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

Nan Jiang*‡, Bangjie Sun*‡, Nayoung Kim†, Terence Sim*, and Jun Han†
*National University of Singapore
†KAIST
‡These authors contributed equally

Abstract*—Unauthorized reuse of digital artworks and creative content has become a widespread concern on social media and content-sharing platforms, where copyrighted images are copied, intentionally modified, and redistributed without permission. Billions of images are estimated to be stolen each day. To protect creators, copyright monitoring services use near-duplicate matching techniques, such as perceptual hashing, to detect unauthorized reuse despite modifications. However, existing approaches remain vulnerable to adversarial perturbations, allowing unauthorized reuse to evade copyright detection and undermining the reliability of monitoring services. A recent defense, CertPHash, aims to enhance adversarial robustness, but for DCT-based perceptual hashing it offers limited robustness gains while substantially reducing hash-bit utilization, which may increase collision risk when operating at scale. This limitation likely arises because CertPHash applies a generic learning-based approach across different perceptual hashing schemes, rather than accounting for the design-specific weaknesses of DCT-based perceptual hashing. In this work, we identify the root cause that makes DCT-based perceptual hashing susceptible to adversarial perturbations, and propose *RD-PHash*, a robustness enhancement tailored to DCT-based hashes. Our experiments show that, for DCT-based hashes such as pHash and PDQ, *RD-PHash* outperforms CertPHash in both hash-bit utilization and adversarial robustness, achieving up to 2.3× higher utilization while reducing the impact of existing adversarial attacks by up to 71%.

1. Introduction

Unauthorized reuse of digital artworks has become a widespread concern on social media and content-sharing platforms, where creators publish artwork online and adversaries copy, intentionally modify, and repost it for financial gain [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. To assist creators in identifying unauthorized reuse of their works at scale, image monitoring services such as TinEye, Pixsy, and Google Images [11], [12], [13] track the online dissemination of protected images on their behalf. These services continuously monitor online image repositories and flag potential unauthorized reuse by identifying images that are visually similar to protected works. Rather than searching for exact duplicates, monitoring services typically rely on near-duplicate matching, which allows unauthorized reuse to be detected even when it undergoes modifications. In practice, near-duplicate matching is commonly implemented using image perceptual hashing (PHash) [14], [15], [16], [17], [18], [19]. Given an input image, PHash outputs a compact binary digest that captures the image's overall structure rather than exact pixel values. Two images are considered a match if their digests differ by no more than

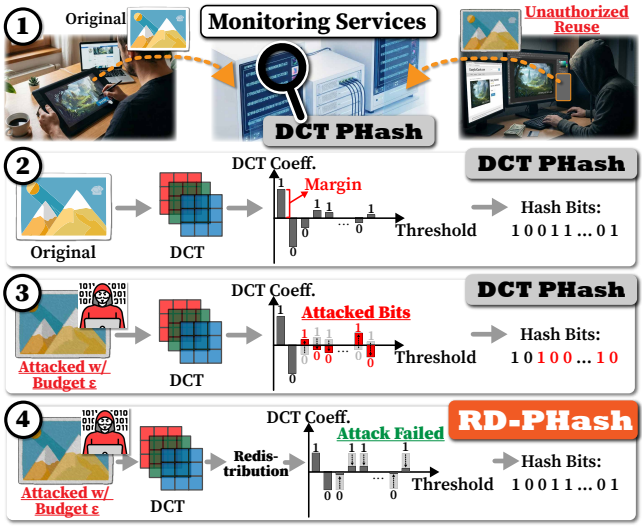

Figure 1: Illustration of a representative attack scenario. (1) Creators publish their works online, while adversaries may repost modified, unauthorized copies. Monitoring services continuously monitor and flag suspected copyright violations using PHash algorithms. (2) DCT-based PHash computes a fixed set of low-frequency DCT coefficients from an image and applies thresholding to convert these coefficients into a binary digest. (3) Under a perturbation quality budget ϵ, bit-flipping attacks add adversarial perturbations to an image, optimizing them to target coefficients with small margins that lie close to the hashing threshold, which are easier to flip. (4) To defend effectively, *RD-PHash* applies an energy-preserving redistribution of low-frequency DCT coefficients, balancing coefficient margins and reducing the number of bits that adversarial attacks can flip under a fixed perturbation budget.

a preset threshold. Critically, as digital artwork and image repositories grow to massive scale, sufficient hash-bit utilization is essential to prevent collisions.

However, PHash algorithms are known to be vulnerable to adversarial *bit-flipping attacks* [20], [21], [22], [23], in which an adversary applies visually imperceptible perturbations to an image to flip individual bits in its hash digest, thereby changing the digest used for matching without altering the image's overall structure. Bit-flipping attacks can lead to two harmful outcomes: (i) *evasion*, where the hash digest of unauthorized reuse differs from that of the protected work by more than the matching threshold, allowing it to evade detection by monitoring services; and (ii) *collision*, where visually dissimilar images produce hash digests that fall within the matching threshold of protected works, undermining the reliability of monitoring services. A recent work, CertPHash [19], aims to improve the robustness of PHash algorithms against adversarial bit-flipping attacks. However, it fails for DCT-based perceptual hashing (e.g., Meta's PDQ), offering limited robustness gains while substantially reducing hash-bit utilization. Our hypothesis is that this stems from applying a generic learning-based defense across diverse PHash designs, instead of addressing the unique vulnerabilities inherent in DCT-based PHash.

In this work, we identify the root cause that makes DCT-based PHash vulnerable to adversarial bit-flipping attacks and introduce *RD-PHash*, a robustness enhancement tailored to DCT-based PHash that mitigates this vulnerability. Figure 1 presents the core idea of *RD-PHash* together with a representative attack scenario. Specifically, DCT-based PHash generates binary digests by thresholding a fixed set of low-frequency DCT coefficients, assigning a bit value of 1 if a coefficient exceeds a threshold and 0 otherwise. We reveal that the vulnerability stems from a **highly imbalanced distribution of margins between DCT coefficients and the threshold** (see Figure 2). This imbalance occurs because only a few low-frequency coefficients have large values, while the majority remain small. Hash bits with small margins are therefore easy to flip, whereas hash bits with large margins are relatively stable. As a result, this imbalance enables an adversary to optimize small perturbations to target vulnerable bits and flip many of them, leading to collision or evasion. To address this vulnerability, we design an algorithm to **re**distribute low-frequency DCT energy across hash bits (hence the name *RD-PHash*), thereby balancing coefficient margins and increasing the overall difficulty of flipping individual bit. *RD-PHash* is lightweight and training-free, and can be integrated directly into existing DCT-based PHash algorithms with minimal effort (e.g., requiring fewer than ten lines of code), enabling practical deployment and operation at scale.

Our evaluation demonstrates that *RD-PHash* improves robustness over prior approaches, including vanilla DCT-based PHash methods (i.e., pHash and PDQ) as well as CertPHash, while preserving hash-bit utilization. Concretely, under the same ℓ_2 attack budget, *RD-PHash* reduces the number of adversarially flipped hash bits by up to 71% compared with vanilla pHash, vanilla PDQ, and CertPHash, across both black-box and white-box attacks. Moreover, *RD-PHash* preserves nearly full hash-bit utilization, achieving up to 2.3× higher utilization than CertPHash. We summarize our contributions as follows:

- We reveal that the core vulnerability of DCT-based PHash stems from the highly imbalanced distribution of low-frequency DCT energy used for hashing.
- We propose *RD-PHash*, a lightweight robustness enhancement that mitigates this vulnerability by redistributing low-frequency DCT energy to increase the cost of flipping bits across the entire hash digest.
- We empirically demonstrate that *RD-PHash* achieves an improved robustness-utilization trade-off over CertPHash, delivering larger hash-bit utilization together with stronger resistance to existing attacks.

2. Background and Related Work

DCT-based Perceptual Hashing Methods. DCT-based perceptual hashing (PHash) maps an image I to a compact binary code such that visually similar images remain close in Hamming space, making it suitable for large-scale near-duplicate detection and copyright monitoring. Specifically, DCT-based PHash converts an input image I to grayscale, resizes it to a fixed resolution of $N \times N$, computes the 2D discrete cosine transform (DCT), and retains only the low-frequency coefficients in the top-left $c \times c$ block. These operations can be expressed as a linear projection

$$\mathbf{z}(\tilde{I}) = W \,\mathrm{vec}(\tilde{I}), \tag{1}$$

where $\tilde{I}$ denotes the preprocessed image after grayscale conversion and resizing, $\mathrm{vec}(\cdot)$ denotes image vectorization, and $W \in \mathbb{R}^{c^2 \times N^2}$ corresponds to the 2D DCT followed by selection of the low-frequency coefficients in the top-left $c \times c$ block. The rows of W are orthonormal, i.e., $WW^\top = I_{c^2}$ [20].

Then, the binary hash $\mathbf{h}(\tilde{I}) \in \{0,1\}^m$ is obtained via median thresholding:

$$h_i(\tilde{I}) = \mathbb{I}\Big[z_i(\tilde{I}) \geq \tau\Big], \quad i = 1, \ldots, m, \tag{2}$$

where $\tau = \mathrm{median}(\mathbf{z})$ and $\mathbb{I}[\cdot]$ denotes the indicator function. Two images I and I' are declared a match if

$$d_H\big(\mathbf{h}(\tilde{I}),\, \mathbf{h}(\tilde{I'})\big) \leq \Delta, \tag{3}$$

where $d_H(\cdot,\cdot)$ denotes the Hamming distance and Δ is a predefined matching threshold.

Classical pHash [14] follows this paradigm with $m = 64$, while Meta's PDQ [15] uses improved preprocessing and a longer digest (typically $m = 256$) to achieve stronger discrimination at large scale.

Adversarial Attacks. Recent work shows that perceptual hashes are vulnerable to adversarial ***bit-flipping attacks***: an adversary applies visually imperceptible perturbations to an image to alter its resulting hash digest [20], [21], [22], [23]. Such bit-flipping attacks could lead to two harmful consequences: (i) *evasion*, where the hash digest of unauthorized

reuse is pushed outside the matching threshold (i.e., differs from that of the protected work by more than the matching threshold), allowing it to avoid detection; and (ii) *collision*, where the hash digest of an unrelated image is pushed within the matching threshold of a protected work, undermining the reliability of monitoring systems.

Bit-flipping attacks targeting evasion can be viewed as *untargeted* attacks, since evasion only requires flipping a sufficient number of bits to push the hash digest outside the matching threshold of the protected image, without targeting any specific bits. In contrast, bit-flipping attacks targeting collision are inherently *targeted*, as they must drive the hash toward a particular target digest and therefore require flipping a specific set of bits. In this work, we restrict our evaluation to untargeted bit-flipping attacks. Because untargeted attacks place fewer constraints on the adversary and allow flexibility in which bits are flipped, they generally require lower perturbation cost than targeted attacks. Consequently, evaluating robustness under untargeted attacks offers a conservative assessment of perceptual hashing robustness.

Robust Perceptual Hashing. To improve the robustness of PHash, several recent works focus on utilizing adversarial bit-flipping attacks via adversarial training [23], [24], [25], [26] or certified training [18], [19], [27], [28], [29], [30]. Among these, CertPHash [19] represents the current state of the art by being the first work providing certified robustness guarantees. However, CertPHash adopts a unified defense to improve robustness of different PHash designs. We observe that such a unified defense may not fully address the vulnerability present in specific PHash designs and can introduce trade-offs between robustness and hash-bit utilization. In this work, we focus on DCT-based hashes and study the root cause of their vulnerability to adversarial bit-flipping attacks. By analyzing the distribution of DCT coefficients used for hashing, we identify a design-specific weakness: low-frequency energy is unevenly distributed across hash bits, leaving many bits with small margins that are easy to flip. We develop *RD-PHash* to mitigate this vulnerability and evaluate its robustness and hash-bit utilization in §5.

3. Threat Model

Adversary Objectives. Following §2, the adversary's objective is to launch *untargeted* bit-flipping attacks. Specifically, the adversary aims to add visually imperceptible perturbations to an image, subject to a fixed quality budget, in order to maximize the Hamming distance between the perceptual hash of the perturbed image and that of the original. Formally, this objective can be expressed as:

$$\max_{\|\delta\|_p \le \epsilon} d_H\big(f(I), f(I+\delta)\big),$$

where δ denotes the image perturbation and ϵ bounds its magnitude under the ℓ_2 norm.

Adversary Capabilities. We consider two adversarial settings:

White-box: In the white-box setting, the adversary has full knowledge of the perceptual hashing algorithm, including the preprocessing steps, DCT computation, redistribution matrix and matching threshold Δ, and can directly optimize the perturbation δ using full knowledge of how perturbations affect the DCT coefficients used for hashing.

Black-box: The adversary has no access to the internal design of the hashing system, but can query the system with perturbed images to obtain hash digests and iteratively update the perturbation δ, subject to a query budget B.

4. Methodology

4.1. Core Weakness of DCT-based PHashes

We reveal that the vulnerability of DCT-based perceptual hashing to adversarial bit-flipping attacks arises from two key factors:

Factor (1). The low-frequency DCT coefficients used to compute the hash exhibit a highly skewed distribution: only a few large coefficients capture most of the image content, while the vast majority are small and carry little information.

Factor (2). Hash bits derived from low-information coefficients encode weaker visual features and can therefore be flipped with minimal visual impact, making them easy targets for adversarial perturbation.

We next present the evidence for Factor (1) by quantifying how the magnitude of each DCT coefficient determines the stability of its corresponding hash bit, through the notion of a per-bit margin, and computing its distribution. For a preprocessed image $\tilde{I}$ and each bit index $i \in \{1, \dots, m\}$, we define the *margin* of the i-th hash bit as

$$\gamma_i \;\coloneqq\; \big|z_i(\tilde{I}) - \tau\big| \;=\; \big|w_i \mathrm{vec}(\tilde{I}) - \tau\big|, \tag{4}$$

where $z_i(\tilde{I})$ denotes the pre-threshold coefficient for bit i, w_i is the i-th row of the projection matrix W, and τ is the threshold used to map the coefficient to a binary hash bit. The margin thus measures the absolute distance of the coefficient from the threshold. Figure 2 visualizes the distribution of margins across images. We observe a highly skewed margin distribution (orange region), in which many bits have small margins, while only a few exhibit much larger margins.

Furthermore, to support Factor (2), we observe a statistically significant monotonic correlation between bit margins and bit-flipping vulnerability. Under a fixed perturbation budget, the number of bits flipped is strongly associated with the average margin of the flipped bits, with Spearman's correlation coefficient $\rho = 0.75$ for PDQ and $\rho = 0.87$ for pHash ($p \ll 0.05$). Therefore, under a fixed ℓ_2 perturbation budget, attacks on images with smaller average margins tend to exhibit a larger number of flipped bits. This empirical relationship indicates that bit margin serves as a proxy for bit-flipping robustness. As a consequence of Factors (1) and (2), many hash bits exhibit inherently small margins, and under a fixed ℓ_2 perturbation budget, attacks can flip a large fraction of these bits, indicating increased vulnerability.

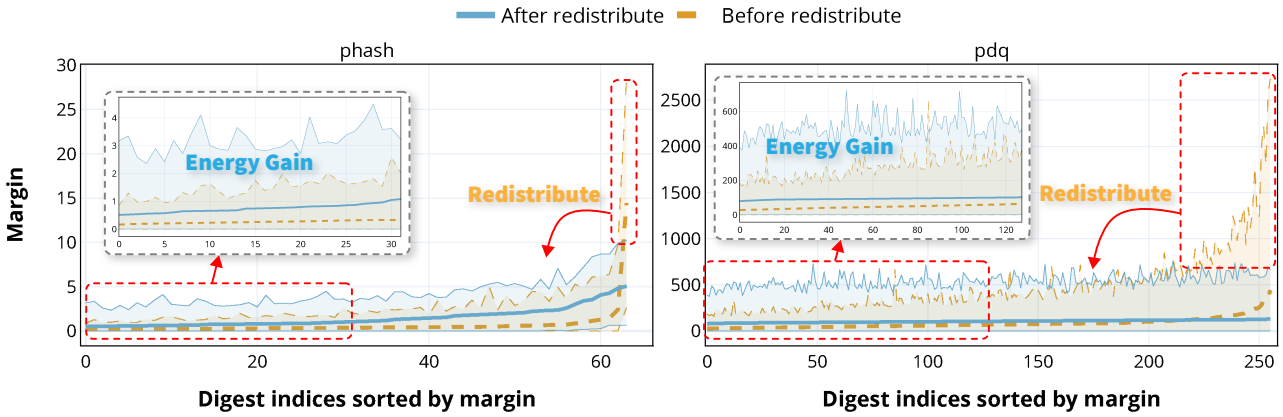

Figure 2: Illustration of *RD-PHash* on pHash and Meta's PDQ. Before redistribution of DCT energy (orange region), the margins between DCT coefficients and the threshold are highly skewed across hash bits, with most bits exhibiting small margins and thus being highly vulnerable to bit-flipping attacks. After redistribution (blue region), *RD-PHash* substantially increases the margins for most bits, thereby improving resistance to attacks. We recommend viewing this figure in color.

4.2. Redistribution of Low-frequency Energy

To address the core weakness identified in §4.1, we propose to redistribute the low-frequency DCT energy used for hashing more uniformly across hash bits, thereby strengthening those associated with small margins. Specifically, our redistribution step forms each new coefficient as a weighted sum of the original low-frequency coefficients:

$$\tilde{z}_i \;=\; \sum_{j=1}^{m} r_{ij}\, z_j, \qquad i = 1, \ldots, m,$$

where the weights $\{r_{ij}\}$ are sampled once and fixed thereafter. Collecting these weights into a matrix $R \in \mathbb{R}^{m\times m}$, the redistribution can be written compactly as

$$\tilde{\mathbf{z}} \;=\; R\mathbf{z}.$$

We choose R to be a uniformly distributed[1] orthonormal matrix, i.e., $R^\top R = I$, and apply the median thresholding rule to $\tilde{\mathbf{z}}$ to obtain the binary digest.

The orthonormality of R is critical. Orthonormality guarantees preservation of the total DCT coefficient energy ($\|\tilde{\mathbf{z}}\|_2 = \|\mathbf{z}\|_2$), and does not introduce artificial amplification or attenuation of energy used for hashing. In addition, it ensures that each output coefficient receives, in expectation, an equal share of the total energy and that no coefficient is systematically amplified or suppressed.

Lemma 1 (Uniform energy allocation)**.** *Fix any $\mathbf{z} \in \mathbb{R}^m$ and consider an orthonormal transform $R \in \mathbb{R}^{m\times m}$ satisfying $R^\top R = I$. Then, taking expectation over the space of all orthonormal transforms, for every $i \in \{1, \ldots, m\}$,*

$$\mathbb{E}\big[\tilde{z}_i^2\big] \;=\; \frac{\|\mathbf{z}\|_2^2}{m}. \tag{5}$$

Consequently, the transform redistributes energy across coordinates without introducing per-index scaling bias in expectation.

Proof. Write $\tilde{z}_i = r_i\mathbf{z}$, where r_i denotes the i-th row of R. Taking expectation over the space of all orthonormal transforms, each row vector r_i is isotropic on the unit sphere, implying

$$\mathbb{E}\big[r_i^\top r_i\big] = \frac{1}{m} I.$$

Therefore,

$$\mathbb{E}\big[\tilde{z}_i^2\big] = \mathbb{E}\big[(r_i\mathbf{z})^2\big] = \mathbf{z}^\top \mathbb{E}\big[r_i^\top r_i\big]\,\mathbf{z} = \frac{1}{m}\|\mathbf{z}\|_2^2,$$

which completes the proof. □

Figure 2 (blue region) illustrates the reduced margin skewness after redistribution. We observe a substantial margin increase for most bits, indicating greater resistance to bit flipping. Importantly, as illustrated in §5.1, this design preserves hash-bit utilization and robustness under benign transformations while substantially increasing the attack budget for both white-box[2] and black-box adversaries to succeed. These findings suggest that margin redistribution is a key ingredient for robust DCT-based perceptual hashing.

5. Evaluation

Through evaluation of *RD-PHash*, we aim to answer the following two research questions:

- **RQ1 (PHash Utility).** To what extent does the proposed scheme preserve the utility of PHash, including resistance to benign transformations and hash-bit utilization?
- **RQ2 (Empirical Robustness).** How robust is the proposed scheme against existing adversarial attacks under black-box and white-box settings?

1. While we sample R uniformly in this work, we leave its optimization for future research (see §6).

2. In the white-box setting, the adversary has full knowledge of R.

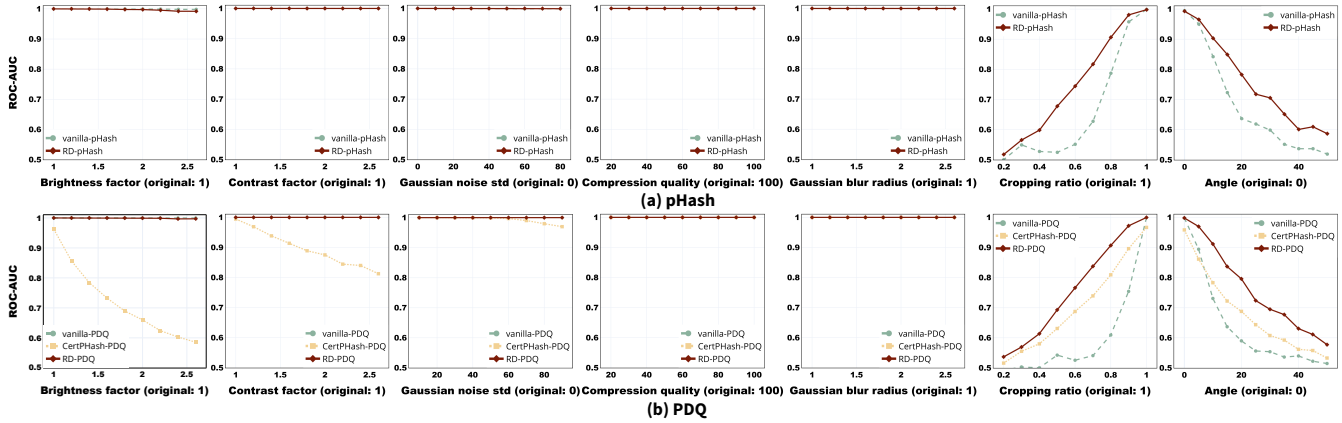

Figure 3: ROC-AUC of perceptual hashing methods under seven benign transformations (i.e., *brightness, contrast, Gaussian noise, JPEG compression, Gaussian blur, cropping, and rotation*) for (a) pHash and (b) PDQ.

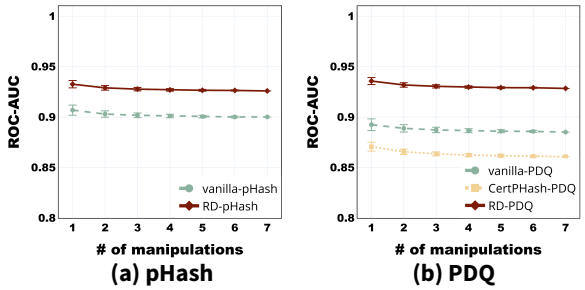

Figure 4: ROC-AUC of perceptual hashing methods under a combination of transformations with increasing number of manipulations for (a) pHash and (b) PDQ.

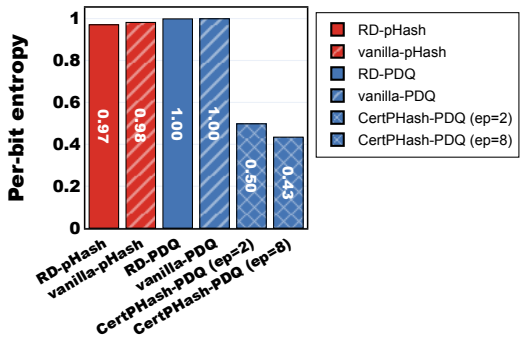

Figure 5: Per-bit entropy for PDQ and pHash with and without *RD-PHash*. For PDQ, we observe that *RD-PHash* preserves nearly full hash-bit utilization while CertPHash exhibits substantially lower utilization.

5.1. RQ1: PHash Utility

Setup. We evaluate *RD-PHash* on 1,000 images randomly sampled from the MS-COCO dataset. For each image I, we generate benignly transformed variants $T(I)$ by varying the intensity of seven common transformations: *brightness, contrast, cropping, Gaussian blur, rotation, Gaussian noise, and JPEG compression*. We then sweep the Hamming-

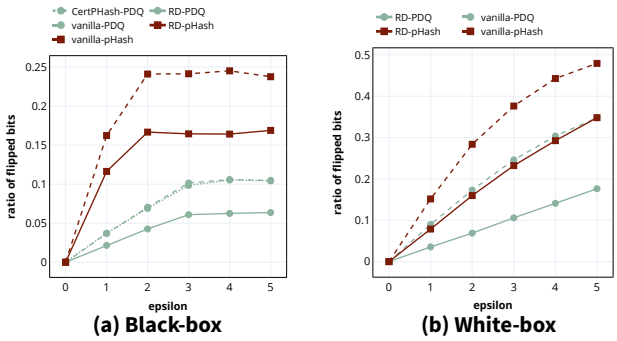

Figure 6: Performance of *RD-PHash* compared to baseline methods in terms of ratio of flipped bits under (a) black-box attacks and (b) white-box attacks. Note that solid lines indicate performance of *RD-PHash* while dotted lines represent performance of baseline methods.

distance threshold and report the ROC-AUC. A higher ROC-AUC indicates stronger robustness to these transformations. We also measure the *per-bit entropy* over all 1,000 images:

$$H_{\text{avg}} = -\frac{1}{L}\sum_{j=1}^{L}\left[p_j \log_2 p_j + (1-p_j)\log_2(1-p_j)\right] \quad (6)$$

where $p_j = \frac{1}{1000}\sum_{i=1}^{1000} h_{i,j}$, $h_{i,j} \in \{0,1\}$ denotes the j-th bit of the hash for the i-th image, and L is the total length.

Robustness to Transformations. Figure 3 depicts the ROC-AUC of *RD-PHash* in comparison with baseline methods, including vanilla pHash, vanilla PDQ, and CertPHash, under seven image transformations. For photometric transformations (*brightness, contrast, blur, noise*) as well as JPEG compression, *RD-PHash* closely matches the vanilla versions of pHash and PDQ, maintaining consistently high ROC-AUC. In contrast, CertPHash exhibits a noticeable decline in ROC-AUC as the brightness and contrast factors increase, suggesting reduced robustness to such photometric changes. For geometric transformations (*cropping, rotation*), *RD-PHash* substantially outperforms vanilla pHash, vanilla PDQ, and

CertPHash, consistently achieving higher ROC-AUC as the cropping ratio and rotation angle grow. We further evaluate robustness under compositions of transformations. Figure 4(a) and (b) present the ROC-AUC of *RD-PHash* and the baselines as the number of combined transformations increases. Again, *RD-PHash* consistently outperforms the original pHash, original PDQ, and CertPHash, with ROC-AUC gains exceeding 5%.

Hash-Bit Utilization. To further assess collision risk at scale, we analyze the hash-bit utilization of pHash- and PDQ-based variants, with results shown in Figure 5. The *per-bit entropy* results indicate that *RD-PHash* preserves nearly full hash-bit utilization. In particular, RD-pHash (0.97) remains close to vanilla pHash (0.98), while RD-PDQ (1.00) is identical to vanilla PDQ (1.00). In contrast, CertPHash-PDQ exhibits substantially lower per-bit entropy (0.50 for ϵ=2 and 0.43 for ϵ=8), indicating a marked reduction in hash-bit utilization.

Answer to RQ1: At the 1,000-image scale, *RD-PHash* preserves the practical utility of perceptual hashing by remaining robust to common image transformations while maintaining nearly full hash-bit utilization, and thus a low collision tendency for perceptually distinct images.

5.2. RQ2: Empirical Robustness

Setup. We prepare four datasets covering both real-world photographs (MS-COCO and ImageNet) and AI-generated images (StableDiffusion-v21 and Amazon-Generated-Images), and evaluate *RD-PHash* with 100 images randomly selected from each dataset, totaling 400 images. For each image, we evaluate robustness under both *black-box* and *white-box* attack settings [20] with perturbation budget $\epsilon \in \{0, 1, 2, 3, 4, 5\}$. Note that the white-box attack has access to the redistribution matrix R. We report the *ratio of flipped bits* in the final hash digest. A lower flipped-bit ratio indicates stronger empirical robustness, since the attack is less effective at altering the hash digest under the same perturbation budget.

Results. Figure 6 shows that *RD-PHash* consistently improves empirical robustness for both pHash- and PDQ-based hashing, outperforming all baseline methods, including vanilla pHash, vanilla PDQ, and CertPHash. Figure 6(a) presents the black-box results. For pHash-based hashing, *RD-PHash* flips fewer bits than vanilla-pHash across all non-zero perturbation budgets. For PDQ-based hashing, RD-PDQ achieves the lowest flipped-bit ratio among all compared PDQ-based variants, whereas both vanilla-PDQ and CertPHash exhibit higher flipped-bit ratios as the perturbation budget increases. Notably, CertPHash-PDQ remains close to vanilla-PDQ and consistently above RD-PDQ. Furthermore, Figure 6(b) presents the white-box results. Even under this stronger threat model, *RD-PHash* continues to outperform vanilla-pHash and vanilla-PDQ by a clear margin. We further observe that the flipped-bit ratio increases with ϵ in the white-box setting, whereas in the black-box setting it tends to saturate as ϵ grows. The overall flipped-bit ratio is also generally higher in the white-box setting, as expected given the stronger adversary.

Answer to RQ2: *RD-PHash* significantly improves the empirical robustness against both black-box and white-box adversarial attacks. While strong white-box attacks cause significant bit flips, *RD-PHash* outperforms baseline methods by suppressing these changes under all tested perturbation budgets.

6. Discussion

Deployment Consideration. Requiring under ten lines of code, *RD-PHash* can be easily integrated into real-world moderation and copyright systems. Furthermore, despite the complex relationship between imperceptibility and the l_2 budget, *RD-PHash* reliably undermines attack stealth by forcing successful perturbations to be visually noticeable.

Scope and Limitations. This work focuses on DCT-based perceptual hashing. We identify a fundamental vulnerability caused by the imbalanced energy distribution of low-frequency coefficients and propose an effective mitigation strategy. While our approach significantly improves robustness over prior defenses, it does not address non-DCT methods like gradient-based or learning-based hashing. These alternative schemes exhibit different vulnerabilities and require separate analysis. Developing robust defenses for these methods while preserving hash-bit utilization remains a key challenge for future research.

Future Work on Certified Robustness and Broader Attack Coverage. Our current results focus on empirical robustness improvements. A natural next step is formal certification of *RD-PHash*, so that robustness guarantees can be stated under explicit perturbation bounds. We are currently extending *RD-PHash* in this direction and plan to combine certification with a more comprehensive attack suite, including a wider range of adaptive attacks and composite transformations. This will help characterize both the guarantees and the practical limits of *RD-PHash*. Additionally, future work will thoroughly analyze the redistribution matrix R to optimize redistribution efficiently and securely.

7. Conclusion

We present *RD-PHash*, a lightweight robustness enhancement for DCT-based perceptual hashing that improves resistance to adversarial bit-flipping attacks without sacrificing hash-bit utilization. More importantly, *RD-PHash* points to a broader lesson for future research: robust defenses should not be built as generic patches, but should instead emerge from a careful understanding of the intrinsic weaknesses of individual hash designs and mechanisms, with mitigation crafted to address those vulnerabilities at their source. In this spirit, we hope this work helps open new directions for principled robustness research in perceptual hashing and, ultimately, advances the protection of digital artworks and creative content in an era marked by their rapid proliferation, transformation, and potential misuse.

References

[1] NANPA, "Global infringement report 2019," 2019.

[2] Copytrack, "How big a problem is copyright infringement?" 2019.

[3] D. Djudjic, "Over 2.5 billion online images are stolen every day, copytrack reports," 2019.

[4] G. Harris, "A lot of photographers find out about image theft when the culprits tag them in social media," 2020.

[5] berify, "[infographic] a snapshot of online image theft," 2018.

[6] C. Cole, "How artists can use copyright law to protect their work and build their legacy," 2024.

[7] R. M. Manar, "9 common reasons for copyright infringement and its effects," 2024.

[8] B. Sutton, "How artists can use copyright law to safeguard their work," 2018.

[9] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models," in *2024 IEEE symposium on security and privacy (SP)*. IEEE, 2024, pp. 807–825.

[10] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, and A. Tran, "Anti-dreambooth: Protecting users from personalized text-to-image synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2116–2127.

[11] TinEye, "Tineye reverse image search," 2026, accessed: 2026-02-27.

[12] Pixsy, "Pixsy: Fight for your images," 2026, accessed: 2026-02-27.

[13] Google, "Google images," 2026, accessed: 2026-02-27.

[14] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," Master's thesis, Upper Austria University of Applied Sciences, Hagenberg Campus, 2010.

[15] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and A. Babenko, "Tmk+pdqf: A test drive of facebook's perceptual hashing algorithms," 2020.

[16] B. Coskun and B. Sankur, "Robust video hash extraction," in *2004 12th European Signal Processing Conference*. IEEE, 2004, pp. 2295–2298.

[17] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," *Master's thesis*, 2010.

[18] J. Madden, M. Bhavsar, L. Dorje, and X. Li, "Robustness of practical perceptual hashing algorithms to hash-evasion and hash-inversion attacks," *arXiv preprint arXiv:2406.00918*, 2024.

[19] Y. Yang, Q. Liu, C. Brix, H. Zhang, and Y. Cao, "Certphash: Towards certified perceptual hashing via robust training," in *34th USENIX Security Symposium (USENIX Security 2025)*. USENIX Association, 2025, pp. 7839–7856.

[20] S. Jain, A.-M. Creţu, and Y.-A. de Montjoye, "Adversarial detection avoidance attacks: Evaluating the robustness of perceptual hashing-based client-side scanning," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 2317–2334.

[21] J. Prokos, N. Fendley, M. Green, R. Schuster, E. Tromer, T. Jois, and Y. Cao, "Squint hard enough: Attacking perceptual hashing with adversarial machine learning," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 211–228.

[22] L. Struppek, D. Hintersdorf, D. Neider, and K. Kersting, "Learning to break deep perceptual hashing: The use case neuralhash," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 58–69.

[23] X. Wang, Z. Zhang, G. Lu, and Y. Xu, "Targeted attack and defense for deep hashing," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2298–2302.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[26] B. Sun, M. C. Chan, and J. Han, "Camprints: Leveraging the" fingerprints" of digital cameras to combat image theft," in *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, 2025, pp. 473–486.

[27] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli, "On the effectiveness of interval bound propagation for training verifiably robust models," *arXiv preprint arXiv:1810.12715*, 2018.

[28] M. N. Mueller, F. Eckert, M. Fischer, and M. Vechev, "Certified training: Small boxes are all you need," *arXiv preprint arXiv:2210.04871*, 2022.

[29] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, "Scaling provable adversarial defenses," *Advances in neural information processing systems*, vol. 31, 2018.

[30] H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning, and C.-J. Hsieh, "Towards stable and efficient training of verifiably robust neural networks," *arXiv preprint arXiv:1906.06316*, 2019.

Appendix A. Ethics Considerations

This work examines the robustness of DCT-based perceptual hashing systems used for copyright monitoring, with the goal of improving their robustness against adversarial bit-flipping attacks. Although we analyze known attack strategies, these are included solely to enable rigorous evaluation and are based on techniques already described in prior work. All experiments are conducted offline on public or synthetic datasets, without interacting with deployed systems or exposing sensitive information. Overall, we believe this work strengthens the reliability of DCT-based perceptual hashing systems while adhering to responsible research practices.