present in real-world scenarios. The results demonstrate that our method exhibits a certain degree of defensive effectiveness against diverse types of physical adversarial traffic signs.

References

[4] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, “Adversarial camouflage: Hiding physical world attacks with natural styles,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1000–1008.

[5] W. Jia, Z. Lu, H. Zhang, Z. Liu, J. Wang, and G. Qu, “Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems,” arXiv preprint arXiv:2201.06192, 2022.

[6] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, “Perceptual-sensitive gan for generating adversarial patches,” in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp.1028–1035.

[7] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, “Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15345–15354.

[8] Giulivi, L., Jere, M., Rossi, L., Koushanfar, F., Ciocarlie, G., Hitaj, B., & Boracchi, G. (2023). Adversarial scratches: Deployable attacks to CNN classifiers. Pattern Recognition, 133, 108985.

[9] Sitawarin C, Bhagoji A N, Mosenia A, et al. Darts: Deceiving autonomous cars with toxic signs [J]. arXiv preprint arXiv:1802.06430, 2018.

[10] Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky. "Deep image prior." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[11] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

同濟大學
TONGJI UNIVERSITY

THNS 2024, November 5-7, 2024

AI & Road Flow

Thank you for your attention!

Trusted Perception Method for Traffic Signs That Are Physically Attacked

Speaker: Qunyao Tan   Email : tqyao@tongji.edu.cn

You can scan the code through Wechat.

We will post team updates in time.

We wholeheartedly welcome the exchange with you.