https://doi.org/10.2352/EI.2024.36.4.MWSF-339

Prediction of Flipped Bits in Robust Image Hashes by Machine Learning

Niklas Bunzel; Fraunhofer SIT / ATHENE / TU-Darmstadt; Darmstadt, Germany
Martin Steinebach; Fraunhofer SIT / ATHENE; Darmstadt, Germany
Marius Leon Hammann; TU-Darmstadt; Darmstadt, Germany
Huajian Liu; Fraunhofer SIT / ATHENE; Darmstadt, Germany

Abstract

There are advantages and disadvantages to both robust and cryptographic hash methods. Integrating the qualities of robustness and cryptographic confidentiality would be highly desirable. However, the challenge is that the concept of similarity is not applicable to cryptographic hashes, preventing direct comparison between robust and cryptographic hashes. Therefore, when incorporating robust hashes into cryptographic hashes, it becomes essential to develop methods that effectively capture the intrinsic properties of robust hashes without compromising their robustness. In order to accomplish this, it is necessary to anticipate the hash bits that are most susceptible to modification, such as those that are affected by JPEG compression. Our work demonstrates that the prediction accuracy of existing approaches can be significantly improved by using a new hybrid hash comparison strategy.

Motivation

For copyright protection and detection of known illegal digital images, robust image hashing can be used. Privacy is an important concern, especially when private images are part of a forensic investigation. Mobile phones, computers, and other devices store many private pictures whose privacy needs to be protected. A suspect may not have any illegal images at all, so his or her privacy must not be compromised. Robust image hashes are very effective at detecting known illegal images. However, they leak information about the original image and cannot be considered privacy preserving. Steinebach et al. propose to combine robust hashes with cryptographic hash functions in a hybrid approach to avoid such information leakage [1].

One property of cryptographic hash functions is the avalanche effect. A small change in a cryptographic hash function's input produces a hash value drastically different from and uncorrelated with the original. As a result, distance metrics such as the Hamming Distance (HD), which are used to compare robust hashes, cannot be applied to cryptographic hashes. As a result, an image must always produce the exact same robust hash, even when attacks such as JPEG compression or rescaling are applied. Otherwise, the cryptographic hashes will not match. (Note: The operations are referred to as "attacks" in this context because they potentially change the hash. These "attacks" are also commonly used in image live-cycles, such as automated normalization on social media platforms). To do this, Steinebach et al. propose the identification of weak bits of robust hashes, which can be neutralized before a cryptographic hash is applied. In this work, we improve existing approaches for predicting weak bits. We also propose and evaluate new prediction approaches based on machine learning.

Predicting flip positions is essential for combining robust and cryptographic hashes. This topic is covered in previous works [2] [1]. There the flipping positions are predicted by their distance to the block median value. In [3] it is shown that this assumption is not reliable enough for effective prediction. In [4] machine learning is used to significantly improve the chances of correct prediction. A prediction will allow to combine robust and cryptographic hashes. In our work we show that the prediction accuracy of existing approaches can be significantly improved by using a new hybrid hash comparison strategy. This will provide superior privacy when identifying images.

Background

Hash-based algorithms are used in various applications, such as image searching, detecting duplicates or near-duplicates, or authenticating images [5] [6] [7] [8]. Hash functions can be divided into cryptographic and robust hashes. They are used for different purposes and have different characteristics.

A cryptographic hash function is designed to provide data integrity, authenticity, and non-repudiation. Its primary purpose is to generate a hash value of a fixed size, or a digest of a message, from an input of any size. Some common cryptographic hash functions include SHA-256 (Secure Hash Algorithm 256-bit), MD5 (Message Digest Algorithm 5), and SHA-3.They are commonly used in various security applications such as password hashing, digital signatures, message integrity checking, and key derivation.

Characteristics of cryptographic hash functions include:

- Deterministic: The same input always produces the same output.
- Fast computation: The hash function should produce the hash value efficiently.
- Resistant to forgery: It is computationally infeasible to find the original input from the hash value.
- Collision resistance: It is computationally infeasible to find two different inputs that produce the same hash value.
- Small changes in the input lead to significant changes in the output (avalanche effect).
- Pseudorandomness: The output should appear random, even if the input has a predictable pattern.

For perceptually similar images, robust image hash functions produce a unique bit string. They operate on an image's perceptual features, not the image file's binary representation. Therefore, they are robust to changes in individual bits. As long as the changes are not perceptually noticeable. This robustness applies to intentional and unintentional image modifications. These can result from malicious attempts to prevent re-identification of an image, or from operations such as compression and scaling, which are often used to reduce the size of an image file during transmission.

Properties of robust hash functions include:

- Deterministic: Like cryptographic hash functions, robust hash functions should produce the same output for the same input.
- Efficient: Robust hash functions are optimized for performance and computational efficiency.
- Uniform distribution: Hash values should be uniformly distributed over the output space.
- Lower collision resistance: While robust hash functions aim for minimal collisions, they may not have the same level of collision resistance as cryptographic hash functions.

Based on machine learning, new attack strategies including preimage and avoidance attacks have been introduced [9] [10] [11]. They show that targeted attacks against robust hashes are possible but still introduce a significant quality loss. It should also be noted that attacks on robust hashing have already been discussed before the advance of machine learning [12] [13].

Perceptual features of images are used in many different robust hash functions [14, 15, 16, 17]. In this work, we base our implementation on the block mean value based perceptual image hash function [18] proposed by Yang et al. in its simplest form:

- Normalize the original image into a preset size and convert it into greyscale.
- Partition the resulting image I into non-overlapping blocks $I_1, I_2, ..., I_N$ where N is the targeted length of the hash bit string.
- Permute the block sequence $\{I_1, ..., I_N\}$ based on a secret key.
- Calculate the mean pixel value of M_i of each block I_i and determine the mean value of this sequence

$$M_d = \text{median}(M_i),\ \forall i \in \{1, 2, ..., N\} \tag{1}$$

- Obtain binary hash value by concatenating the individual hash bits:

$$h(i) = \begin{cases} 0, & M_i < M_d,\ \forall i \in \{1, 2, ..., N\} \\ 1, & M_i \geq M_d,\ \forall i \in \{1, 2, ..., N\} \end{cases} \tag{2}$$

The robust hash applied in this work is the ForBild block hash presented by [19]. It is the result of an evaluation of image hashing methods [20].

Robust Hashing and Privacy

For some time, there has been a discussion whether an image can be derived from its robust hash. While this seems to be unrealistic in the sense that a high quality image is reproduced, the structure-preserving nature of many robust hashes may allow to reproduce a general idea of the image content. Recent works based on machine learning show some progress in this respect[1] . It is still not clear if the reproduced content depends on the training data or only on the hash. Still, this developments makes it even more important to investigate methods to support the privacy of robust hashing.

There are also other approaches to combine privacy and robust hashing. Concepts have been discussed to use secure matching protocols able to compute the hamming distance of two binary strings [21] [22]. Here the hashes are not anonymized, but their comparison is executed in a privacy-preserving manner. This could be a challenge for scenarios where robust hashing is applied as it requires multiple parties that may not efficiently established in e.g. an upload filter.

Hybrid Hash

Robust hashes can be matched not only if they are identical. They can also be similar. This similarity is measured by the Hamming distance. Therefore, small image changes, e.g. JPEG compression or scaling, result in a robust hash still similar to the original robust hash. However, depending on how much the image is attacked, individual hash bits of these attacked images change their value. This bit-flipping behavior, analyzed by Steinebach et al., renders cryptographic hashes of robust hashes ineffective due to the avalanche effect [3].

Before a cryptographic hash function can be applied, weak bits must be predicted and neutralized. Steinebach et al. use the normalized distance between the pixel value of a block and the normalized median of an image to predict such bits. Combining robust hashing, neutralizing weak bits, and a cryptographic hash function is called Hybrid Hash. In this work, we further analyze bit flipping during robust image hashing to improve existing heuristic predictions. We also propose machine learning approaches to predict weak bits.

Concept

Our aim is to predict the bits of a robust hash that are likely to flip due to an operation the hash should be robust against. For robust hashing, flipping of individual bits is not a problem as they define similarity between hashes as two hashes having less different bits than a given threshold when they use Hamming distance as a measure. For hybrid hashes flipping bits is an important challenge as the robust hash is converted into a cryptographic hash and thereby loses its potential to be compared with the other hash in a similar manner. Machine learning can help to predict the flipping positions. We see this as a classification problem: a hash bit is either classified as stable or as likely to flip. As supervised learning is best suited for classification tasks, and we can generate labeled training data on demand by attacking images and comparing their hashes with the original ones, we use supervised learning to train our classifiers.

Two strategies can be followed to re-identify hybrid hashes: double prediction [4] and single prediction. We consider a database of known images K and a set of unknown images S that is being examined.

During double prediction, the database containing hybrid hashes of known images $H(K)$ is computed by predicting weak bits for every image $k \in K$. For higher accuracy, this should be done for multiple attack types like different JPEG quality factors (qf). These weak bits are then neutralized, and the hybrid hashes are computed. The same approach can be applied for every image

[1] https://www.anishathalye.com/2021/12/20/inverting-photodna/

$s \in S$ in order to compute the hybrid hashes database $H(S)$ for examined images. Therefore double prediction can be implemented with negligible performance overhead compared to block hash matching. The performance can vary based on the chosen predictor.
During single prediction, weak bits are only predicted for either K (Original Prediction) or S (Suspect Prediction). Predicted weak bits $P(X)$ are stored in addition to the resulting hybrid hashes $H(X)$ for the chosen image database $X \in \{K,S\}$. Images y from the other database $Y \in \{K,S\}, Y \neq X$ are not fed into a prediction function. For every image y:

1. the neutralized robust hash $N_p(y)$ is computed for every $p \in P(X)$
2. the hybrid hash $H_p(N_p(y))$ is computed

and the resulting hybrid hashes are stored. This introduces a significant performance and memory overhead compared to double prediction and block hash matching. On the other hand, single predictions yield a higher recall score than double predictions with a comparable precision score. In this work, we focus on single prediction strategies.

K-Nearest Neighbors

We implement a single-label classifier that classifies robust hash blocks individually based on the observed flipping likelihood corresponding to [4]. Each block can either flip (encoded as 1) or not flip (encoded as 0). The KNN classifier predicts a class label P_N for every block B_N, for $N = 1,...,256$. We also compare the results to a classifier that considers the JPEG quality factor (qf) as a third feature. The qf is approximated using the dc coefficient of each image.

As most blocks of a robust hash do not flip, most samples belong to class 0, which causes the dataset to be imbalanced. This is fixed by re-sampling the training data so that both classes are represented equally.

Single-Label Classification

In accordance with [4] we use a Deep Neural Network (DNN) with three hidden layers, the Adam optimizer and binary cross-entropy as loss function. The results are superior to all previous predictions in both recall and overprediction.

Again, we seek to improve both by using the qf as an additional feature. The predictions using the qf achieve a nearly perfect recall score and for qf > 10 a significantly lower overprediction.

Evaluation

For the evaluation of our single prediction re-identification approach we use 2000 randomly selected images of a cheerleading team from the galaxy data set [23]. The images in this data set show:

- Various amounts of humans
- Humans in various poses
- Humans of various appearance
- Humans in various environments

They were taken with different cameras and different resolutions. For our evaluation, we randomly selected the images and divided them into two data sets of 1,000 images each. One of them contains known images and the other unknown images.

The following attacks are performed on all images:

- JPEG compression with qf $\in \{90, 80, ..., 10\}$
- Scaling with scaling factors $\in \{0.5, 0.75, 0.9, 1.1, 1.5\}$

They are likely to be the result of unintentional manipulation. This often occurs when an image is transmitted or uploaded but exceeds the maximum file size.
Because the data set does contain several attacked images and the respective original images, double compression and double scaling are performed implicitly but not evaluated explicitly.

We compare both machine-learning approaches to the relative distance approach utilized in the literature [2]. Here the prediction is based on the distance between the individual hash bit and the median used as a threshold for bit value assignment. The concept here is that the close a value is to the median, the more likely it is to flip due to an attack. To represent how many FPs have to be neutralized during cryptographic hashing are predicted per image we define Overprediction as

$$\text{Overprediction} = \frac{FP}{TP + FP + TN + FN} \quad (3)$$

Performance

In addition to the prediction quality discussed in the next section, we also had a look at the computation time of the strategies. To compare the performance of the previously described approaches, we measure the median time it takes to hash and predict 110 images. The measurement is given relative to the time it takes to compute a standard robust block hash. The evaluation is performed on an Intel(R) Xeon(R) CPU @2.00GHz CPU and with 52 GB memory. No GPU was used for machine learning optimization. Table 1 shows that both KNN and DNN are slower than the standard block hash. We evaluated both strategies regarding the complexity in Table 2. We can see that the single prediction strategy produces less predictions as the double prediction strategy and neutralizes more bits. Therefore the generated hybrid hashes tend to be more robust.

Hash	Factor
Block Hash	1
Relative Distance	1.01
KNN	1.29
DNN	1.59

Table 1: Performance of different prediction approaches.

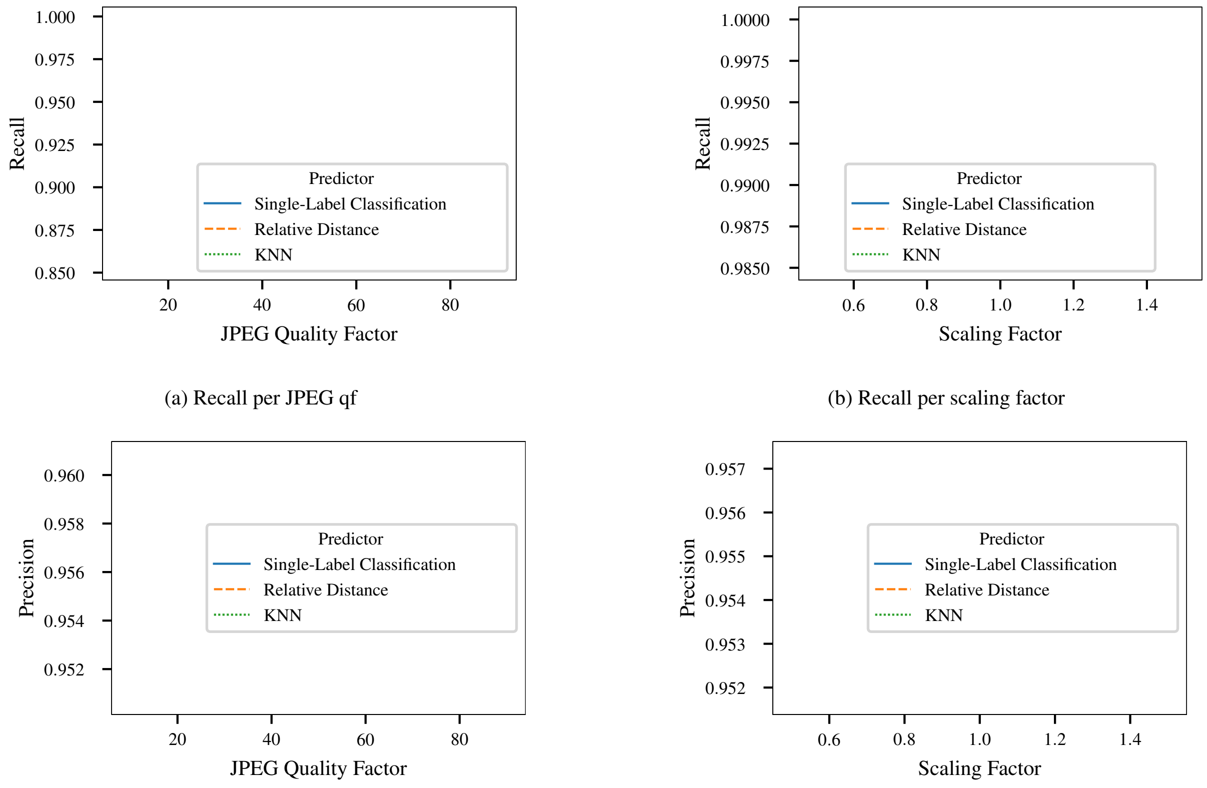
Prediction Strategy	Predictions	Neutralizations
Double Prediction	$O(n+m)$	$O(n+m)$
Single Prediction	$O(\min(n,m))$	$O(n \times m)$

$n \in K$and $m \in S$

Table 2: Performance of different prediction approaches.

Results

The inclusion of the JPEG quality factor improves the performance of the classifiers significantly as stated in [4], but the whole re-identification process this might not be true for all prediction strategies. For our new single prediction strategy the inde-



(a) Recall per JPEG qf

(b) Recall per scaling factor

(c) Precision per JPEG qf

(d) Precision per scaling factor

Figure 1: Recall and precision for qf-independent single prediction re-identification. Single Label Classification is the DNN approach

pendence of the quality factors can be advantageous for example for the recall against a scaling attack.

The results for single prediction, shown in figures 1 and 2, show a high recall score between 0.85 and 1. Single-label classification achieves the highest recall score for JPEG compression and decent results for scaling. The precision remains lower than that of Steinebach et al. and Breidenbach et al. but is still reasonably high at 0.95 and higher.

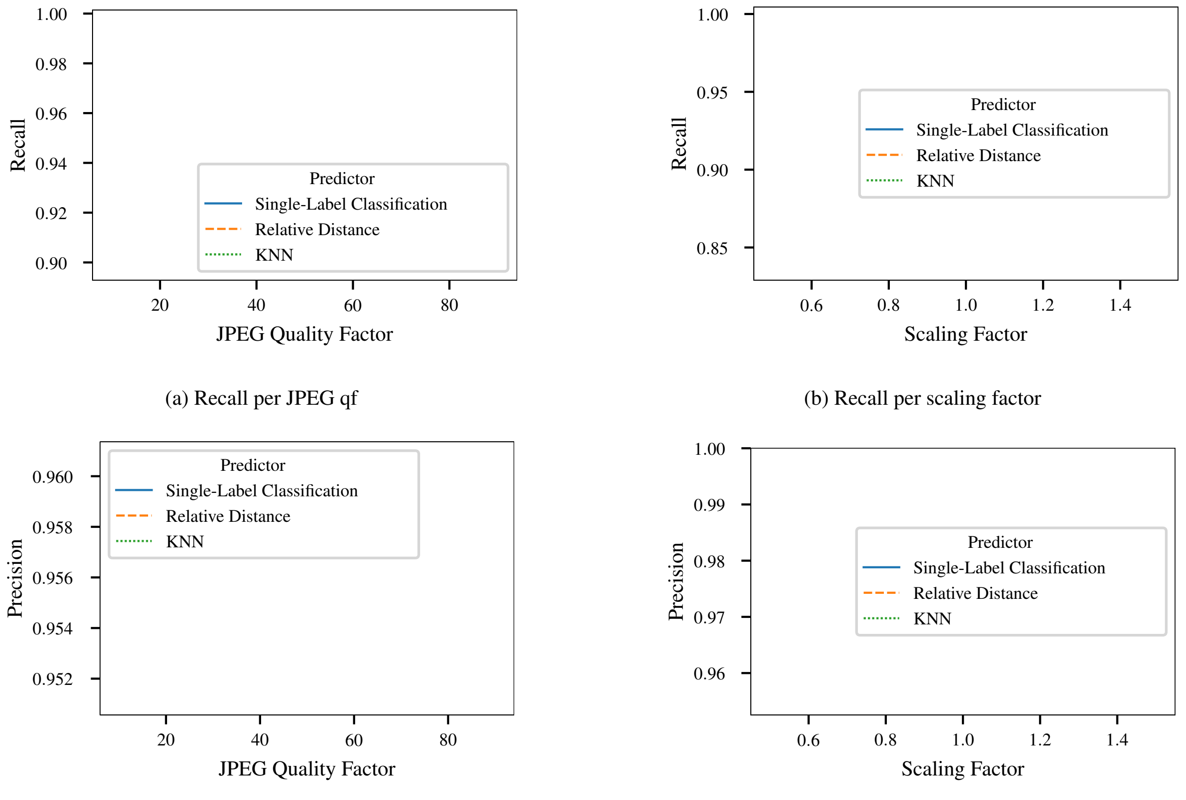
Overall, the results show that we can predict weak bits in an image. However, predictions between an original and an attacked image differ marginally, which results in false negatives. Furthermore, single-label classification yields the best recall results irrespective of the prediction strategy used. The KNN classifier, however, consistently results in the lowest recall score of all predictors.

We compared our new single prediction approach against the double prediction approach in Figures 3 and 4 for quality factor dependent and independent variants. We can see that our single prediction approach performs significantly better in regards to recall and slightly worse in regards to precision. The low recall of the double prediction approach caused by the avalanche effect is not desirable. A combined approach can be used where double prediction is performed for an initial re-identification. The performance-intensive re-identification using single prediction can be run afterward while the initial results are analyzed. As a result, initial results are available quickly, while more accurate results are being determined. The hybrid hashes obtained during double prediction re-identification can be used to improve the performance of the single prediction re-identification.

We also can see that the approaches are robust against a certain amount of scaling. Figure 4 (b) and (d) show no significant change of performance depending on the scaling factor. For JPEG compression, single prediction is also very stable with respect to recall as can be seen in figure 4 (a). Precision under JPEG compression seems to be the most unstable prediction. In Figure 4 (c) we can see that only KNN double prediction works equally well independently of the QF. KNN single prediction on the other hand shows significant changes that do not follow the expectation of improving performance with increasing QF.

Summary

Our prediction algorithms use image properties to predict weak bits in robust hashes with high recall. In particular, the DNN classifier achieves high recall combined with low overprediction. Approximating the qf of an image increases recall at low qf and decreases overprediction at high qf. The flipped bit prediction rate of machine learning-based predictors is superior to that of heuristic approaches when using the approximated qf as a feature. Overall, we show that given an image, we can predict weak bits with a recall close to one and an overprediction of less than 10%. Our new single prediction strategy outperforms the previously proposed double prediction strategy for all classifiers under scaling and JPEG compression attacks in terms of recall for



(a) Recall per JPEG qf

(b) Recall per scaling factor

(c) Precision per JPEG qf

(d) Precision per scaling factor

Figure 2: Recall and precision for qf-dependent single prediction re-identification. Single Label Classification is the DNN approach

quality dependent and independent variants of the classifiers. In regards to precision the single prediciton strategy is slightly inferior. In terms of computation times the double prediction strategy out performs our single prediction.

Future Work

As mentioned in the concept section, both double and single prediction strategies are possible. We evaluated and compared both in this work, in conclusion our single prediction strategy outperforms the double prediction in recall and is about equal in precision. Therefore, having a better accuracy at the cost of having a much higher computation time. In future work, a combination of both strategies should be evaluated to combine their advantages. One idea would be to use double prediction as a fast pre-filter and single prediction as the final evaluation given a defined threshold result of dual prediction.

Acknowledgment

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Martin Steinebach, Sebastian Lutz, and Huajian Liu. Privacy and robust hashes. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–8, 2019.

[2] Uwe Breidenbach, Martin Steinebach, and Huajian Liu. Privacy-enhanced robust image hashing with bloom filters. In Melanie Volkamer and Christian Wressnegger, editors, *ARES 2020: The 15th International Conference on Availability, Reliability and Security, Virtual Event, Ireland, August 25-28, 2020*, pages 56:1–56:10, New York, NY, USA, 2020. ACM.

[3] Martin Steinebach. A close look at robust hash flip positions. *Electronic Imaging*, 2021(4):345–1, 2021.

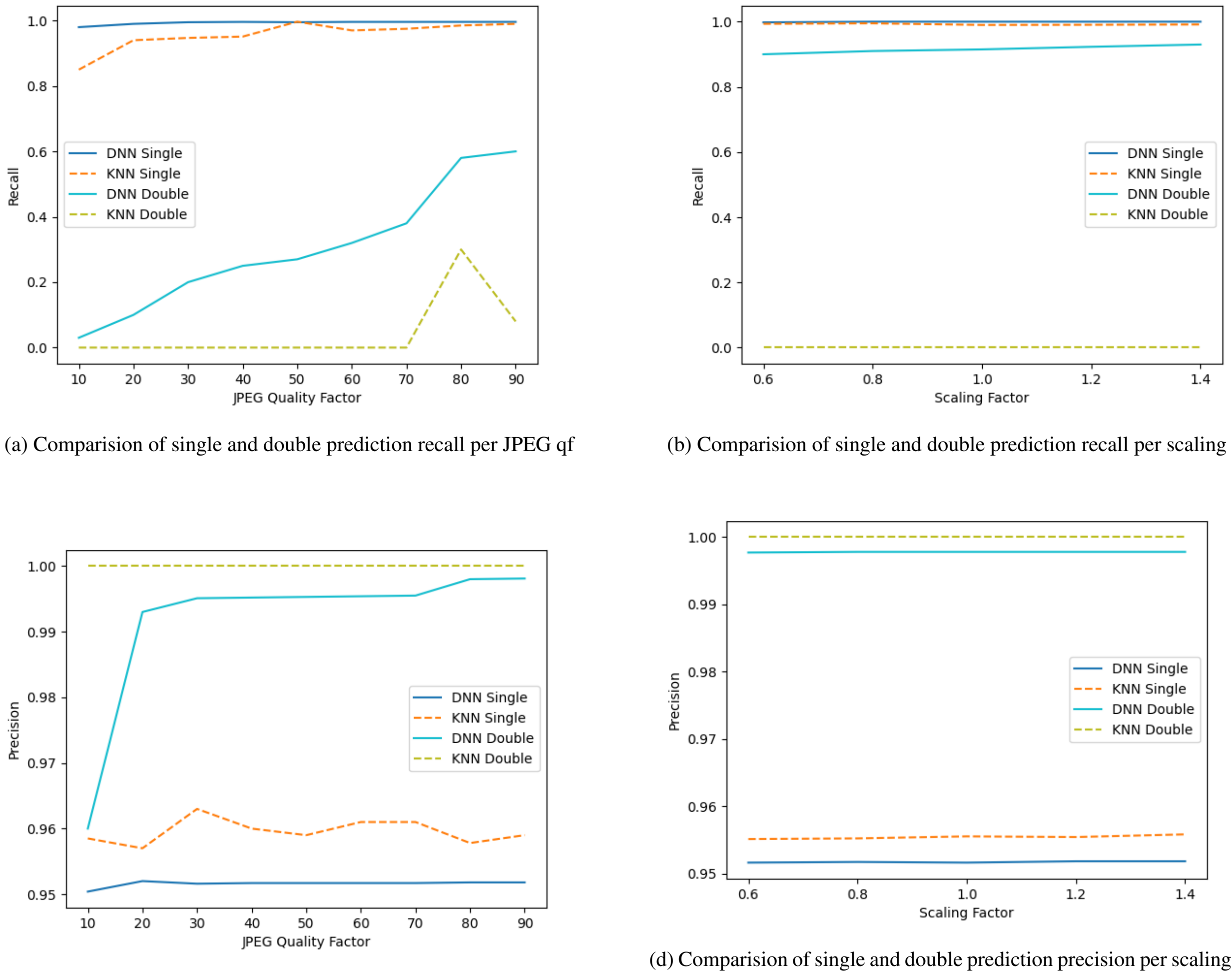
[4] Marius Leon Hammann, Martin Steinebach, Huajian Liu, and Niklas Bunzel. Predicting positions of flipped bits in robust image hashes. *Electronic Imaging*, 35:375–1, 2023.

[5] Andrea Drmic, Marin Silic, Goran Delac, Klemo Vladimir, and Adrian S. Kurdija. Evaluating robustness of perceptual image hashing algorithms. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 995–1000. IEEE, 2017.

[6] Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran. Automatic image filtering on social networks using deep learning and perceptual hashing during crises.

[7] Ling Du, Anthony T.S. Ho, and Runmin Cong. Perceptual hashing for image authentication: A survey. *Signal Processing: Image Communication*, 81:115713, 2020.

[8] Martin Steinebach, Huajian Liu, and York Yannikos. Efficient cropping-resistant robust image hashing. In *2014 Ninth International Conference on Availability, Reliability and Security*, pages 579–585. IEEE, 2014.



(a) Comparision of single and double prediction recall per JPEG qf

(b) Comparision of single and double prediction recall per scaling factor

(c) Comparision of single and double prediction precision per JPEG qf

(d) Comparision of single and double prediction precision per scaling factor

Figure 3: Recall and precision for qf-independent single and double prediction re-identification

[9] John Prokos, Tushar M. Jois, Neil Fendley, R. Schuster, Matthew Green, Eran, Tromer, and Yinzhi Cao. Squint hard enough: Evaluating perceptual hashing with machine learning. *IACR Cryptol. ePrint Arch.*, 2021:1531, 2021.

[10] Shubham Jain, Ana-Maria Crețu, and Yves-Alexandre de Montjoye. Adversarial detection avoidance attacks: Evaluating the robustness of perceptual hashing-based client-side scanning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2317–2334, 2022.

[11] Lukas Struppek, Dominik Hintersdorf, Daniel Neider, and Kristian Kersting. Learning to break deep perceptual hashing: The use case neuralhash. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 58–69, 2022.

[12] Li Weng and Bart Preneel. Attacking some perceptual image hash algorithms. In *2007 IEEE International Conference on Multimedia and Expo*, pages 879–882. IEEE, 2007.

[13] Oleksiy Koval, Sviatoslav Voloshynovskiy, Patrick Bas, and François Cayre. On security threats for robust perceptual hashing. In *Media Forensics and Security*, volume 7254, pages 157–169. SPIE, 2009.

[14] Zhenjun Tang, Xianquan Zhang, Xuan Dai, Jianzhong Yang, and Tianxiu Wu. Robust image hash function using local color features. *AEU - International Journal of Electronics and Communications*, 67(8):717–722, 2013.

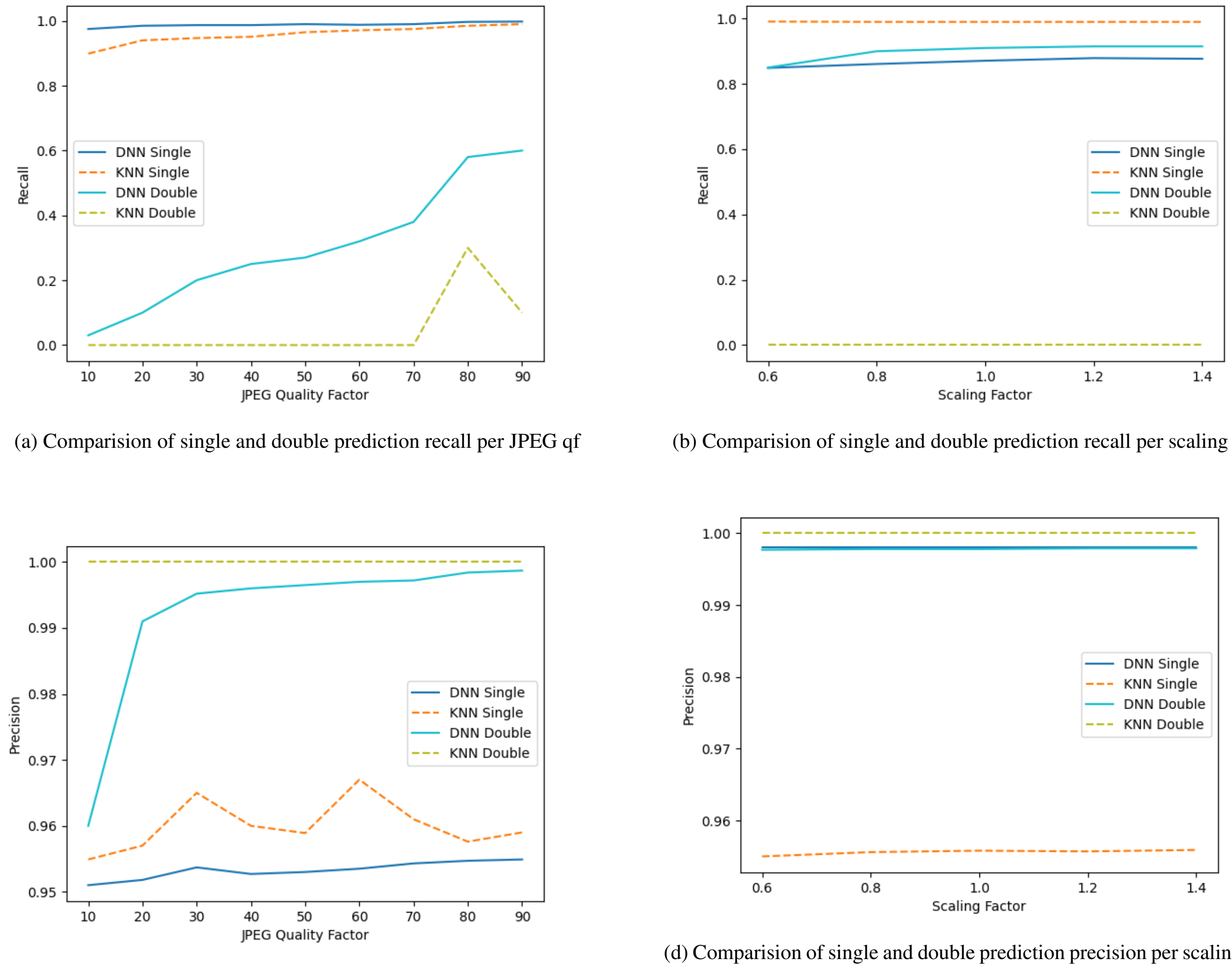
[15] Zhenjun Tang, Fan Yang, Liyan Huang, and Xianquan Zhang. Robust image hashing with dominant dct coefficients. *Optik*, 125(18):5102–5107, 2014.

[16] Zhenjun Tang, Lv Chen, Xianquan Zhang, and Shichao Zhang. Robust image hashing with tensor decomposition. *IEEE Transactions on Knowledge and Data Engineering*, 31(3):549–560, 2019.

[17] Rui Sun and Wenjun Zeng. Secure and robust image hashing via compressive sensing. *Multimedia Tools and Applications*, 70, 06 2012.

[18] Bian Yang, Fan Gu, and Xiamu Niu. Block mean value based image perceptual hashing. In *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia*, IIH-MSP '06, page 167–172, USA, 2006. IEEE Computer Society.

[19] Martin Steinebach. Robust hashing for efficient forensic analysis of image sets. In Pavel Gladyshev and Marcus K. Rogers, editors, *Digital Forensics and Cyber Crime*, volume 88 of *Lecture Notes of the*



(a) Comparision of single and double prediction recall per JPEG qf

(b) Comparision of single and double prediction recall per scaling factor

(c) Comparision of single and double prediction precision per JPEG qf

(d) Comparision of single and double prediction precision per scaling factor

Figure 4: Recall and precision for qf-dependent single and double prediction re-identification

Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 180–187. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[20] Christoph Zauner, Martin Steinebach, and Eckehard Hermann. Rihamark: perceptual image hash benchmarking. In Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp III, editors, *Media Watermarking, Security, and Forensics III*, SPIE Proceedings, page 78800X. SPIE, 2011.

[21] Ayman Jarrous and Benny Pinkas. Secure hamming distance based computation and its applications. In *Applied Cryptography and Network Security: 7th International Conference, ACNS 2009, Paris-Rocquencourt, France, June 2-5, 2009. Proceedings 7*, pages 107–124. Springer, 2009.

[22] Masaya Yasuda, Takeshi Shimoyama, Jun Kogure, Kazuhiro Yokoyama, and Takeshi Koshiba. Secure pattern matching using somewhat homomorphic encryption. In *Proceedings of the 2013 ACM workshop on Cloud computing security workshop*, pages 65–76, 2013.

[23] Martin Steinebach, Huajian Liu, and York Yannikos. Forbild: Efficient robust image hashing. In *Media Watermarking, Security, and Forensics 2012*, volume 8303, pages 195–202. SPIE, 2012.

Author Biography

Niklas Bunzel received his B.Sc. and M.Sc. degrees in computer science and IT security from Technical University Darmstadt 2015 and 2020, respectively. He is currently a PhD student at the TU-Darmstadt and a research scientist at Fraunhofer Institute for Secure Information Technology (SIT) and the National Research Centre for Applied Cybersecurity - ATHENE. His major research interests include artificial intelligence, IT security and steganography.

Prof. Dr. Martin Steinebach is the manager of the Media Security and IT Forensics division at Fraunhofer SIT. In 2003 he received his PhD at the Technical University of Darmstadt for this work on digital audio watermarking. In 2016 he became honorary professor at the TU Darmstadt.

Marius Hammann received his M.Sc. in IT Security and M.Sc. in Computer Science from TU Darmstadt in 2023. In his current position as IT Security Engineer at Aareon Group, he focuses on threat intelligence and digital forensics.

Huajian Liu received his B.S. and M.S. degrees in electronic engi-

neering from Dalian University of Technology, China, in 1999 and 2002, respectively, and his Ph.D. degree in computer science from Technical University Darmstadt, Germany, in 2008. He is currently a senior research scientist at Fraunhofer Institute for Secure Information Technology (SIT). His major research interests include information security, digital watermarking, robust hashing and digital forensics.