USENIX Security ‘24

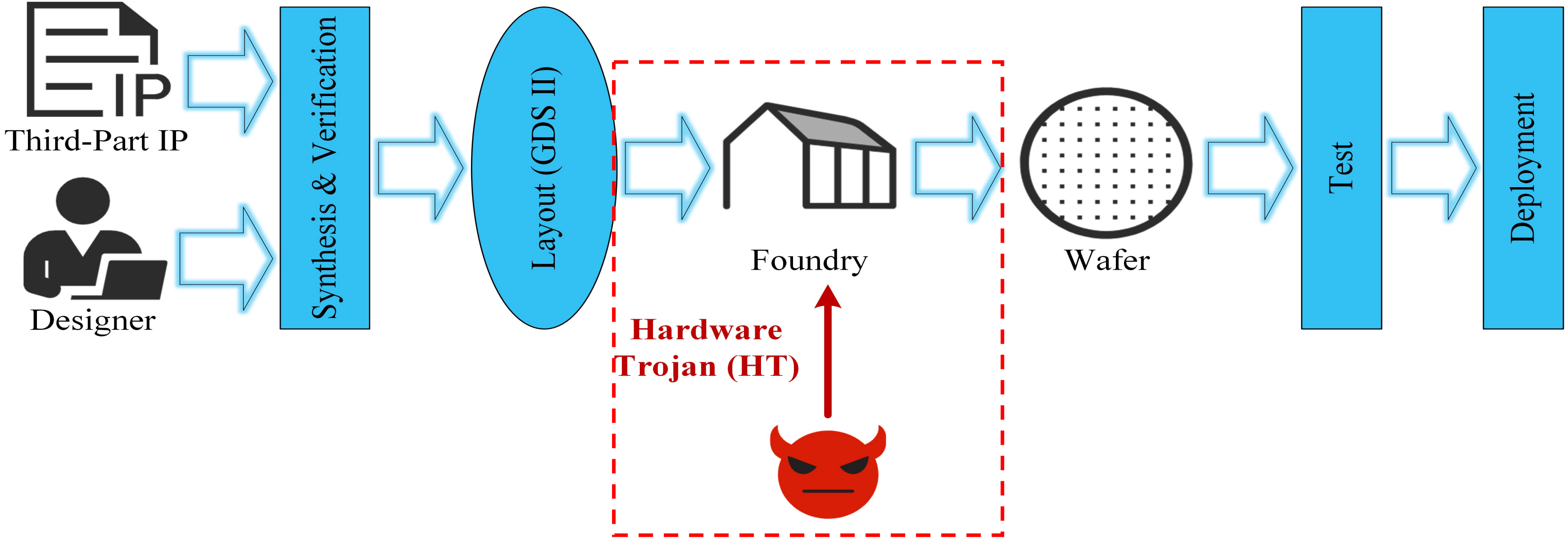
Improving the Ability of Thermal Radiation Based Hardware Trojan Detection

Ting Su, Yaohua Wang*, Shi Xu, Lusi Zhang, Simin Feng, Jialong Song, Yiming Liu
Yongkang Tang, Yang Zhang, Shaoqing Li, Yang Guo, Hengzhu Liu

National University of Defense Technology

Outlines

Hardware Trojan Threat



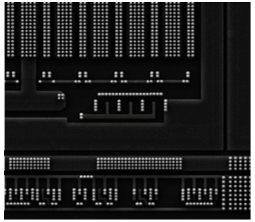
[1] Kaiyuan Yang, Matthew Hicks, Qing Dong, Todd Austin, and Dennis Sylvester. A2: Analog Malicious Hardware. In 2016 IEEE Symposium on Security and Privacy (SP), pages 18–37, San Jose, CA, 2016. IEEE.

[2] Timothy Trippel, Kang G. Shin, Kevin B. Bush, and Matthew Hicks. ICAS: An Extensible Framework for Estimating the Susceptibility of IC Layouts to Additive Trojans. In 2020 IEEE Symposium on Security and Privacy (SP), pages 1742–1759, San Francisco, CA, USA, 2020. IEEE.

[3] Tiago D. Perez and Samuel Pagliarini. Hardware Trojan Insertion in Finalized Layouts: From Methodology to a Silicon Demonstration. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42(7):2094–2107, 2023.

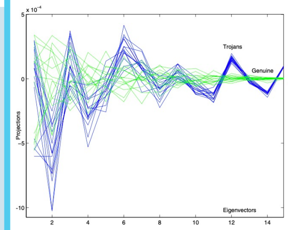
HT Detection Methods

Side-channel Analysis Techniques

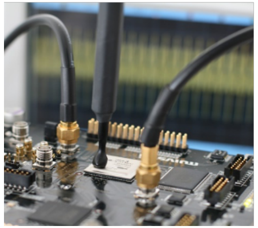


Reverse Engineering

High Cost
Complex Process



Power

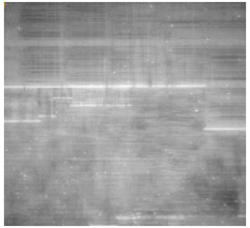


Electromagnetic Radiation

Requiring the **golden chip or testing vectors**
Limited by **IC size, process variation and noises**

Optical Light

Weak Penetration



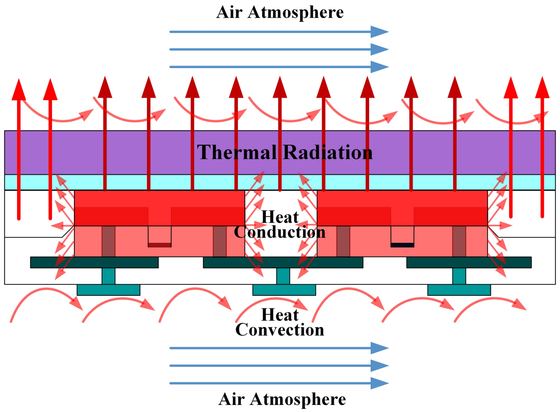
Thermal Radiation

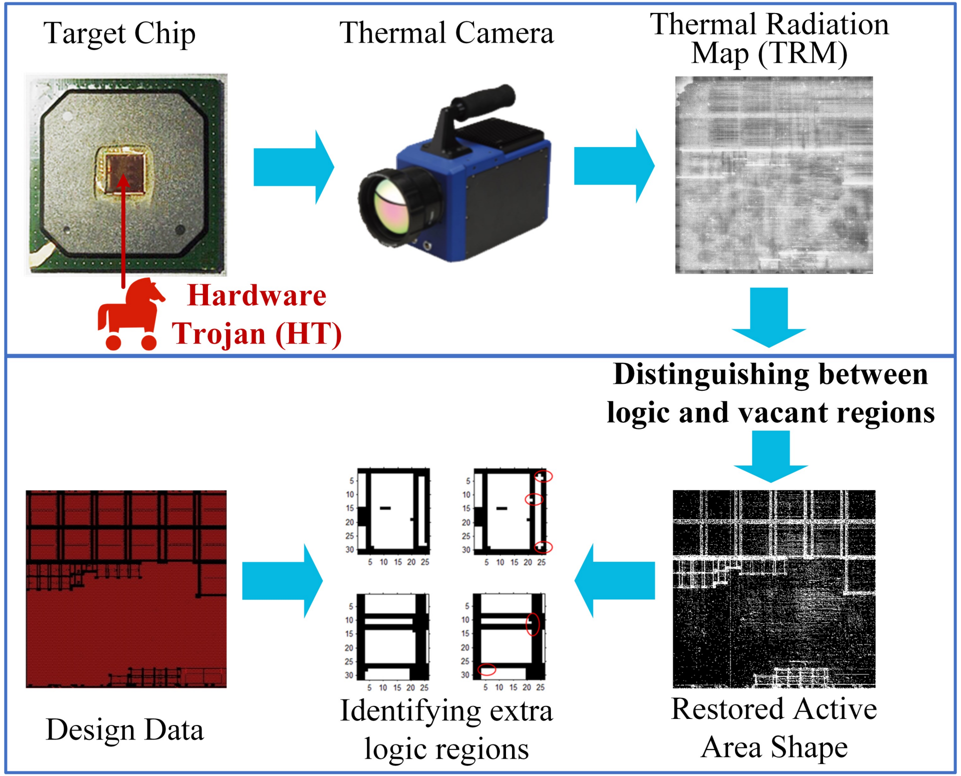
Good penetration
Pixel-level Resolution

Thermal Radiation (TR) Based Detection

➢ Advantages

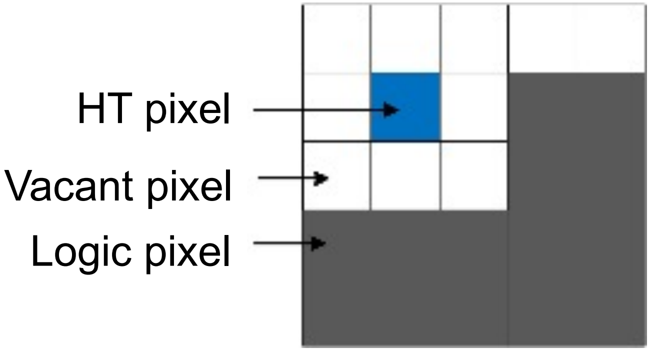
- High detection resolution
- Process variation resistant
- Adaptability for large ICs
- Golden chip free
- HT activation free



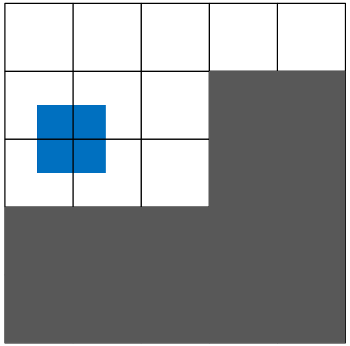


Previous TR-based Methods

- Nazma et al. [TCAD-2014]: Shows promising detection ability, but relies on **stronger simulation tools**
- Tang et al. [TVLSI-2019]: Can only identify **the ideal HT** that fully occupies at least one pixel on the TRM



The ideal HT



The HT spreads into multiple pixels

Outlines

Ⅰ Background

Ⅱ Motivation

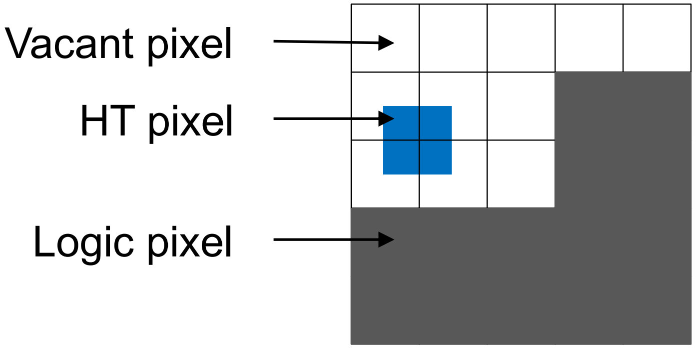
Ⅲ NICE Mechanism

Ⅳ Experiment & Conclusion

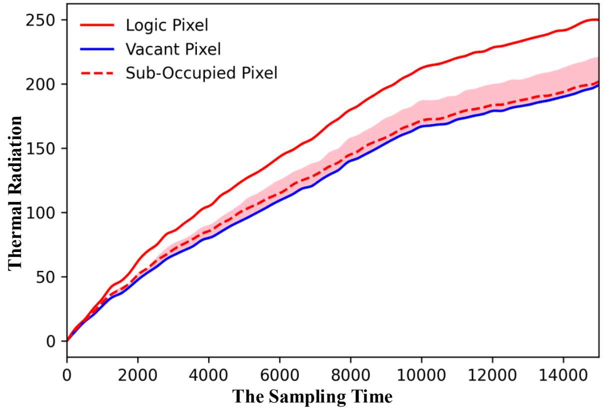
Observation

➢ Sub-pixel HT

- We can not ensure precise alignment of the HT boundaries with the pixels
- Each infected pixel is easily blurred as either a logic or vacant area



How to detect them



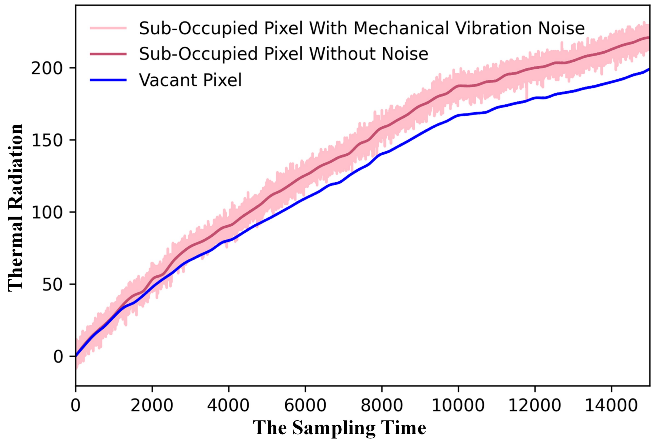
Observation

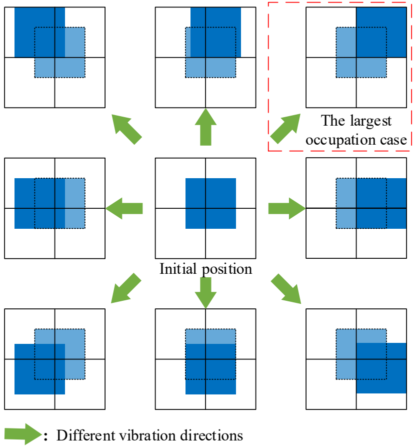
➢ Two sides of mechanical vibration

Cons: It complicates the TR distinction between sub-occupied and vacant pixels

Pros: It **can vary the pixel occupation** of HTs

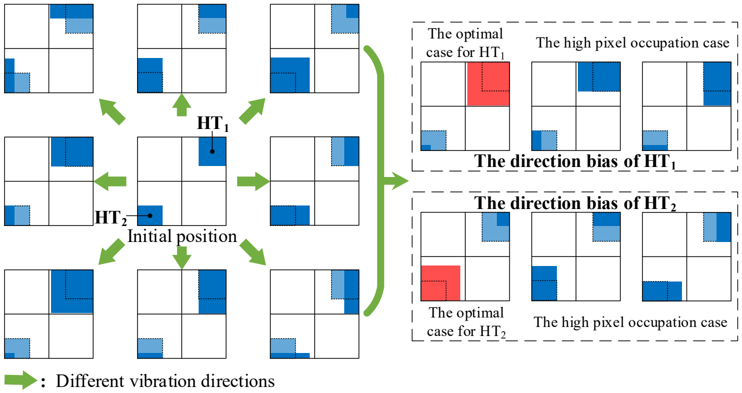
Mechanical vibration from thermal cameras





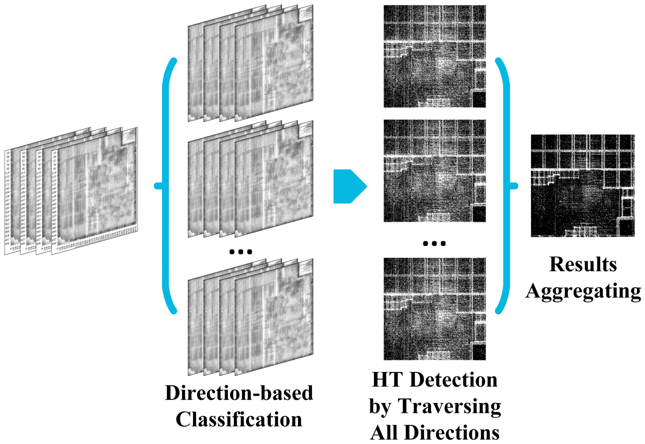
Our Goals

- We want to find out the vibration direction that can enhance the TR distinction, thereby effectively detecting sub-pixel HT



Single direction cannot uniformly optimize detection across all HTs

Transforming The Problem

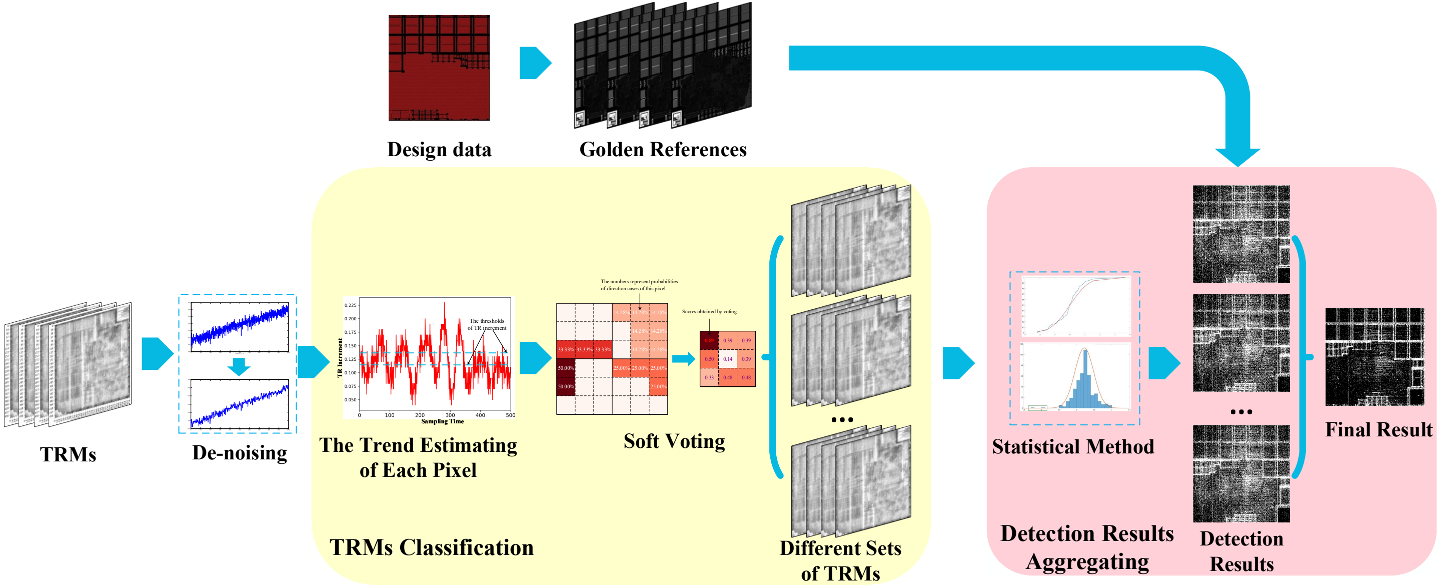


Detecting potential HTs by traversing all directions

Outlines

Overview

Direction-based TRMs Classification → HT detection by traversing all directions → Results aggregation for possible HT pixels

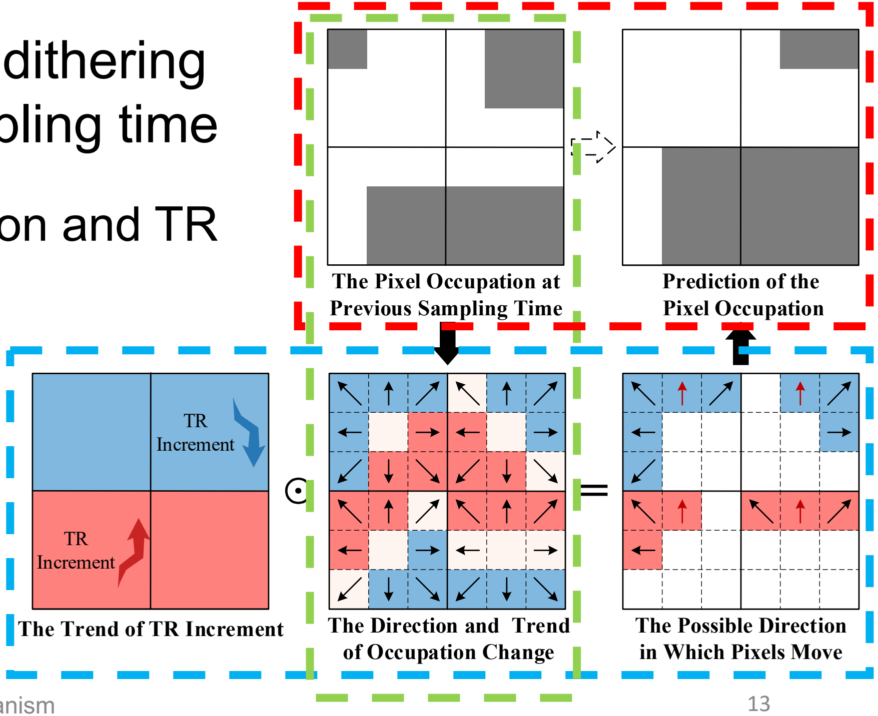


Noise Based Pixel Occupation Enhancement (NICE)

Direction-based TRMs Classification

Direction-based TRMs Classification → HT detection by traversing all directions → Results aggregation for possible HT pixels

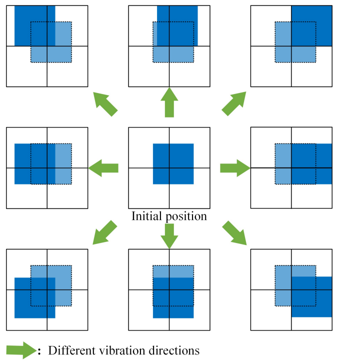
- This procedure entails identifying the dithering direction within the pixel at each sampling time
  - The correlation between pixel occupation and TR increment
  - The convergence of all pixels dithering



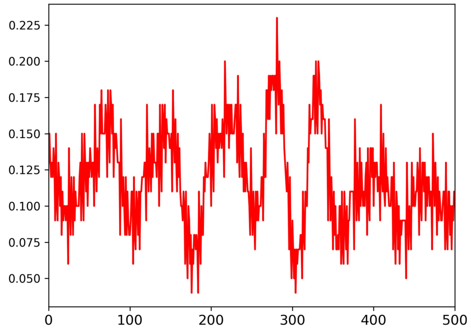
Estimating Possible Directions for Each Pixel

Direction-based TRMs Classification → HT detection by traversing all directions → Results aggregation for possible HT pixels

- **STEP Ⅰ**: Formulated a linear regression model
  - Pixel Occupation $\boldsymbol{X_{pixel}}$: Calculated from IC layout containing occupation information for each pixel
  - TR Increment Data $\Delta \boldsymbol{I}$: Extracted from TRM sequences



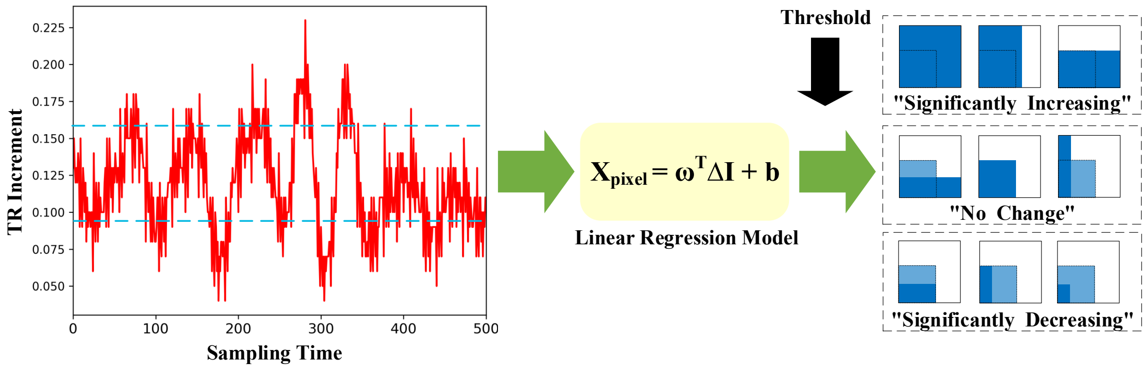
$$\mathbf{X}_{\mathbf{pixel}} = \omega^{T} \Delta \mathbf{I} + b$$



Estimating Possible Directions for Each Pixel

Direction-based TRMs Classification → HT detection by traversing all directions → Results aggregation for possible HT pixels

- **STEP Ⅱ:** Determining trends of pixel occupation over time
  - INPUT: TR Increment Data ΔI of each pixel
  - OUTPUT: Determined pixel occupation X'_{pixel} at every sampling time
- **STEP Ⅲ**: Estimating possible dithering directions



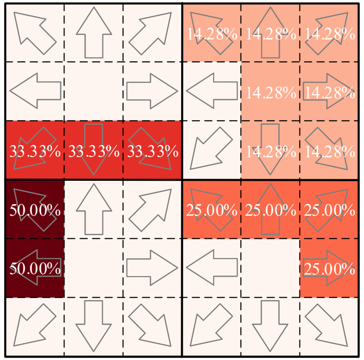
Classifying TRMs into Different Direction Sets

Direction-based TRMs Classification → HT detection by traversing all directions → Results aggregation for possible HT pixels

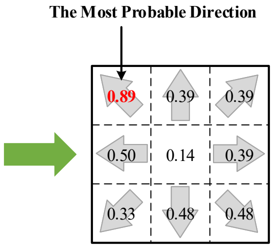
- Calculating the probabilities p_{ij}^{dk} of possible directions dk of each pixel ij
- Determining the most probable direction $Prob_{max}$ through a weighted average

$$Prob_{max} = \max_{1 \le k \le n} \{ \sum_{i=1}^{M} \sum_{j=1}^{N} p_{ij}^{dk} \}$$

$$p_{ij}^{dk} = \begin{cases} 0 & , dk \in possible\ directions \\ \frac{1}{number\ of\ possible\ directions} & , dk \notin possible\ directions \end{cases}$$



The Possible Directions of Four Pixels

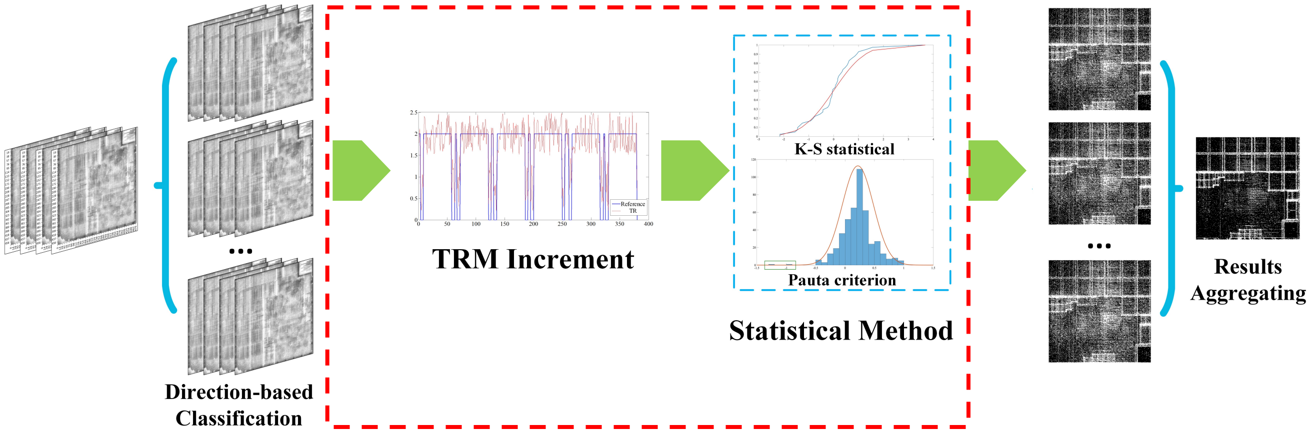


Results of Soft Voting

HT Detecting and Results Aggregating

Direction-based TRMs Classification → **HT detection by traversing all directions** → Results aggregation for possible HT pixels

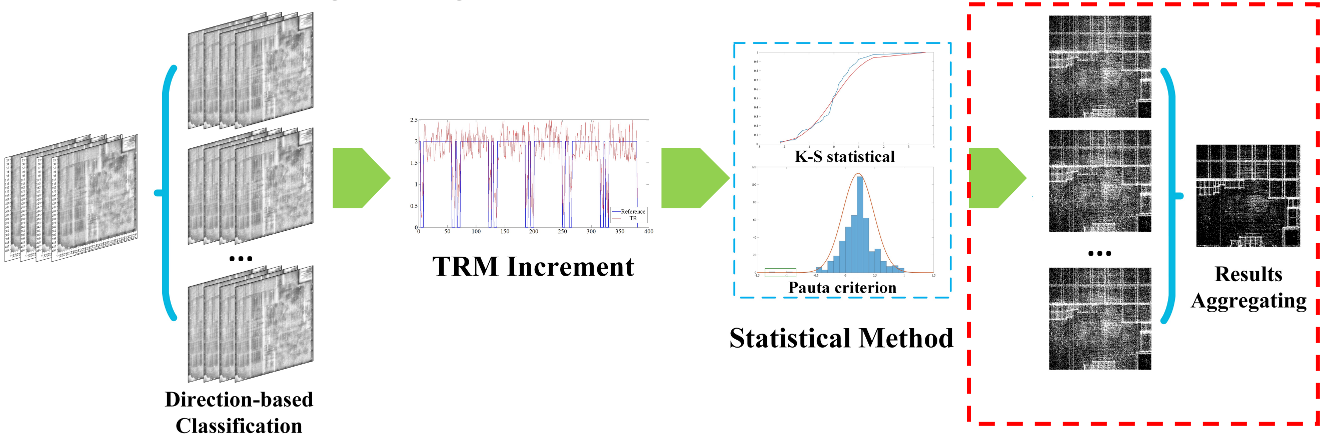
- The TRMs set in each direction is processed to distinguish between logic and vacant regions through the K-S statistic and the Pauta criterion
- Comparing with the golden references, extra HT pixels can be identified



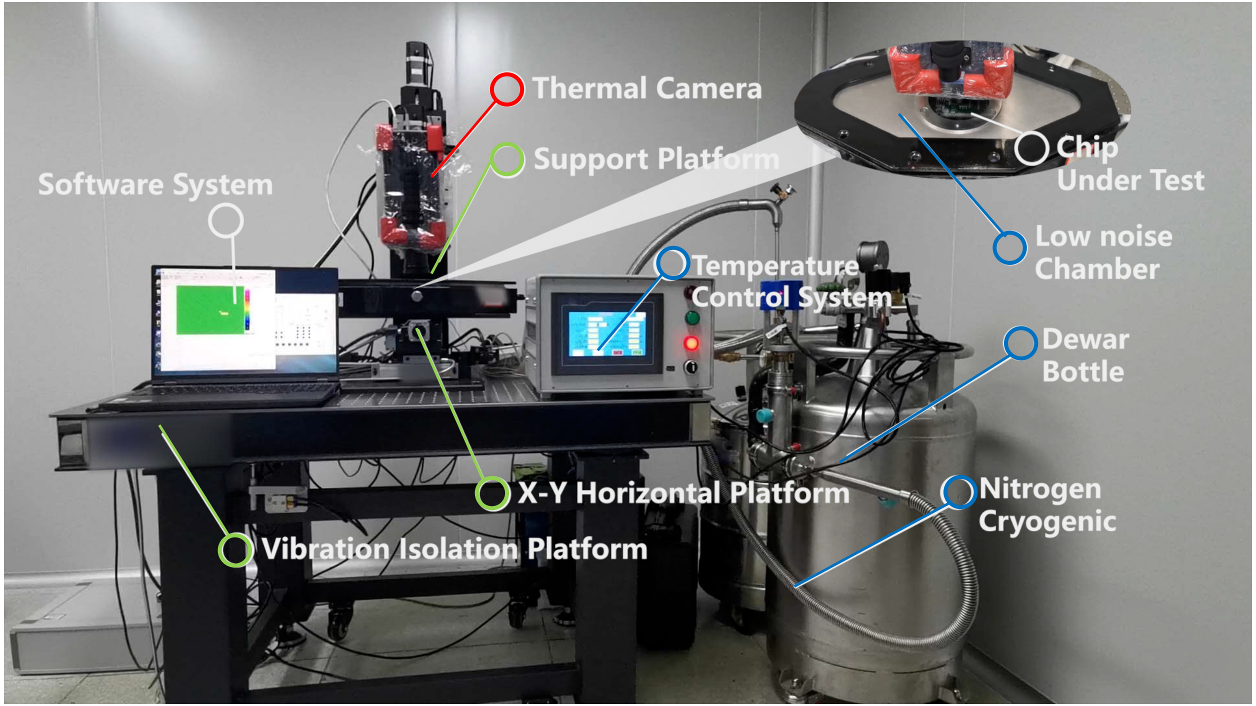
HT Detecting and Results Aggregating

Direction-based TRMs Classification ⇨ HT detection by traversing all directions ⇨ **Results aggregation for possible HT pixels**

- Typically, any extra logic pixels detected in any directions should be considered as HTs
- In particular, the result need to be corrected, when extra logic pixels corresponds to logic regions in most references in other directions



NICE System Implementation

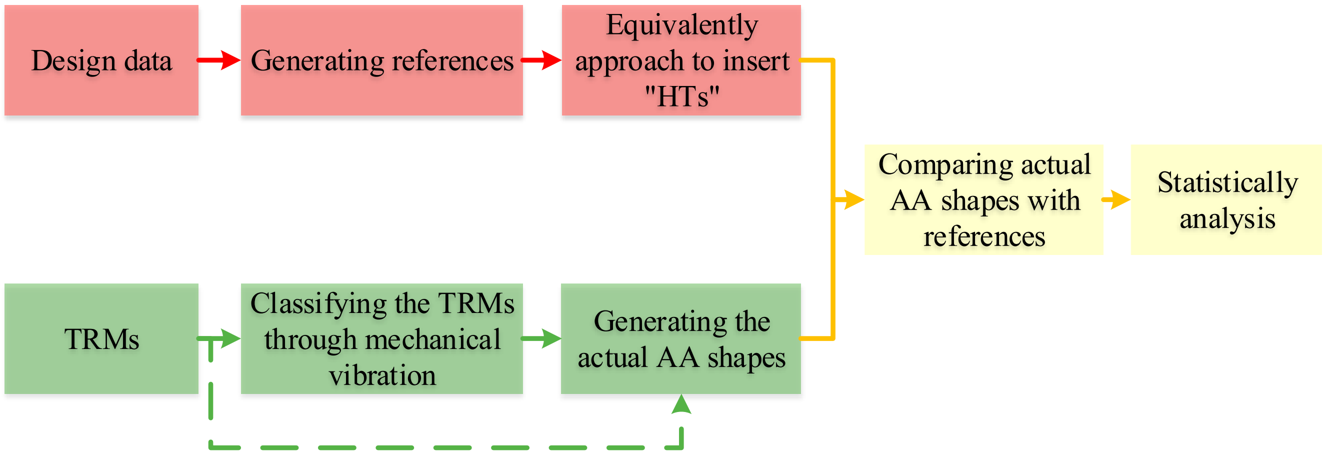


Outlines

- Ⅰ Background
- Ⅱ Motivation
- Ⅲ NICE Mechanism
- **Ⅳ Experiment & Conclusion**

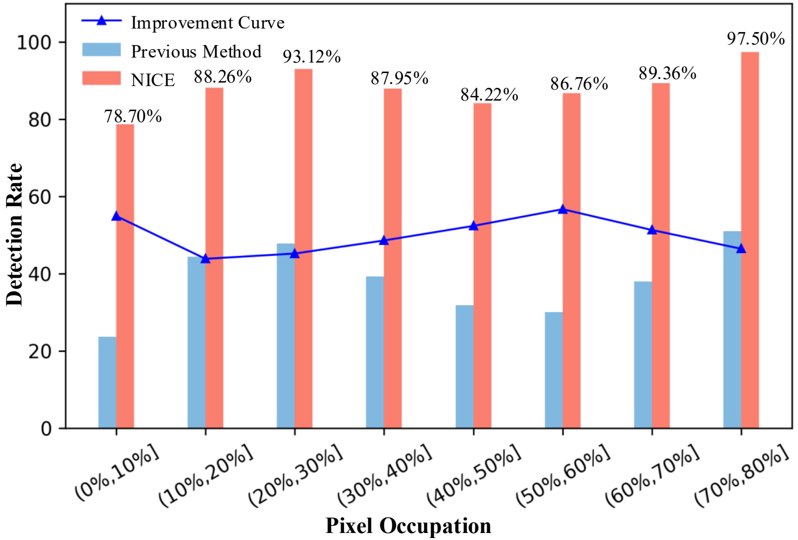
Experiment Scheme

- The equivalently approach is employed to implement "HT"



Experiment for Sub-pixel HT Detection

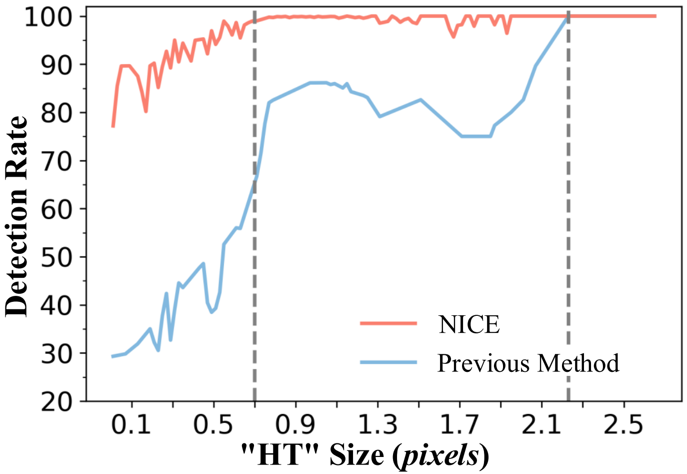
- NICE can detect sub-pixel HTs with a **detection rate** of up to **91.82%** and a false alarm rate below 9%, representing a **performance improvement** of more than **47%** over the previous method



	Previous method	NICE (single set)	NICE (final result)						
			Thresholds:	1%	2.5%	5%	10%	15%	20%
Detection rate	44.36%	67.48%		45.26%	87.77%	91.81%	90.53%	83.20%	84.30%
False alarm rate	15.90%	13.42%		12.56%	16.18%	8.44%	9.85%	13.13%	10.09%

Performance Across Different HTs

- NICE can push the detection boundary of TR-based methods from **more than two pixels** to only **0.7 pixels**

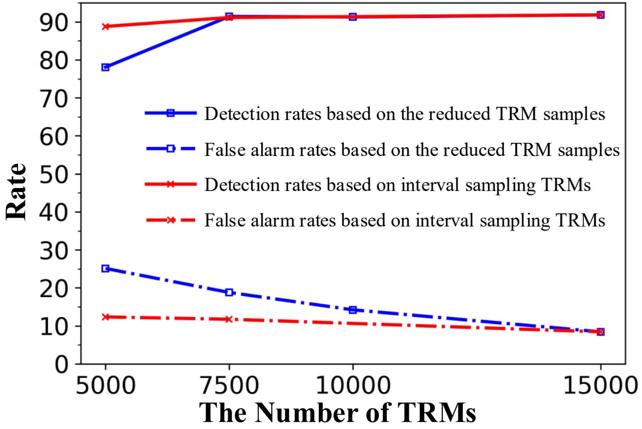


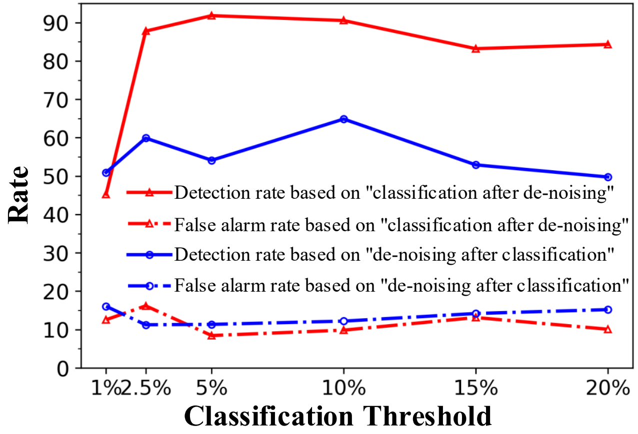
Camera resolution	Chip technology	Number of equivalent gates	Detection boundary Previous	Detection boundary NICE
15μm*15μm	130nm	≈ 55	⩾110	⩾39 ✓
	65nm	≈ 117	⩾234	⩾82 ✓
	40nm	≈ 239	⩾478	⩾167
	28nm	≈ 446	⩾892	⩾312
	14nm	≈ 868	⩾1736	⩾608
12μm*12μm	130nm	≈ 36	⩾72 ✓	⩾25 ✓
	65nm	≈ 75	⩾150	⩾52 ✓
	40nm	≈ 153	⩾306	⩾107 ✓
	28nm	≈ 286	⩾572	⩾200
	14nm	≈ 556	⩾1112	⩾389
5μm*5μm	130nm	≈ 6	⩾12 ✓	⩾4 ✓
	65nm	≈ 13	⩾26 ✓	⩾9 ✓
	40nm	≈ 27	⩾54 ✓	⩾19 ✓
	28nm	≈ 50	⩾100	⩾35 ✓
	14nm	≈ 96	⩾192	⩾67 ✓

Works	HTs	Number of equivalent gates
Siddik et al. [8]	PUF-based HT	≈ 125
Deng et al. [13]	A2(130nm)	≈ 43
Dharsee et al. [14]	Jinn	125
Jain et al. [23]	TAAL(32nm)	≈ 189
Kumar et al. [25]	edAttack(15nm)	≈ 37
Lin et al. [26]	TSC	⩽ 100
Trippel et al. [46]	A2(45nm)	91
	Key Leak(45nm)	187
Trust-Hub	AES-T400	≈ 90
	AES-T600	≈ 100
	AES-T700	≈ 80
	AES-T800	≈ 230
	AES-T900	≈ 840
	AES-T1000	≈ 80
	AES-T1100	≈ 80
	AES-T1200	≈ 840
	AES-T2000	≈ 80

Sensitivity Analysis

- **Number of TRMs:** NICE can achieve steady performance, even when the number of samples is decreased to 50%
- **Classification Thresholds:** NICE is robust enough for different thresholds
- **White Noise:** NICE also outperforms previous methods, as the effects of classification thresholds and white noise are combined





Conclusion

- A novel method exploiting **the potential of noise** for TR-based HT detection
- It can detect sub-pixel HTs with **high performance**, without needing a **golden chip** and special **test vectors**
- It can enable a **more flexible** and **cost-effective** selection of thermal cameras for TR-based HT detection

USENIX Security ‘24

Improving the Ability of Thermal Radiation Based Hardware Trojan Detection

Thank you for your time and attention!

Contact information:

suting@nudt.edu.cn, yaowangeth@gmail.com