RD-PHash: A Robustness Enhancement for DCT-Based Perceptual Hashing Against Adversarial Bit-Flipping Attacks

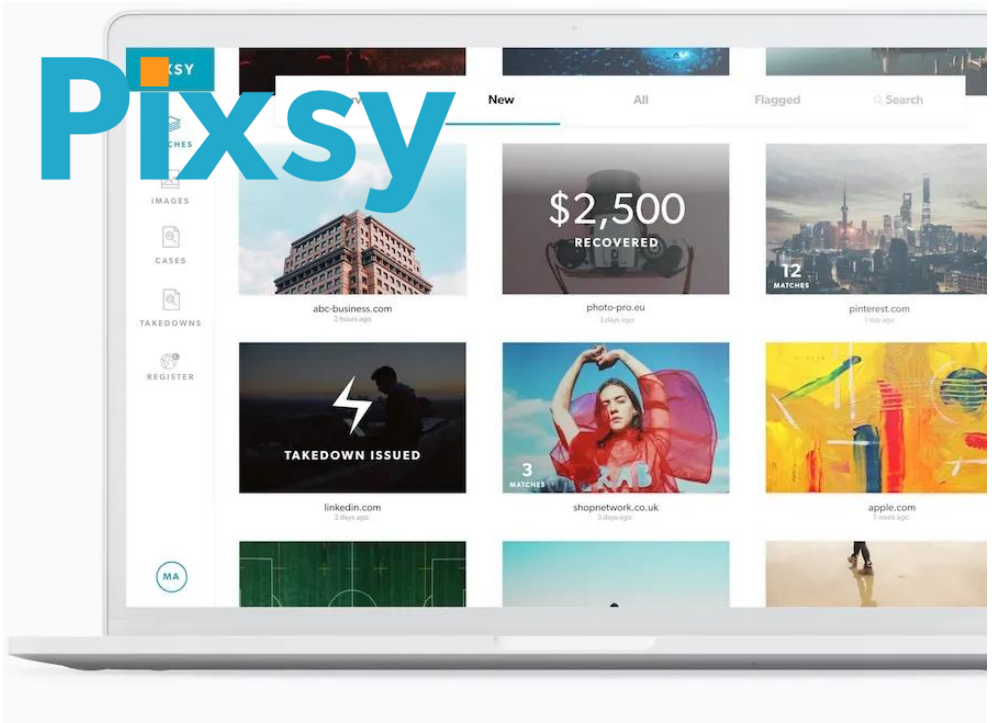
Nan Jiang[*‡], Bangjie Sun[*‡], Nayoung Kim[†], Terence Sim[*], Jun Han[†]

[*] National University of Singapore, [†] KAIST

[‡] *Equal contribution*

Perceptual Hashing Protects Digital Trust

- Perceptual hashing (PHash) is widely adopted in online services
- Find image **copies** and **edited versions** across platforms



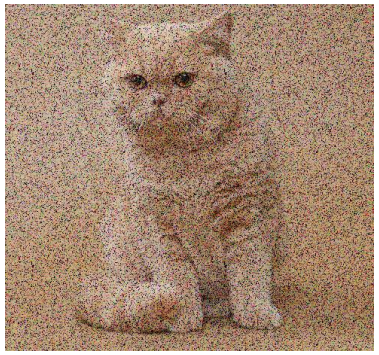
An example copyright monitoring service

An example image search engine

Background: Perceptual Hashing (PHash)

- Given an input image, PHash computes a (binary) **hash digest**
- Two **visually similar** images yield sufficiently **close** hash digests

Original



Edited

Meta Microsoft

Perceptual Hashing

Perceptual Hashing

Hash Bits:

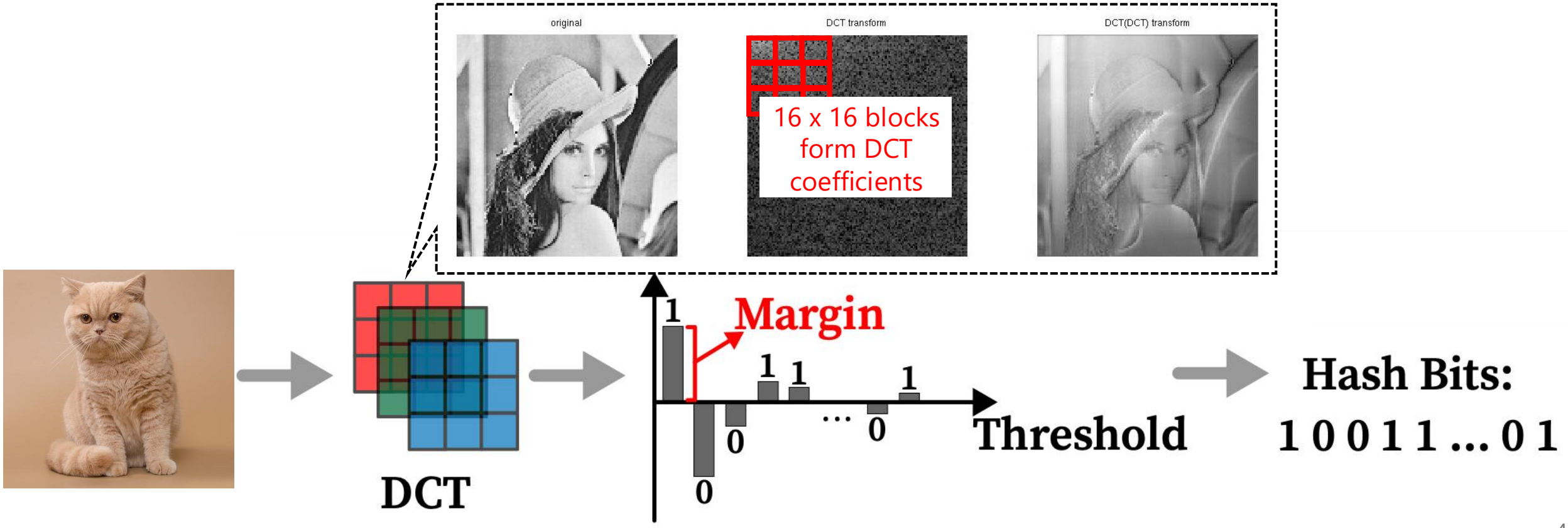
0 1 0 1 1 … 0

0 1 0 1 0 … 0

Small hash distance (same image!)

Background: DCT-Based PHash

- Discrete Cosine Transform (DCT) is used to process the image
- Low-frequency DCT coefficients captures the **image content**



Vulnerability of DCT-Based PHash

- **Bit-flipping attack:** Add **small** perturbations to images to flip **many** hash bits
- **Evade detection** of image copies **without severely degrading quality**

Prokos et al. [Security'23]

(a) Starting Image
L_2 Dist: 0

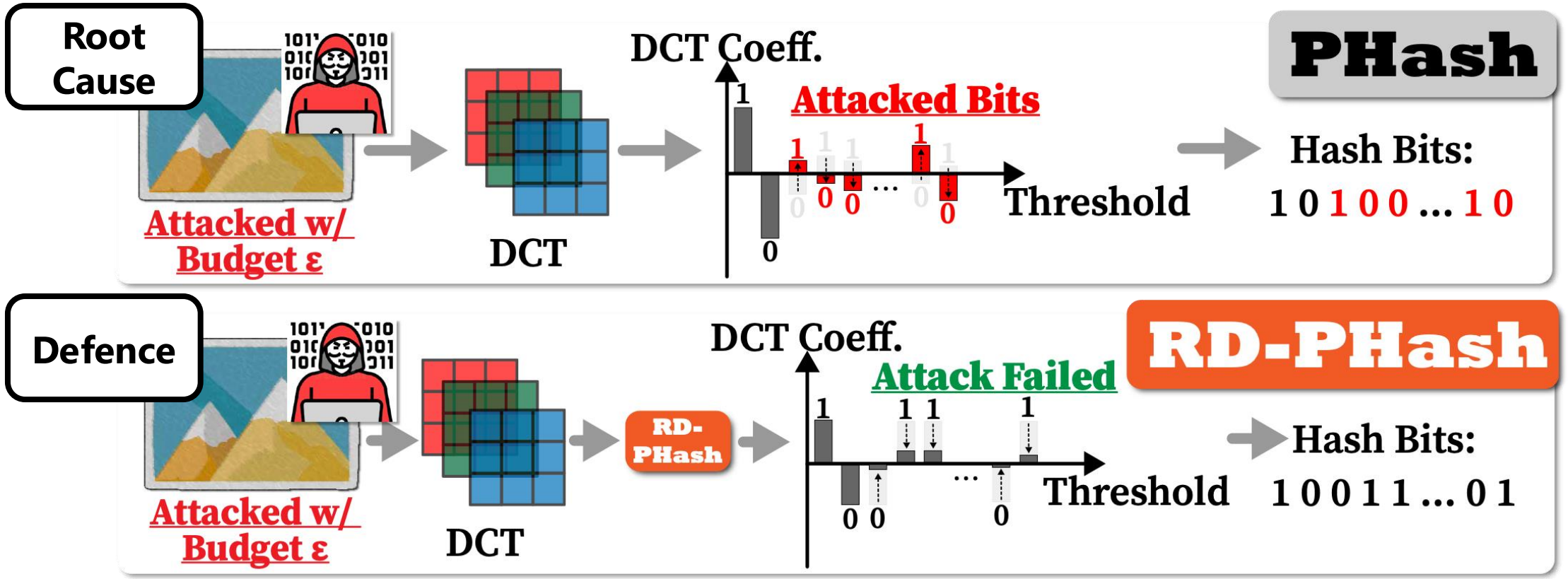
(b) $\Delta_d = 1800$ (BL)
L_2 Dist: 15.2

(c) $\Delta_d = 4000$
L_2 Dist: 40.2

Can we make DCT-based PHash more robust by reducing the number of flipped bits for the given amount of perturbation?

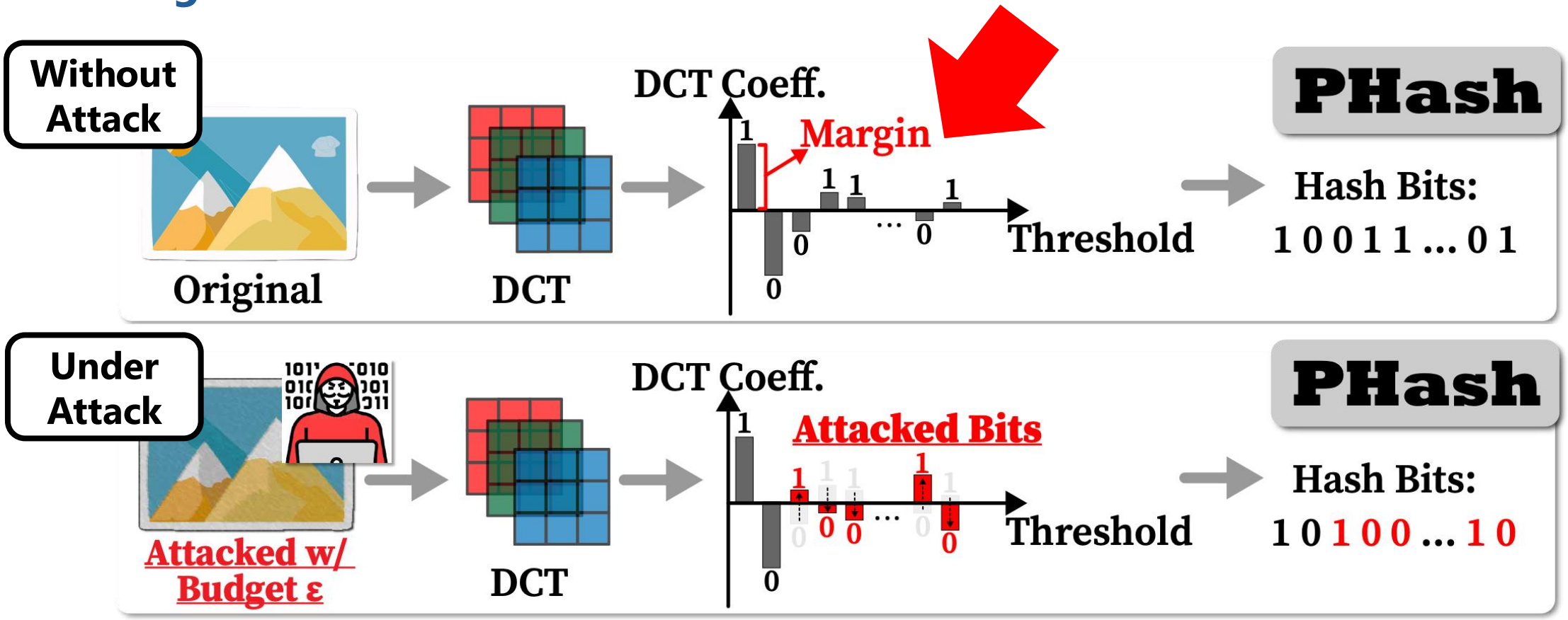
Our Work: *RD-PHash*

- (1) Analyze the root cause of DCT-based PHash's vulnerability
- (2) Introduce a lightweight enhancement to reduce the number of vulnerable bits



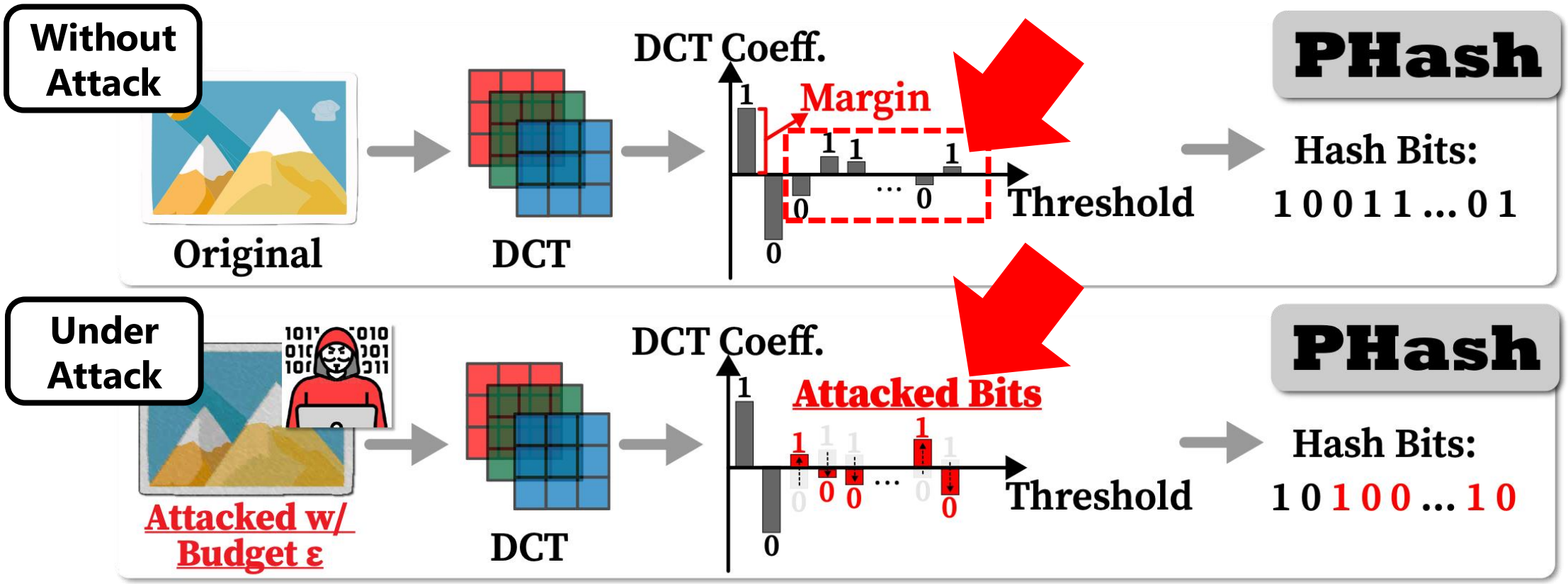
(1) Root Cause of DCT-Based PHash's Vulnerability

- **Perturbations** required to flip a hash bit **increase monotonically** with **margins** between the DCT coefficient and the threshold



(1) Root Cause of DCT-Based PHash's Vulnerability

- **Extremely imbalanced distribution** of margins introduces many bits with small margins that an attacker could easily flip with small perturbations



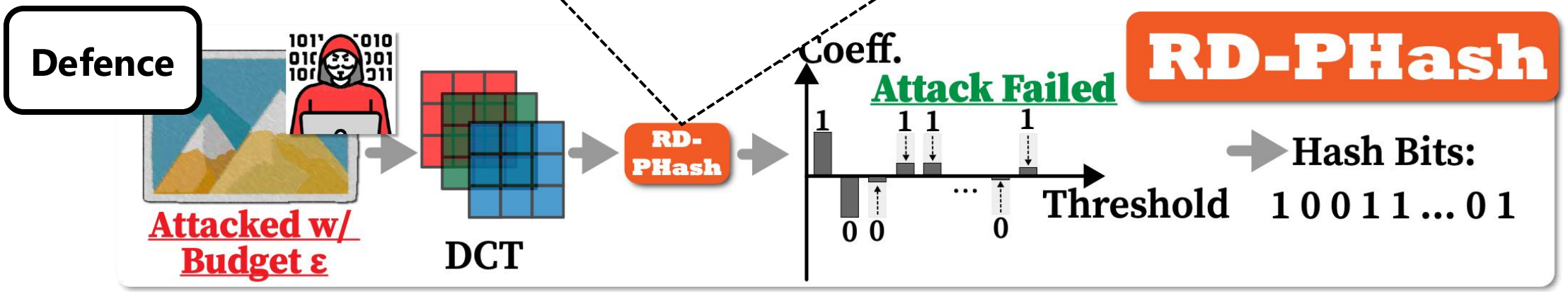
(2) Mitigation: Our Lightweight Solution

- With < 10 lines of code change, we can patch existing PHash methods
- **Key: *Redistribute* the margins** by applying an **orthonormal matrix**

Before redistribution: $DCT\ Coefficients = \boldsymbol{W} \cdot Img$

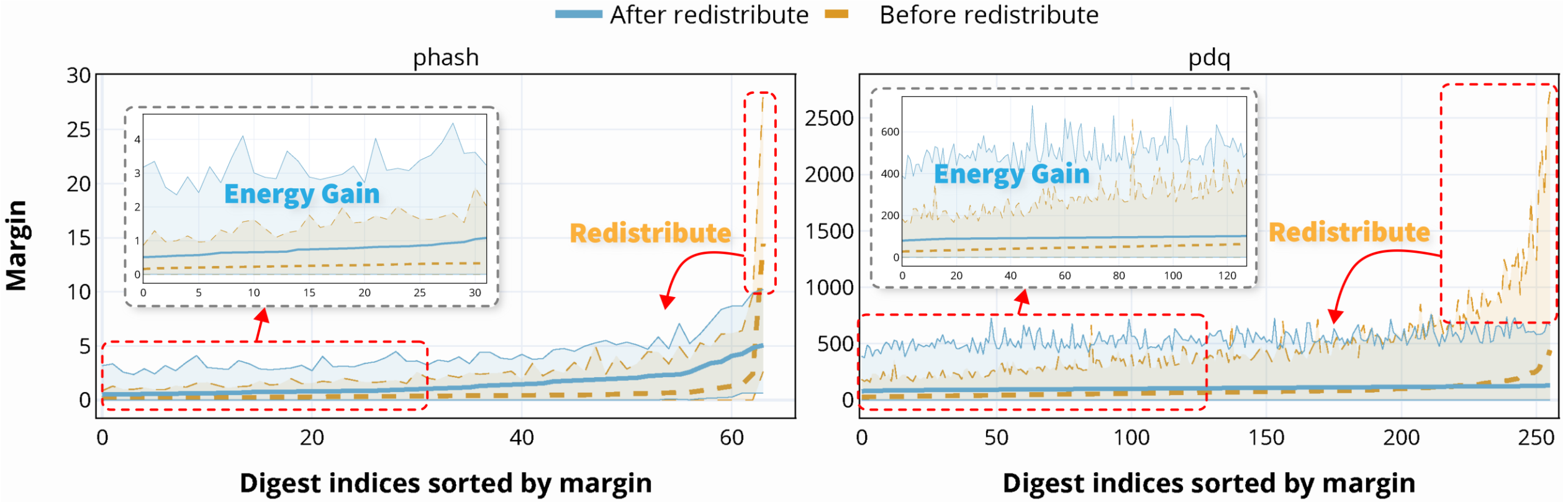
After redistribution: $DCT\ Coefficients = \boldsymbol{RW} \cdot Img$

Sampled from a Uniform distribution



(2) Mitigation: Analysis

- Margin redistribution effectively increases the amount of margins **at the tail**



Evaluation: Setup and Research Questions

Setup:

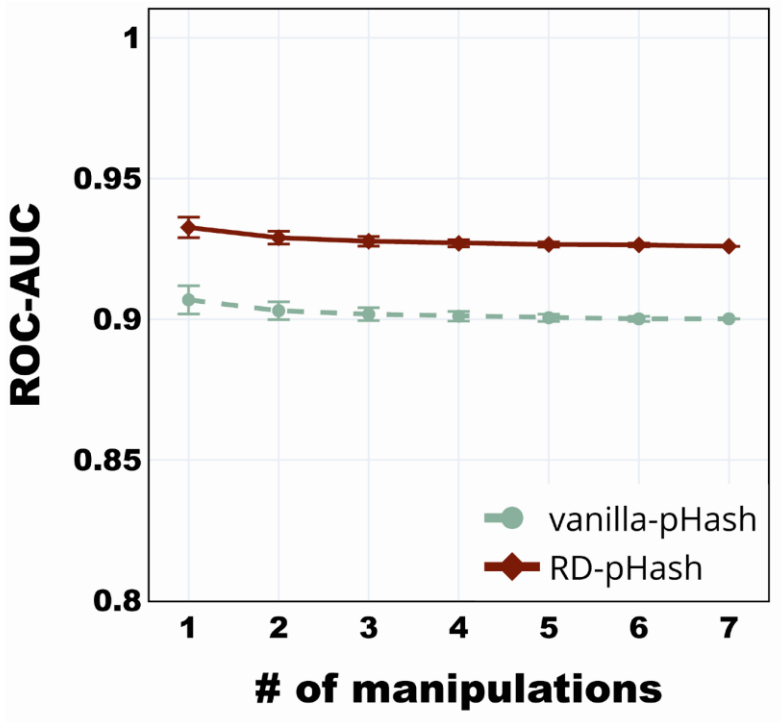
- **PHash schemes:** Phash and Meta's PDQ
- **Attacks:** one *black-box* and one *white-box* attack
- **Baseline:** CertPHash [Security'25]

Research questions:

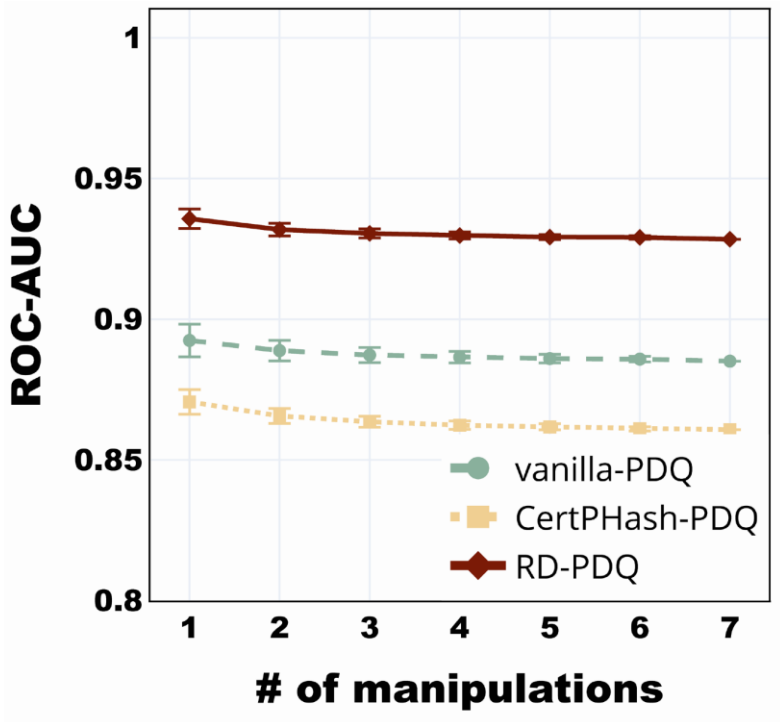
- **RQ1 (PHash Utility):** Resistance to transformations and hash-bit utilization
- **RQ2 (Empirical Robustness):** Robustness against perturbations under black-box and white-box settings

Evaluation: RQ1 PHash Utility

- *RD-PHash* outperforms vanilla versions of PHash schemes and CertPHash in terms of its **resistance to benign transformations**



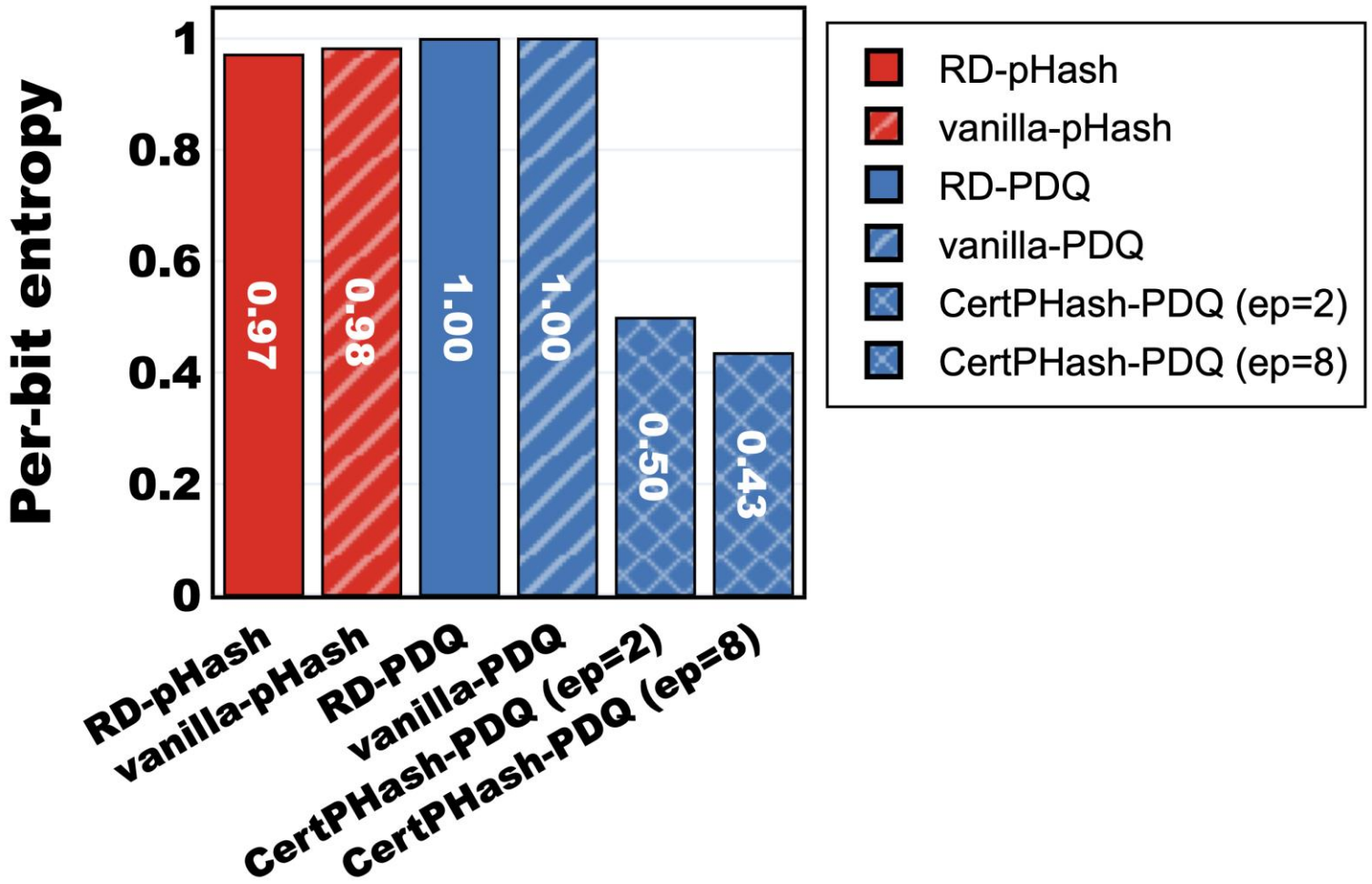
(a) pHash



(b) PDQ

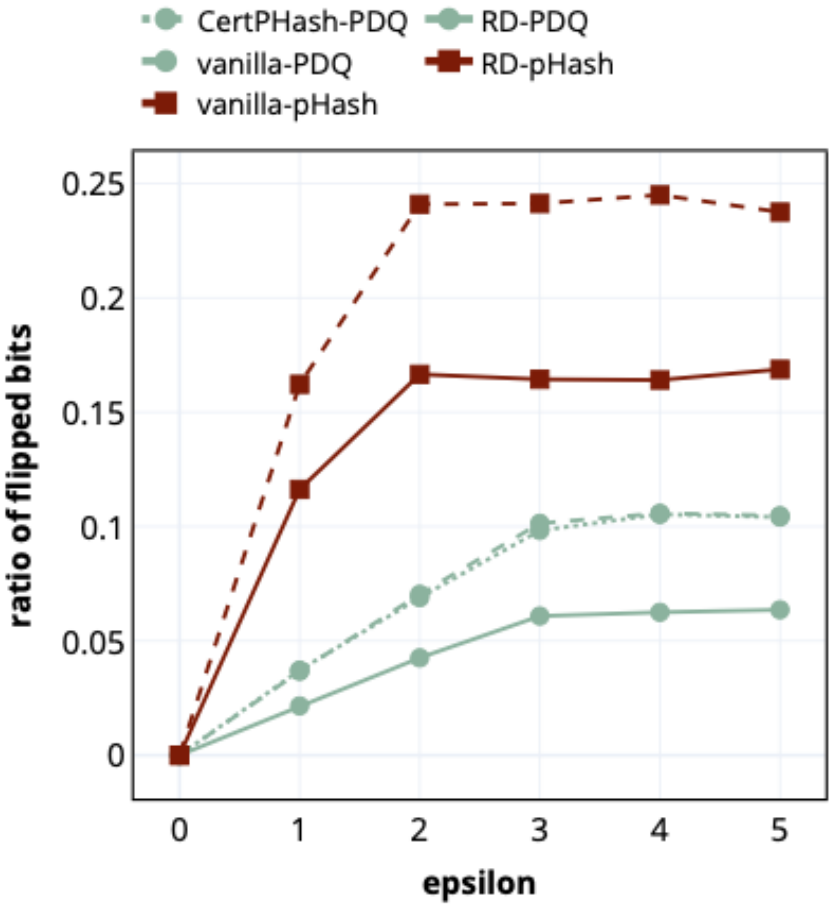
Evaluation: RQ1 PHash Utility

- *RD-PHash* preserves **hash-bit utilization**

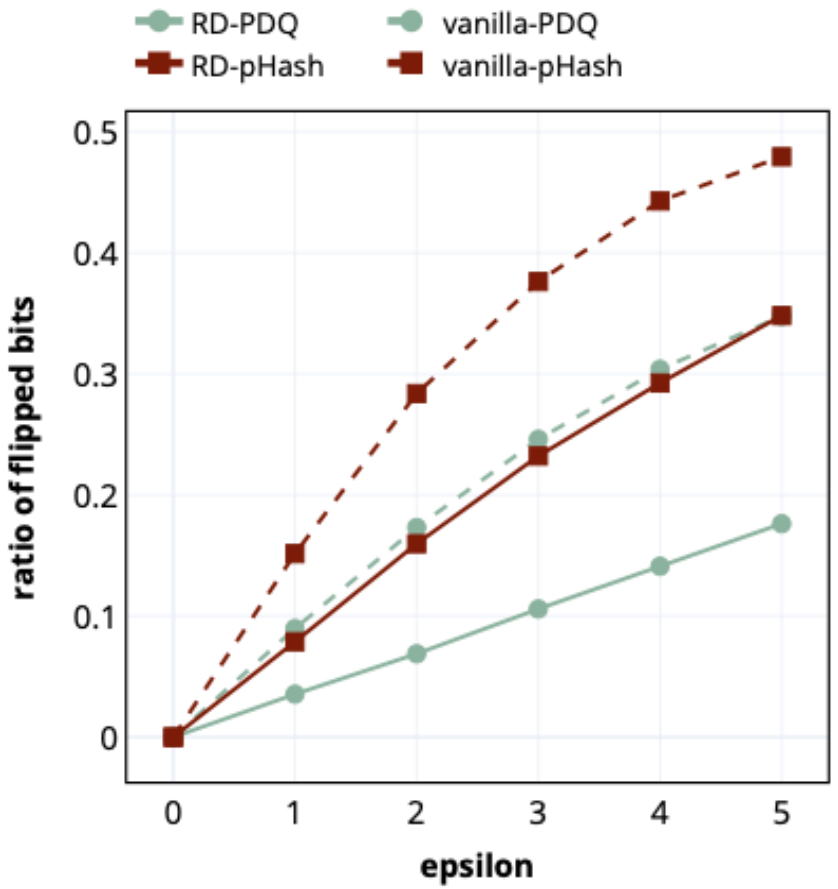


Evaluation: RQ2 Empirical Robustness

- *RD-PHash* outperforms vanilla versions of PHash schemes and CertPHash



Black-box setting



White-box setting

Discussion

- We plan to further extend RD-PHash in three ways

(1) Matrix Optimization

- Instead of random sampling, optimization of ***R*** could be investigated

(2) Certified Robustness

- Formal certification of RD-PHash, so that robustness guarantees can be stated under explicit perturbation bounds

(3) Scope Expansion

- Include neural-network-based hashes and more comprehensive attacks

Conclusion

- *RD-PHash* improves robustness against adversarial bit-flipping attacks
- Spur future research in designing robust PHash