Adversarial example detection Bayesian game

H. Zeng, B. Chen, K. Deng, A. Peng

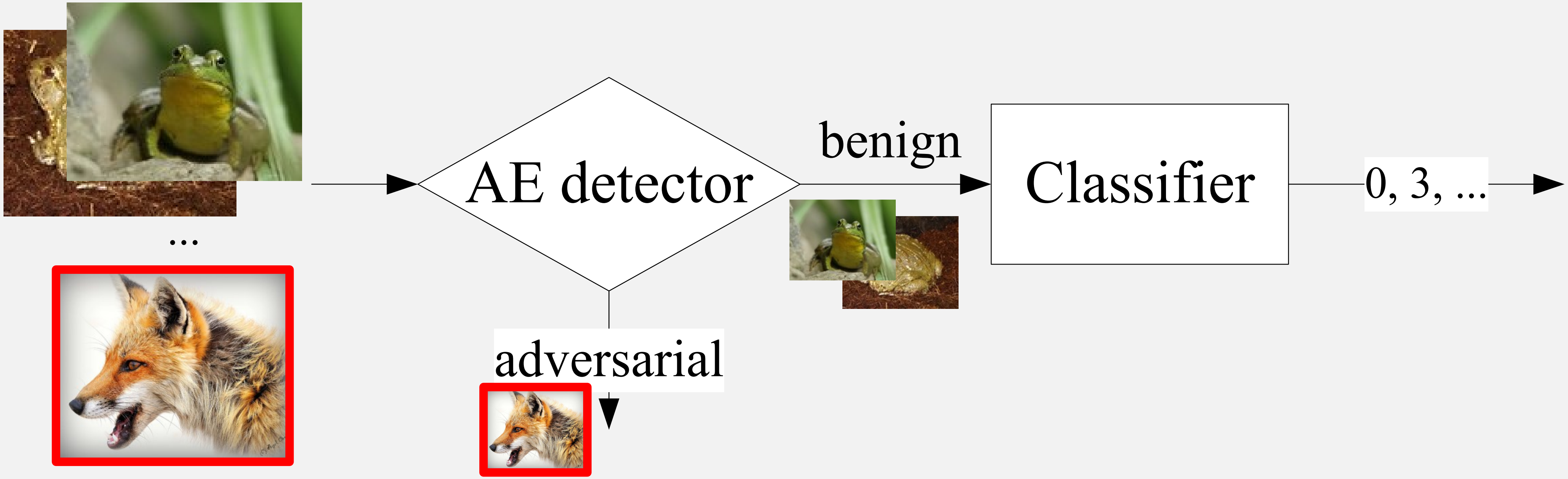
ICIP2023, Kuala Lumpur 2023.10

Contents

1 Motivation

Motivation

A popular defense strategy against adversarial examples (AE) is detect-then-reject.

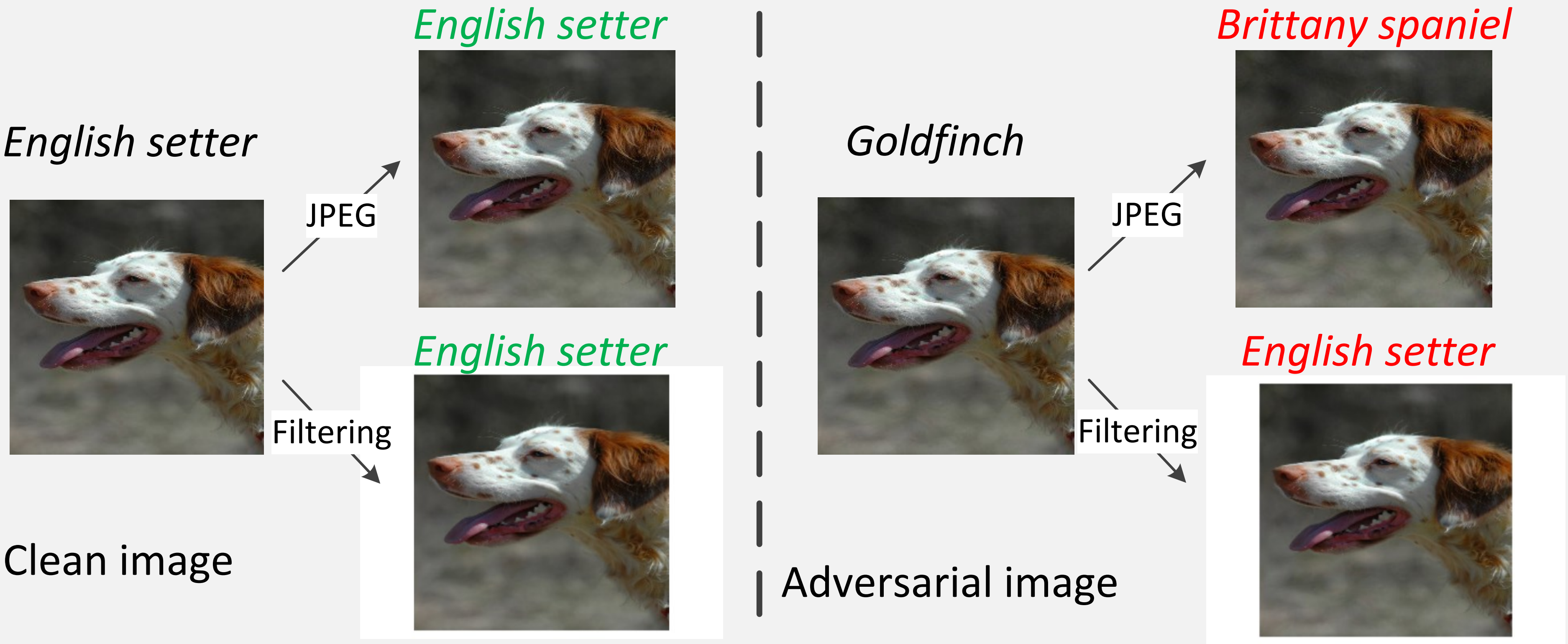


Motivation

Existing detectors are based on the following two assumptions about AE:

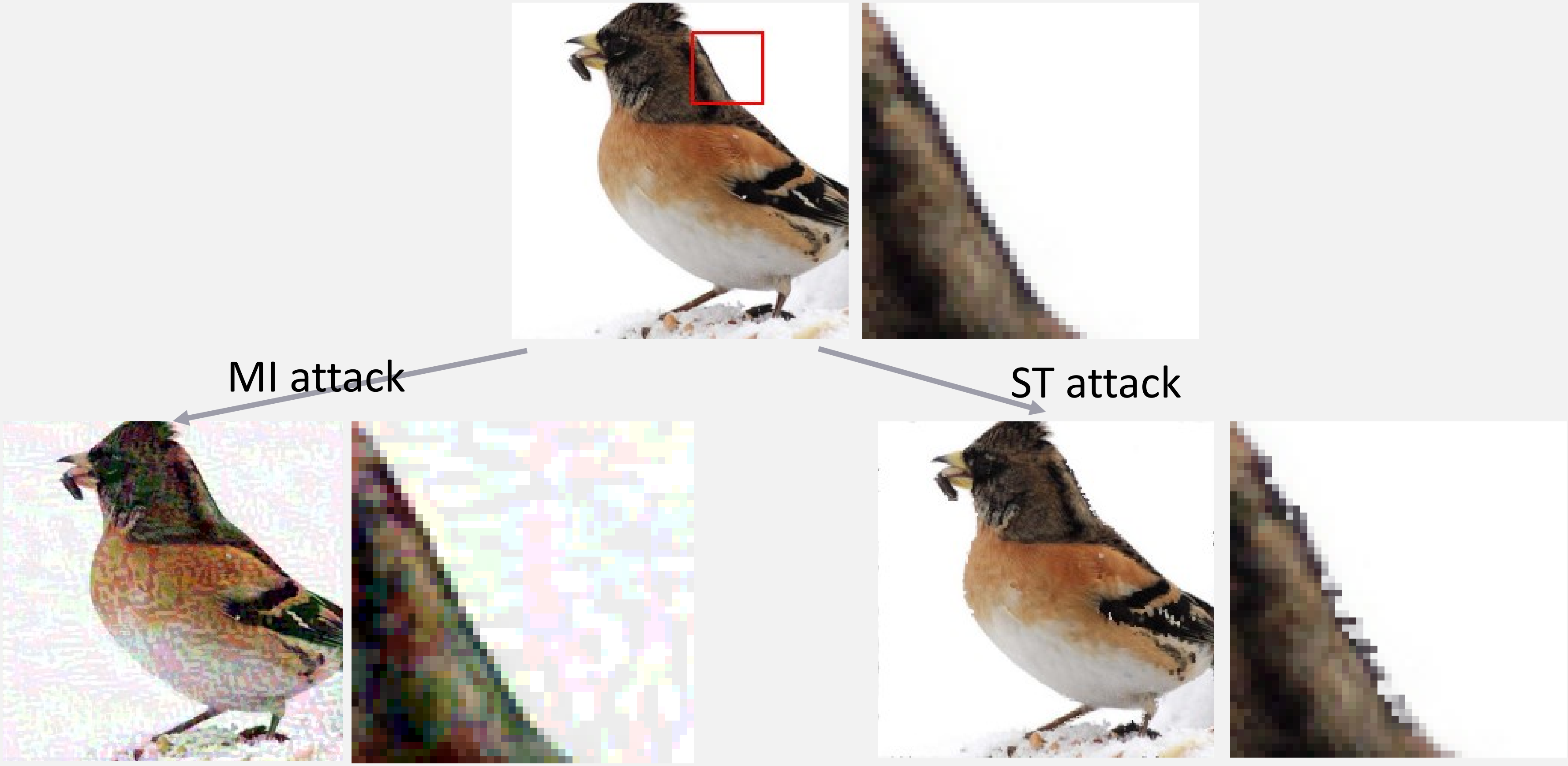
1) Compared to natural images, AEs are more sensitive to disturbance:

$$F(I') \neq F(P(I'))$$



Motivation

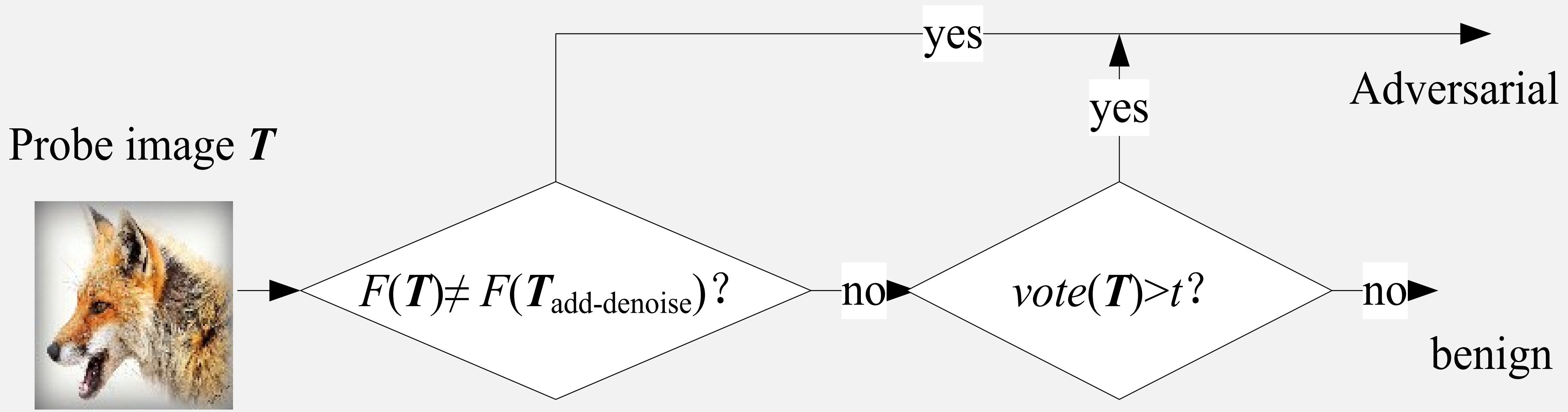
2) Adversarial perturbation disrupts local dependences of natural images.

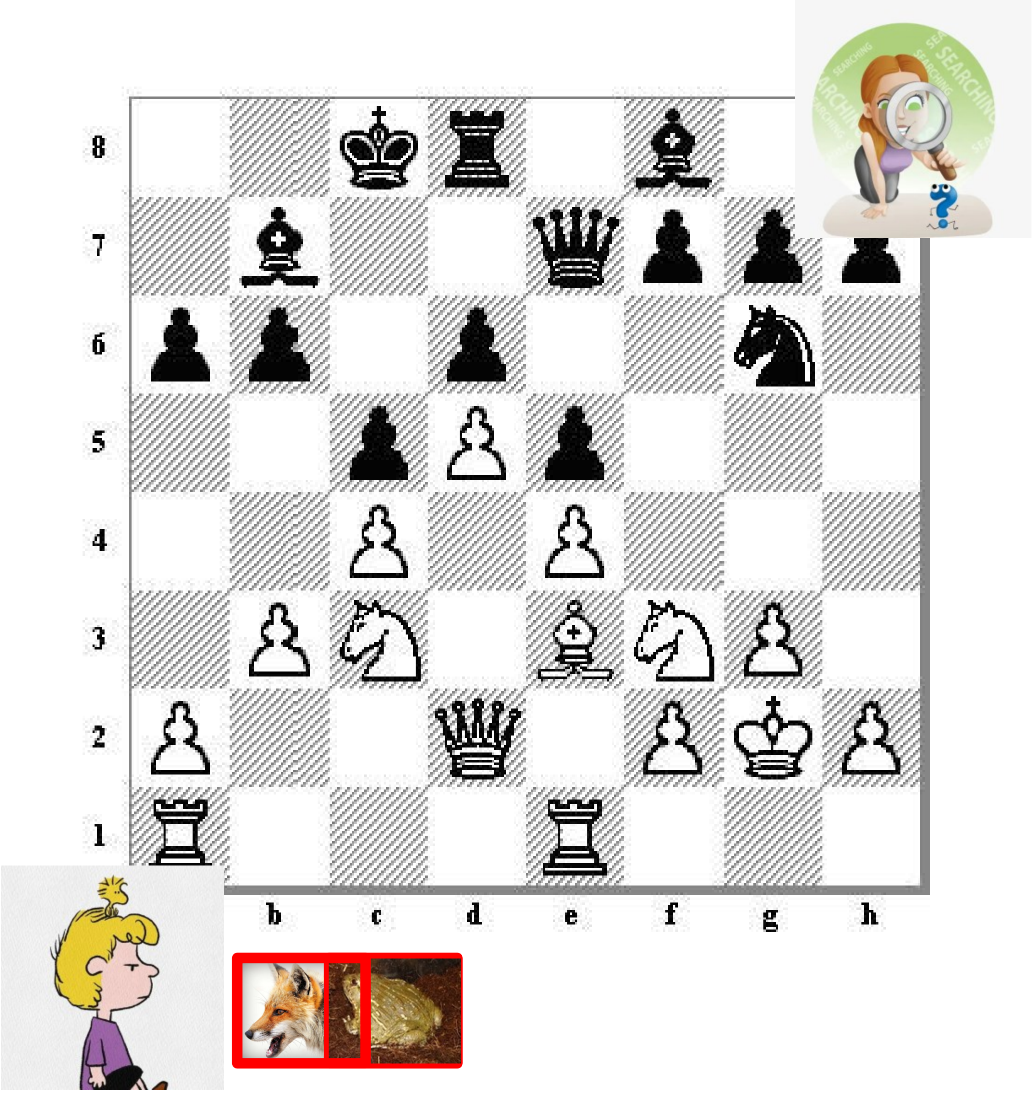


Motivation

The detectors based on these two assumptions are **complementary:**

The first type detectors are good at revealing AEs of weak strength, whereas the second type detectors are suitable for detecting AEs of large budget.



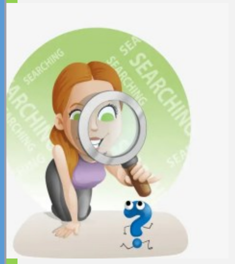


2 Game model

Game model

The players’ knowledge

1. The detectors adopted by the investigator;
2. The attack methods available to the attacker, and prior belief of them;
3. The investigator’s strategy space;
4. The attacker’s strategy space;
5. The payoff matrix.



The exact attack adopted.

Game model

Definition: AE-detection $(S_I, S_A, \Omega, \boldsymbol{p}, \boldsymbol{U})$ game is a zero sum, incomplete information game played by the investigator and the attacker, featured by the following strategies and payoff:

1) S_I: The investigator's strategy space, i.e., P_{fa}^1 that can be allocated to $\delta^1()$.
2) S_A: The attacker's strategy space, i.e., the attacking strength r in generating AEs.
3) Ω: The set of attack methods.
4) $\boldsymbol{p}$: The prior belief about the probability measure of Ω. $\boldsymbol{p} = [p_1, p_2, \cdots, p_N]$ and $\sum_l p_l = 1$.
5) $\boldsymbol{U}$: The payoff matrix, which is defined as the total detection rate of the two-step test: $U(P_{fa}^1, r) = P_d(P_{fa}^1, r)$

3 Results

Results

Experimental settings

Classification model: a pre-trained ResNet18 model

Dataset: 10000 images from ImageNet validation dataset.

Training: 7000, Testing: 3000.

Attacks: IFGSM, MI, $\epsilon \in \{1, 2, 4, 6, 8\}$

C&W, and Spatially transformed (ST), $k \in \{0, 5, 10, 15, 20\}$

Defense: $\delta^1()$-Noise addition-then-denoising test [1].

$\delta^2()$-SRM-based test [2].

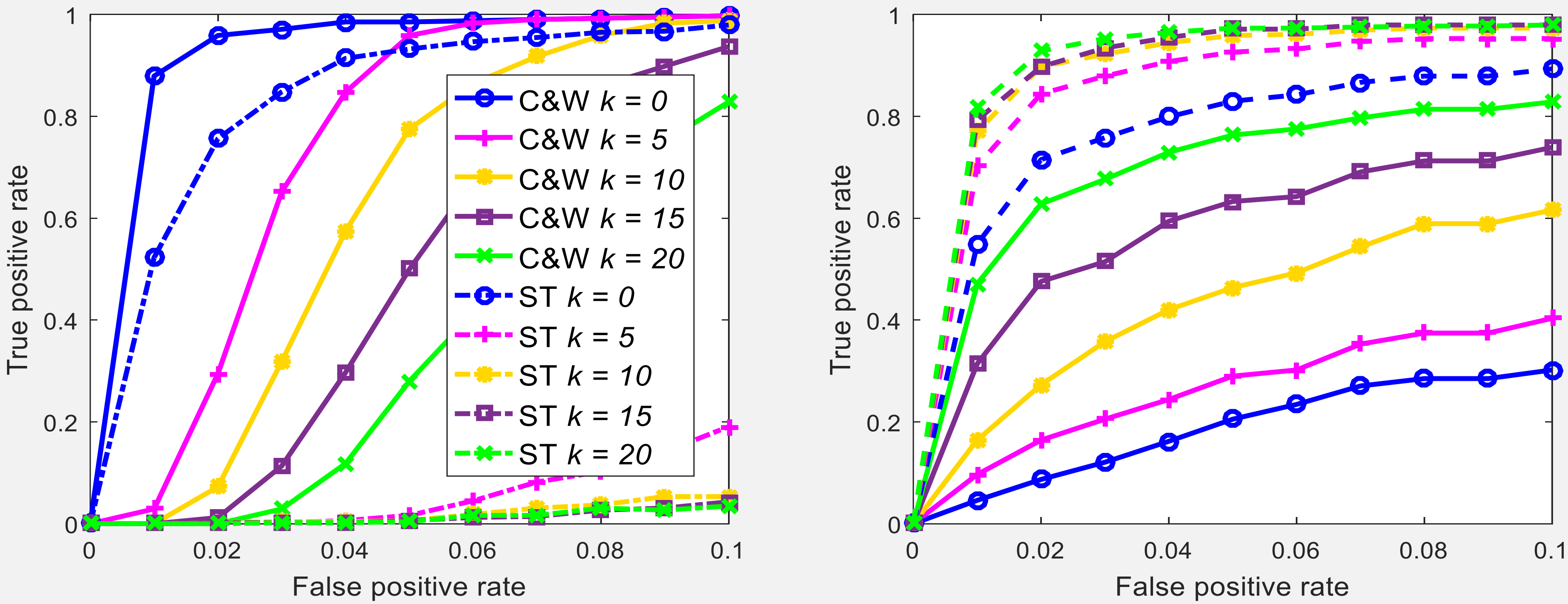
[1] K. Deng, A. Peng, W. Dong, H. Zeng, "Detecting C&W adversarial images based on noise addition-then-denoising," ICIP2021, pp. 3607–3611.

[2] J. Liu, W. Zhang, Y. Zhang, et al., "Detection based defense against adversarial examples from the steganalysis point of view," CVPR2019, pp. 4820–4829

Results

ROCs of the two single tests

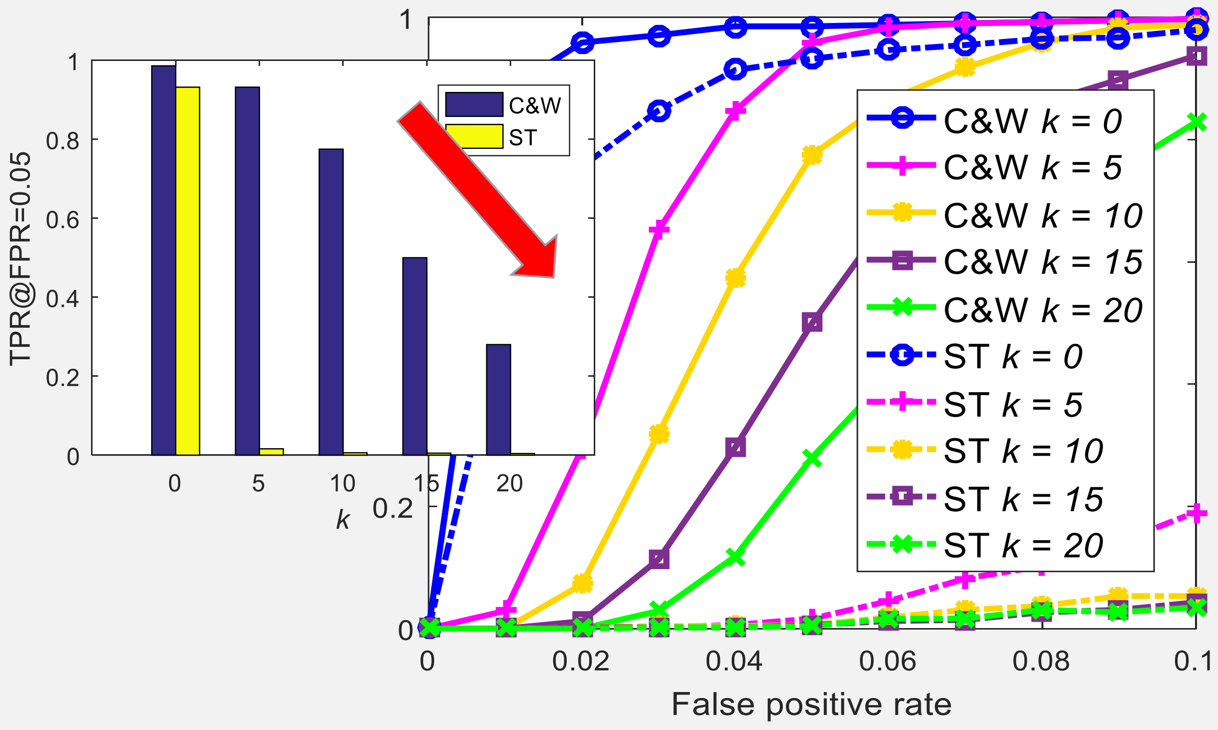
A strong complementarity between δ^1(Left) and δ^2(Right)

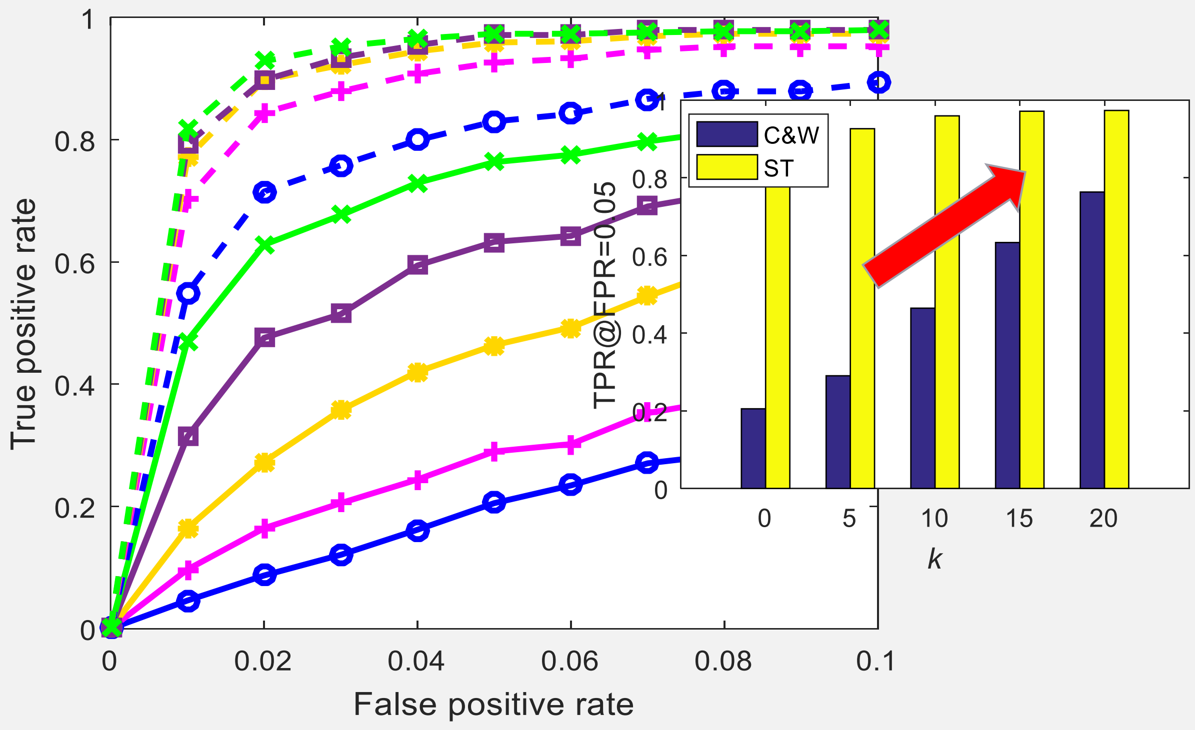


Results

ROCs of the two single tests

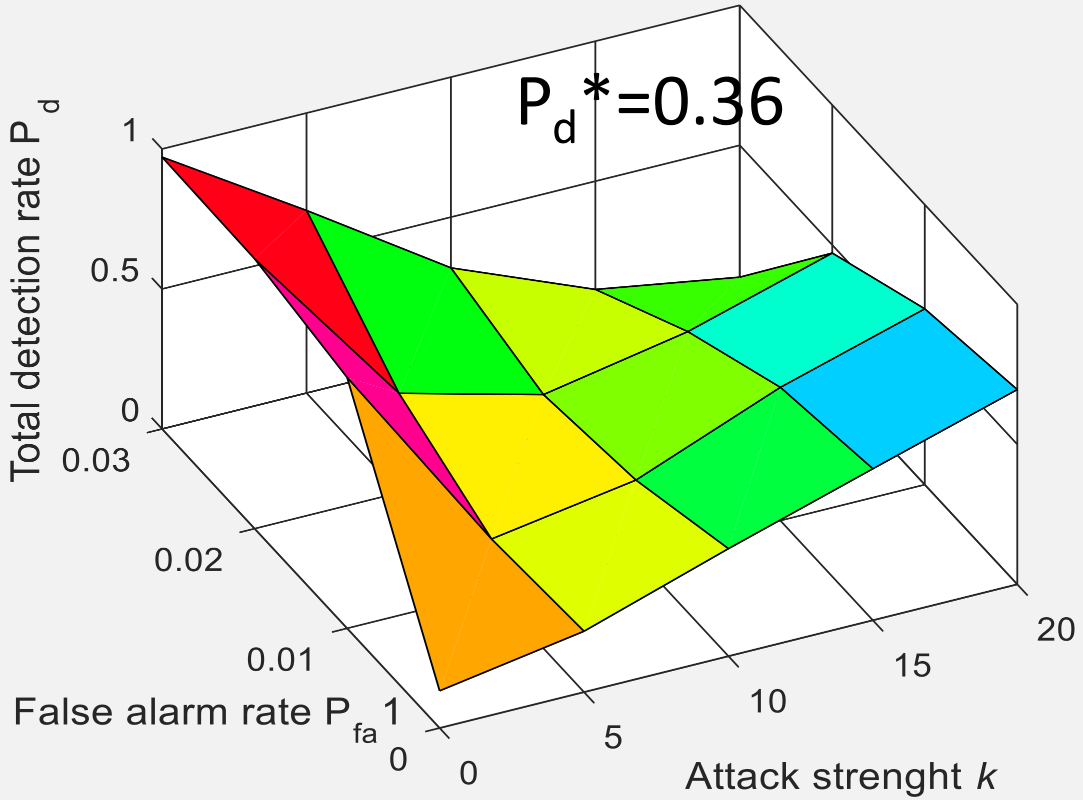
A strong complementarity between δ^1(Left) and δ^2(Right)

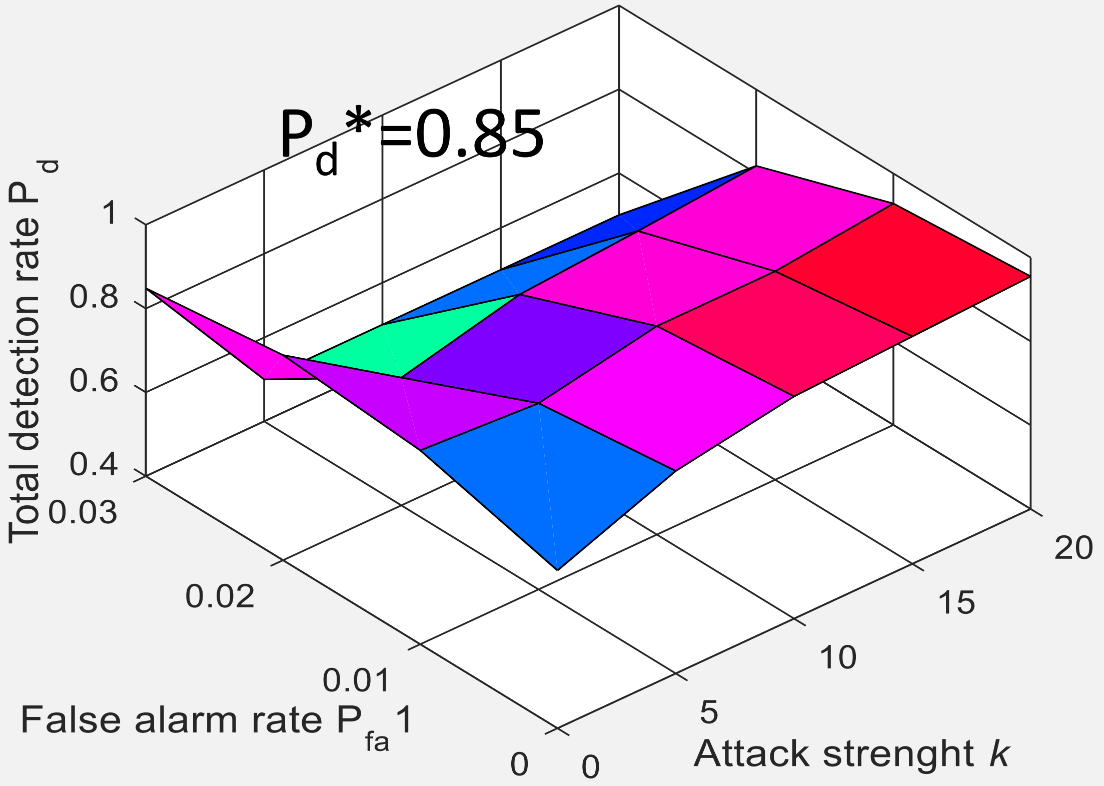




Results

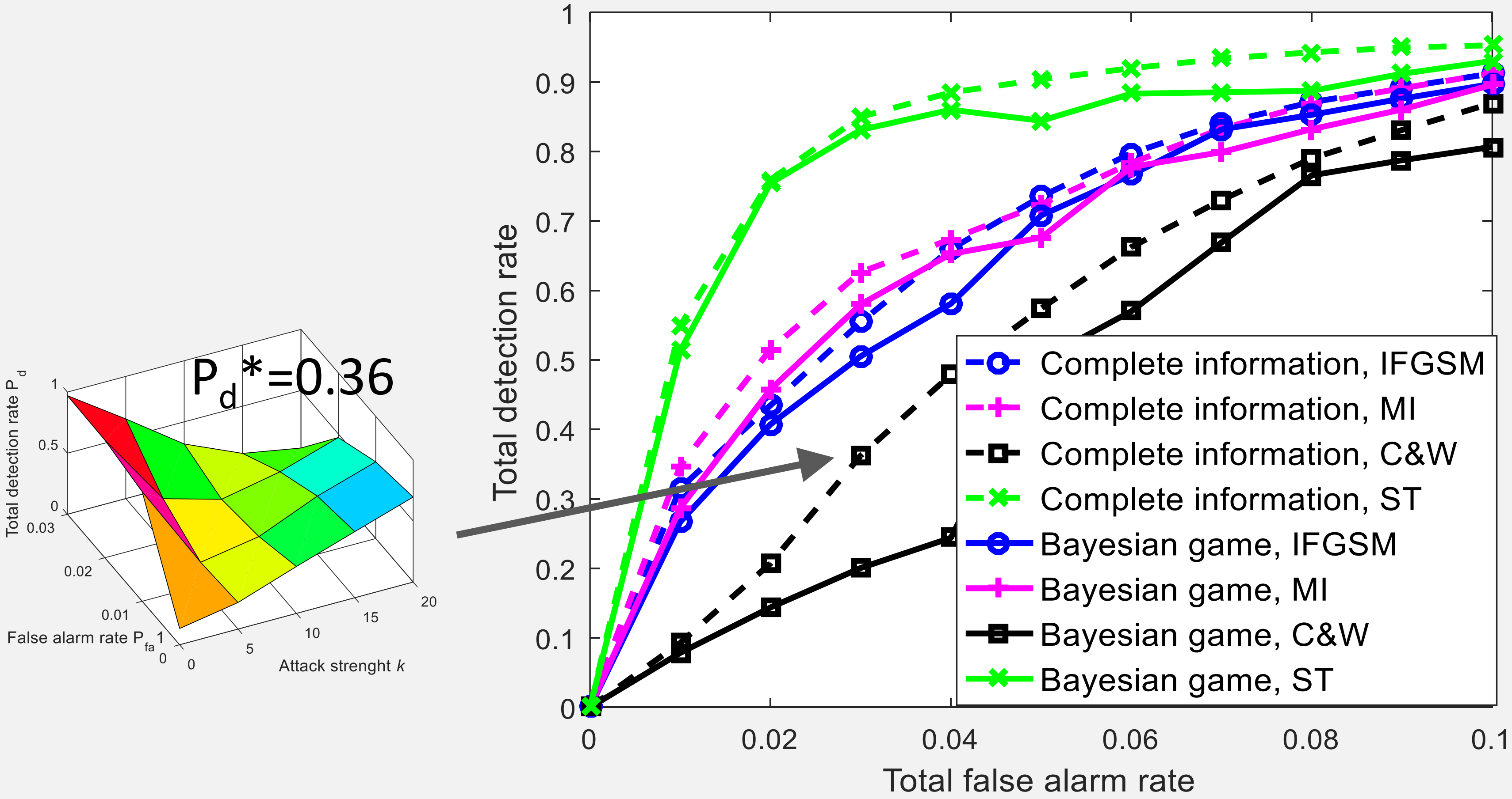
P_d matrix of the two-step test at $P_{fa} = 0.03$. (L) C&W, (R) ST.





Results

Nash equilibrium ROCs, $p = [0.25, 0.25, 0.25, 0.25]$ for the Bayesian game.



4 Summarization

Summarization

1) Game theory is used to model the interplay between AE generation and detection. Under this framework, we can compare the security of different attacks in a more systematic way.

2) Bayesian game is used to model the information asymmetry in this interplay, which makes our analysis more realistic .

Thanks for attention

Codes: *https://github.com/zengh5/AED_BGame*