# Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning

**Shubham Jain***
Imperial College London
`s.jain@imperial.ac.uk`

**Ana-Maria Creţu***
Imperial College London
`a.cretu@imperial.ac.uk`

**Yves-Alexandre de Montjoye**
Imperial College London
`demontjoye@imperial.ac.uk`

## Abstract

End-to-end encryption (E2EE) in messaging platforms enables people to securely and privately communicate with one another. Its widespread adoption however raised concerns that illegal content might now be shared undetected. Client-side scanning based on perceptual hashing has been recently proposed by governments and researchers to detect illegal content in E2EE communications. We propose the first framework to evaluate the robustness of perceptual hashing-based client-side scanning to detection avoidance attacks and show current systems to not be robust. We propose three adversarial attacks—a general black-box attack and two white-box attacks for discrete cosine transform-based algorithms–against perceptual hashing algorithms. In a large-scale evaluation, we show perceptual hashing-based client-side scanning mechanisms to be highly vulnerable to detection avoidance attacks in a black-box setting, with more than 99.9% of images successfully attacked while preserving the content of the image. We further show several mitigation strategies, such as expanding the database with hashes of images modified using our attack, or increasing the detection threshold, to be ineffective against our attack. Taken together, our results shed serious doubts on the robustness of perceptual hashing-based client-side scanning mechanisms currently proposed by governments, organizations, and researchers around the world.

## 1 Introduction

More than two billion people across the world use end-to-end encryption (E2EE)-enabled platforms such as Signal and WhatsApp [21, 4], exchanging more than 100 billion messages daily on WhatsApp alone [20]. E2EE provides strong privacy protection to users, preventing governments, hackers, and platform providers themselves to access the content of their communications. Governments and organizations have however raised concerns that E2EE is preventing the detection of illegal content [14, 16] such as online disinformation, child sexual abuse media, or terrorism related content, as required by law in many countries [15, 2].

Client-side scanning has been proposed to detect the sharing and storage of illegal content on E2EE-enabled platforms by industrial practitioners, researchers and policy makers [1, 6, 9, 17, 13]. Here, a signature of a visual media (image, video) would be computed on the user's device and then compared against the database of signatures of known illegal images. If a match is found, the user would be flagged and/or the unencrypted content automatically shared for further review. Designed to be robust to small changes to the media, as well as transformations like rotation and rescaling, perceptual hashing algorithms are used to generate the signature. Several variations of this scheme have been proposed, e.g., Apple's proposal sends a cryptographic voucher containing the encrypted images that the system would be able to decrypt only if the number of matches reach a predefined threshold [1].

*The first two authors contributed equally and are listed in a random order.

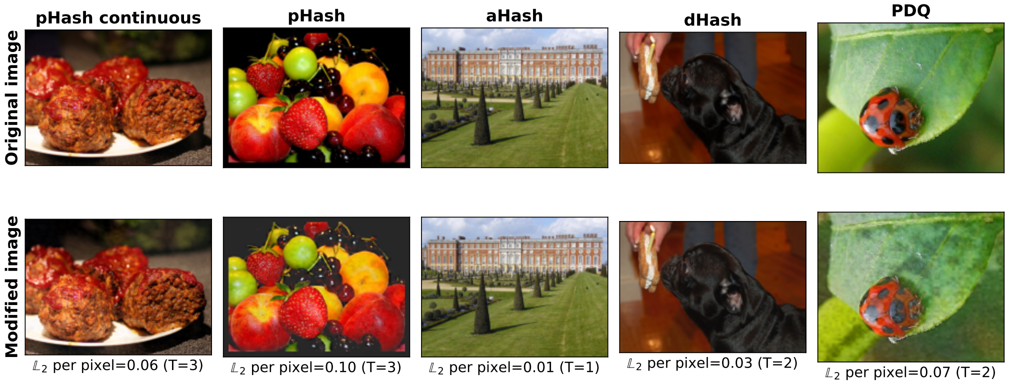
pHash continuous
pHash
aHash
dHash
PDQ
Original image
Modified image
$\mathbb{L}_2$ per pixel=0.06 (T=3)
$\mathbb{L}_2$ per pixel=0.10 (T=3)
$\mathbb{L}_2$ per pixel=0.01 (T=1)
$\mathbb{L}_2$ per pixel=0.03 (T=2)
$\mathbb{L}_2$ per pixel=0.07 (T=2)

Figure 1: Examples of original and modified images with $\mathbb{L}_2$ perturbation per pixel for different hashing functions. All the images evade detection while maintaining a high visual similarity to the original image.

We propose the first framework to evaluate the robustness of perceptual hashing-based client-side scanning to a novel *detection avoidance attack*. Our framework assesses the threat posed by adversarial attacks aiming to minimally modify an image so as to avoid detection while preserving its content. Our framework also takes into account the diversity of modifications produced by the attack in terms of distances between signatures. The more diverse the perturbations, the harder it is for the system to mitigate by expanding the database with the hashes of adversarially modified images.

We propose three adversarial attacks against perceptual hashing algorithms, a general black-box attack, inspired by previous work in adversarial ML, e.g., [10, 19, 22], and two novel and optimal white-box attacks exploiting the linearity and orthonormality of discrete cosine transform (DCT)-based perceptual hashing algorithms. The attacks are designed to minimally perturb the image while avoiding detection. In particular, they produce a wide range of perturbations preventing easy mitigation strategies such as expanding the database with modified images.

The full paper is publicly available as a preprint at `https://arxiv.org/abs/2106.09820`.

## 2 Attack model

**Perceptual hashing algorithms.** Perceptual hashing algorithms compute a signature of an image without having to share it, by extracting features that remain invariant under small modifications, such as resizing, or noise addition. Formally, a perceptual hashing algorithm $h : \mathcal{I} \rightarrow \mathcal{O}^l$ is a deterministic function mapping a multimedia $X \in \mathcal{I}$ to a fixed-size vector representation, the *hash*, usually consisting of bits ($\mathcal{O} = \{0, 1\}$) or real numbers ($\mathcal{O} = \mathbb{R}$). The similarity between two media $X, X' \in \mathcal{I}$ is quantified by computing the *distance* between the hashes according to a metric $d$.

**Client-side scanning.** Perceptual hashing-based client-side scanning for illegal image detection consists of a database $\mathcal{D} = \{X_1, \ldots, X_N\}$ of $N$ images $X_i \in \mathcal{I}, \forall 1 \leq i \leq N$, a perceptual hashing algorithm $h$, a distance $d$, and a threshold $T > 0$. Given an image $X \in \mathcal{I}$, the detection system computes the distance between the hash $h(X)$ and the hashes of images in the database $h(X_i)$. The image is flagged if there exists $1 \leq i \leq N$ such that $d(h(X_i), h(X)) \leq T$. We note that our framework is applicable more broadly to visual media (e.g., videos).

**Attack model.** We propose an adversarial attack against perceptual hashing-based client-side scanning, which we call *detection avoidance attack*. We assume that a malicious agent, *the attacker*, is in possession of a image from the database, henceforth denoted *original image* $X \in \mathcal{D}$. The attacker's goal is to minimally modify $X$ into $X'$ such that its content is preserved while avoiding detection. More specifically, the attacker's goal is to modify $X$ via an additive perturbation $\delta$ such that the modified image: (1) is valid, i.e., $X' = X + \delta \in \mathcal{I}$ meaning that no pixel goes out of bounds, (2) evades detection, i.e., $d(h(X), h(X + \delta)) > T$, where $T$ is the threshold used by the detection system and (3) the perturbation is minimal in terms of visual dissimilarity. We assume that the attacker knows the distance $d$ and the threshold $T$.

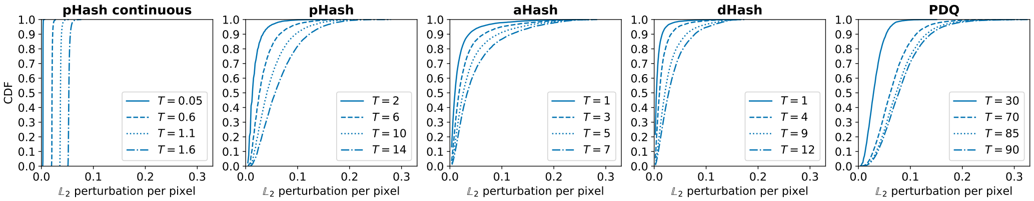

Figure 2: Cumulative distribution function (CDF) for the $\mathbb{L}_2$ perturbations per pixel for different perceptual hashing algorithms and thresholds. A lower perturbation indicates higher visual similarity between the modified and original images. The perturbation increases slowly with the threshold, but remains small in all cases.

**Perturbation diversity requirement.** Additionally, the attack should be resistant to simple defenses that the system could implement, such as expanding the database with hashes of modified images. The detection system might indeed gain knowledge of the attack. The attack should therefore produce a wide range of random perturbations that cannot be predicted and added to the database.

**Attacker access.** We consider two levels of attacker access to the perceptual hashing algorithm: black-box and white-box access. The assumption of black-box access is realistic in the context of perceptual hashing-based client-side scanning, as the most feasible and practical implementation would be to deploy the algorithm on the user device in a compiled form along with the messaging application [5]. The assumption of white-box access can be satisfied as the result of reverse-engineering work and the limited number of available perceptual hashing algorithms available.

## 3 Attack methodology

**Black-box attack.** Our black-box attack seeks a perturbation $\delta$ maximizing the distance $f(\delta) := d(h(X), h(X+\delta))$ between the hashes of original and modified images $X$ and $X' = X + \delta$ while ensuring the perturbation $||\delta||_2$ is smaller than a fixed constant $\epsilon$. More specifically our attack seeks a solution to the following optimization problem:

$$\text{Find: } \max_{\delta} \min(T, f(\delta)) \tag{1}$$

$$\text{s.t.: } ||\delta||_2 \leq \epsilon \tag{2}$$

$$\delta \in \mathcal{I} - X \tag{3}$$

The objective is either $T$ or $f(\delta)$; when equal to the former, the program stops, while when it is equal to the latter, i.e., $f(\delta) \leq T$, we perform gradient ascent in search for a better solution. To ensure that the perturbation is as small as possible, we start with a very small admissible perturbation $\epsilon$ and gradually increase it when the above program fails to find a solution within a reasonable amount of steps.

Our black-box methodology is based on Natural Evolutionary Strategy (NES) [19, 22] and is similar to Ilyas et al. [10]. We use NES techniques to estimate the gradient of our objective function. We update the perturbation $\delta$ with the sign of the gradient as used previously in adversarial attacks against ML classifiers [8]. To ensure constraints are satisfied, we use projected gradient ascent [12].

**White-box attack.** We also develop a principled attack to devise a minimum perturbation for DCT-based perceptual hashing algorithms, such as pHash and Facebook's PDQ [11]. The Discrete Cosine Transform (DCT) [3] is a very popular image compression algorithm. Our attack exploits the linearity and orthonormality of DCT. By using the Euclidean distance to measure both the input perturbation and the distance between original and modified hashes, we show that minimal perturbations can be found as linear combinations of eigenvectors derived from the linear transform. We further show that the minimal perturbation needed to exceed threshold $T$ is equal to $T$ exactly.

## 4 Results

We perform a large-scale extensive evaluation on the ImageNet dataset [18] of five commonly used perceptual hashing algorithms: pHash (continuous and discrete) [7, 23], dHash, aHash and PDQ [11].

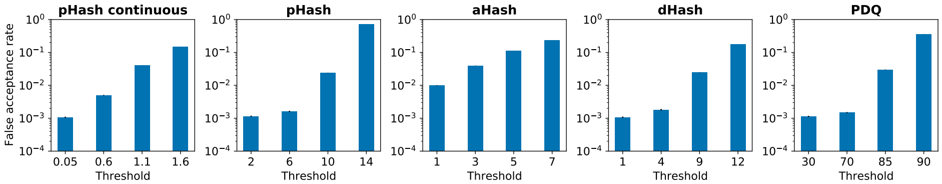

Figure 3: False Positive Rate (FPR) of a detection system for different perceptual hashing algorithms and thresholds. The database size is $N = 100K$, the number of images used to estimate the FPR is $M = 500K$ and the number of attacked images is $N' = 1K$.

We show that our black-box attack manages to successfully attack all images ($N = 10K$) for all perceptual hashing algorithms but one, reaching an attack effectiveness of 99.9%.

Fig. 1 shows examples of images successfully perturbed using our attack along with the threshold used and resulting $\mathbb{L}_2$ perturbation per pixel. The modified image always preserves the visual content of the original image with PDQ seemingly requiring the most visually perceptible modifications. Fig. 2 further shows that a perturbation with small $\mathbb{L}_2$ is enough to successfully attack most images. We obtain similar results for the commonly used LPIPS distance [24].

We further study the robustness of a detection system against our black-box attack for a wide variety of thresholds, and find that our modified images almost always avoid detection. For instance, we observe that for database of size $100K$, $100\%$ and $98.1\%$ of adversarially modified images evaded detection by PDQ algorithm for the recommended thresholds $T = 30$ and $T = 85$ respectively.

We study several mitigation strategies against our attack and show them to be ineffective. First, we observe that larger thresholds slightly reduce the number of modified images that evade detection, but as we show in Fig. 3 the false positive rates become very large. We further estimate that for a small prevalence rate of illegal content ($p = 10^{-4}$), the number of images wrongly flagged daily could vary from a few millions to more than 1B images as the threshold increases.

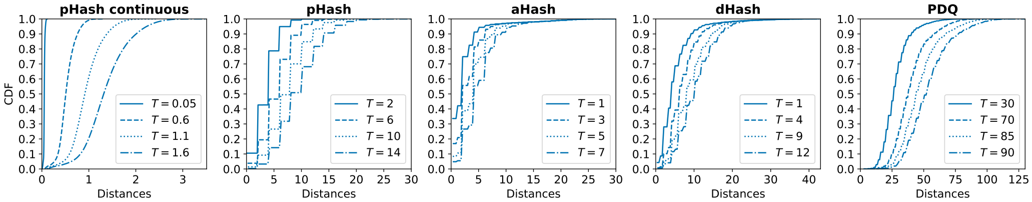

Figure 4: Pairwise distances between hashes of the multiple modified images generated for the same image, for different algorithms and thresholds. $D = 50$ modified images are generated for each ($N' = 100$) original image.

Second, we modify our attack to generate different perturbations resulting in diverse output hashes in each run of the attack. Fig. 4 shows that measures such as expanding the database with hashes of adversarially modified illegal images would prove to be ineffective. Indeed, the attacker could still generate a perturbation that might evade detection.

We analyze the optimality of our black-box attack in terms of finding the minimum perturbation. We compare results with the white-box attack against the DCT algorithm. We show that more than 95% of perturbations from our black-box attack are within twice the optimal limit (in $\mathbb{L}_2$ norm).

## 5 Conclusion

Taken together, our results strongly suggest that perceptual hashing-based client-side scanning is highly vulnerable to small modifications in all scenarios considered, and that simple mitigation strategies like increasing the threshold or expanding the database would be ineffective. This sheds serious doubts on the robustness of currently proposed client-side scanning mechanisms based on perceptual hashing [1, 9, 6, 17, 13].

## References

[1] Csam detection - technical summary 2021.

[2] Online disinformation | shaping europe's digital future.

[3] N Ahmed, T Natarajan, and K R Rao. Discrete cosine transform. *IEEE TRANSACTIONS ON COMPUTERS*, page 4, 1974.

[4] Whatsapp Blog. Two billion users – connecting the world privately. `https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately/`, 2020. Accessed on June 6, 2021.

[5] Jon Callas. Thoughts on mitigating abuse in an end-to-end world. page 16, Jan 2020.

[6] European Commission. Technical solutions to detect child sexual abuse in end-to-end encryption communications, 2020.

[7] B. Coskun and B. Sankur. Robust video hash extraction. In *Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004.*, page 292–295, Apr 2004.

[8] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[9] Himanshu Gupta and Harsh Taneja. Whatsapp has a fake news problem—that can be fixed without breaking encryption. `https://www.cjr.org/tow_center/whatsapp-doesnt-have-to-break-encryption-to-beat-fake-news.php`, Aug 2018. Accessed on June 6, 2021.

[10] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, page 2137–2146. PMLR, Jul 2018.

[11] John Kerl. The TMK+PDQF video-hashing algorithim and the PDQ image hashing algorithm. `https://github.com/facebook/ThreatExchange/blob/master/hashing/hashing.pdf`, 2020. Accessed on June 7, 2021.

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[13] Jonathan Mayer. Content moderation for end-to-end encrypted messaging. *Princeton University*, 2019.

[14] NCMEC. Ncmec's statement regarding end-to-end encryption. `https://www.missingkids.org/blog/2019/post-update/end-to-end-encryption`, Mar 2019. Accessed on June 6, 2021.

[15] Home Office. Interim code of practice on online child sexual exploitation and abuse (accessible version). `https://www.gov.uk/government/publications/online-harms-interim-codes-of-practice/interim-code-of-practice-on-online-child-sexual-exploitation-and-abuse-accessible-version`, Dec 2020. Accessed on June 6, 2021.

[16] Priti Patel, William Barr, Peter Dutton, Andrew Little, Bill Blair, India, and Japan. Internation statement: End-to-end encryption and public safety. `https://www.gov.uk/government/publications/international-statement-end-to-end-encryption-and-public-safety`, Oct 2020. Accessed on June 6, 2021.

[17] Julio CS Reis, Philipe Melo, Kiran Garimella, and Fabrício Benevenuto. Can whatsapp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Kennedy School Misinformation Review*, 2020.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[19] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv:1703.03864 [cs, stat]*, Sep 2017.

[20] Manish Singh. Whatsapp is now delivering roughly 100 billion messages a day. `https://social.techcrunch.com/2020/10/29/whatsapp-is-now-delivering-roughly-100-billion-messages-a-day/`, Oct 2020. Accessed on June 6, 2021.

[21] Manish Singh. Signal's brian acton talks about exploding growth, monetization and whatsapp data-sharing outrage. `https://social.techcrunch.com/2021/01/12/signal-brian-acton-talks-about-exploding-growth-monetization-and-whatsapp-data-sharing-outrage/`, Jan 2021. Accessed on June 6, 2021.

[22] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, Jan 2014.

[23] Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. Master's thesis, 2010. `https://www.phash.org/docs/pubs/thesis_zauner.pdf`.

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.