Squint Hard Enough

Attacking Perceptual Hashing with Adversarial Machine Learning

About Me

Jonathan Prokos - research@prokos.us

Masters in Security Informatics from JHU ‘22

Security Research Engineer at Two Six Technologies

> Focused on program analysis (binary RE/VR) + adversarial machine learning attacks/defenses

Authors

Jonathan Prokos - Johns Hopkins University, now Two Six Technologies

Neil Fendley - Johns Hopkins University Applied Physics Laboratory

Matthew Green - Johns Hopkins University

Roei Schuster - Vector Institute

Eran Tromer - Tel Aviv University and Columbia University

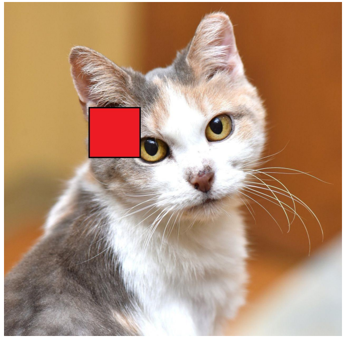
Tushar Jois - Johns Hopkins University, now City College of New York

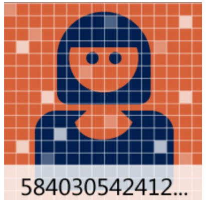
Yinzhi Cao - Johns Hopkins University

Background & Motivation

What is a Perceptual Hash Function (PHF)?

- Locality Sensitive
- Embeds image semantics

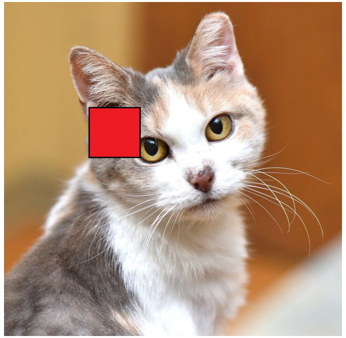


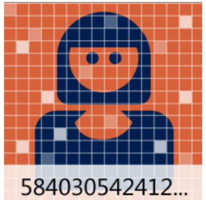


PHF	0b07008009...	0c07008409...	1519179f15...
SHA	57ead5f6f8...	97d071d6e6...	042a3db811...

What is a Perceptual Hash Function (PHF)?

- Locality Sensitive
- Embeds image semantics

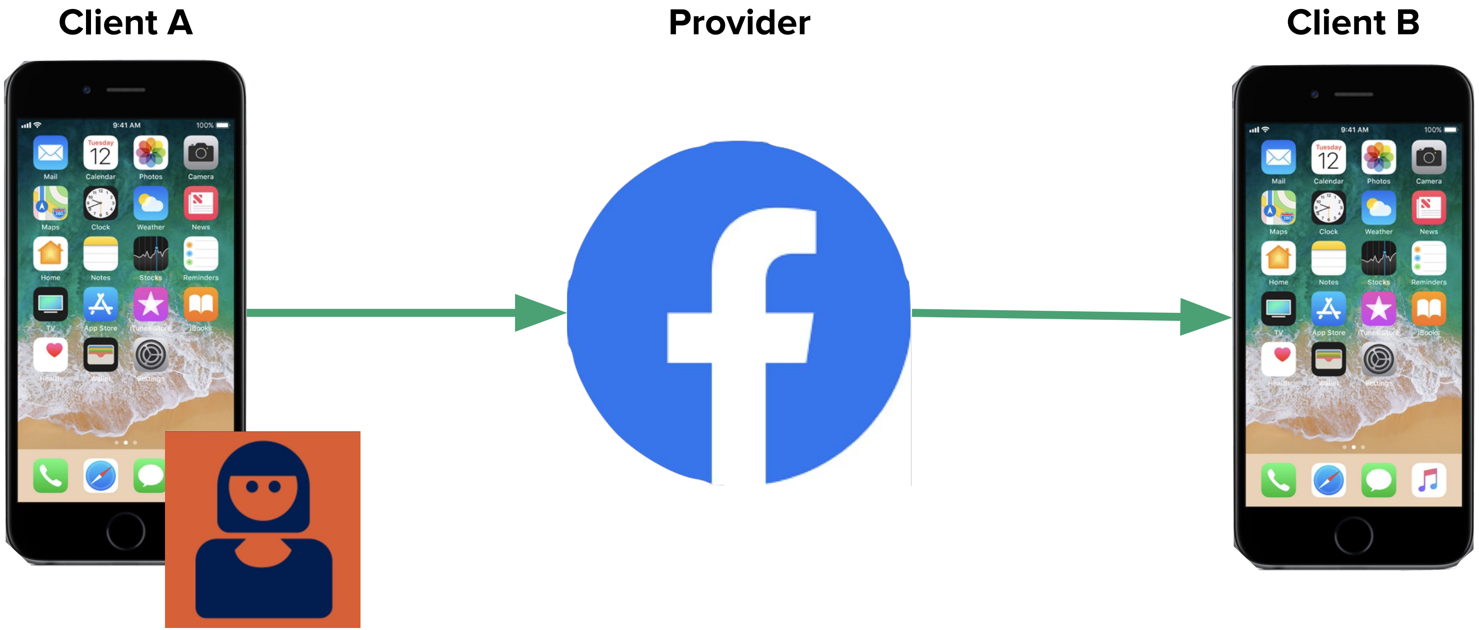




PHF	0b07008009...	0c07008409...	1519179f15...
SHA	57ead5f6f8...	97d071d6e6...	042a3db811...

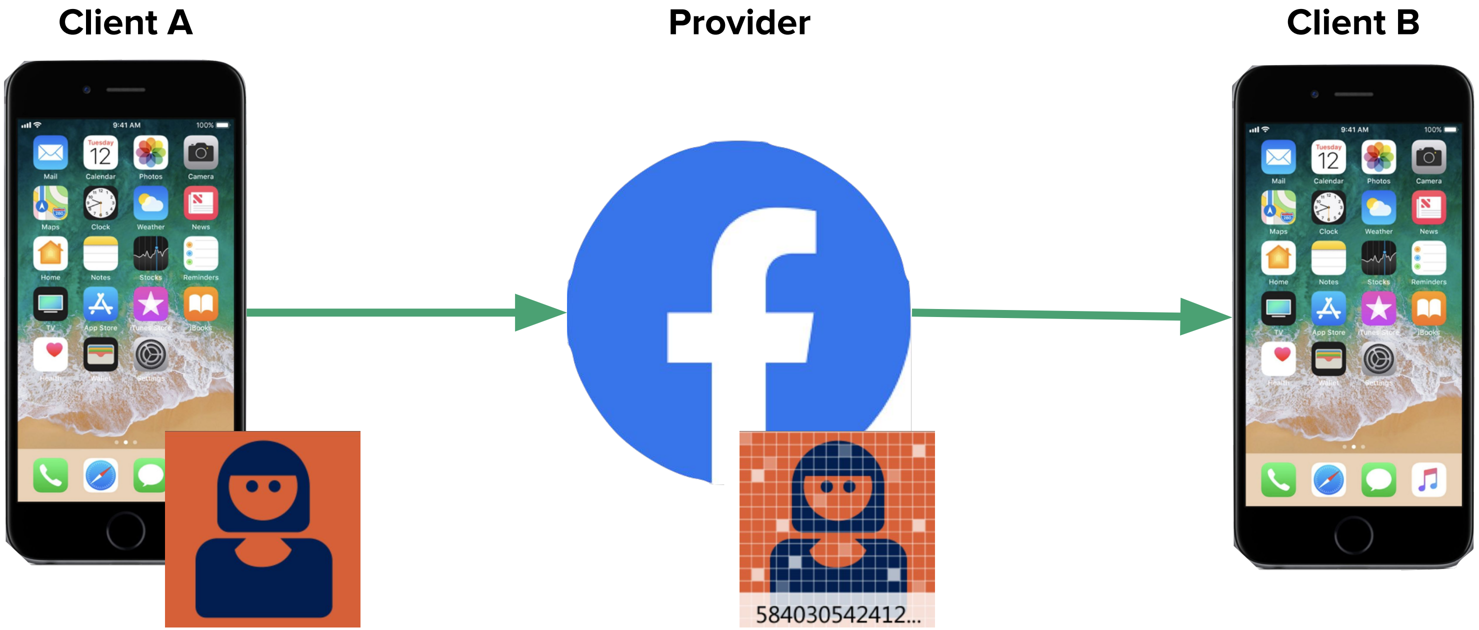
How Are PHFs Used?

How Are PHFs Used?



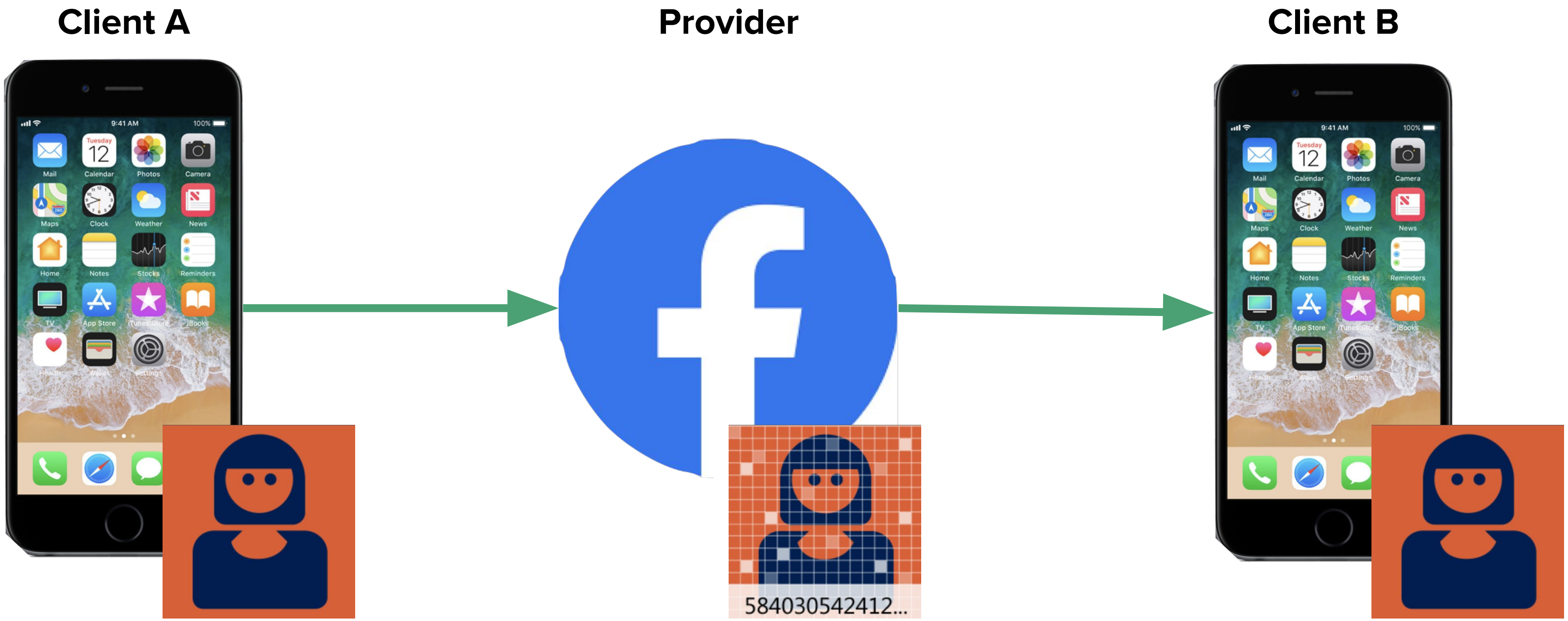
Client A sends image to FB

How Are PHFs Used?



FB uses PHF and hash corpus to check for illicit content

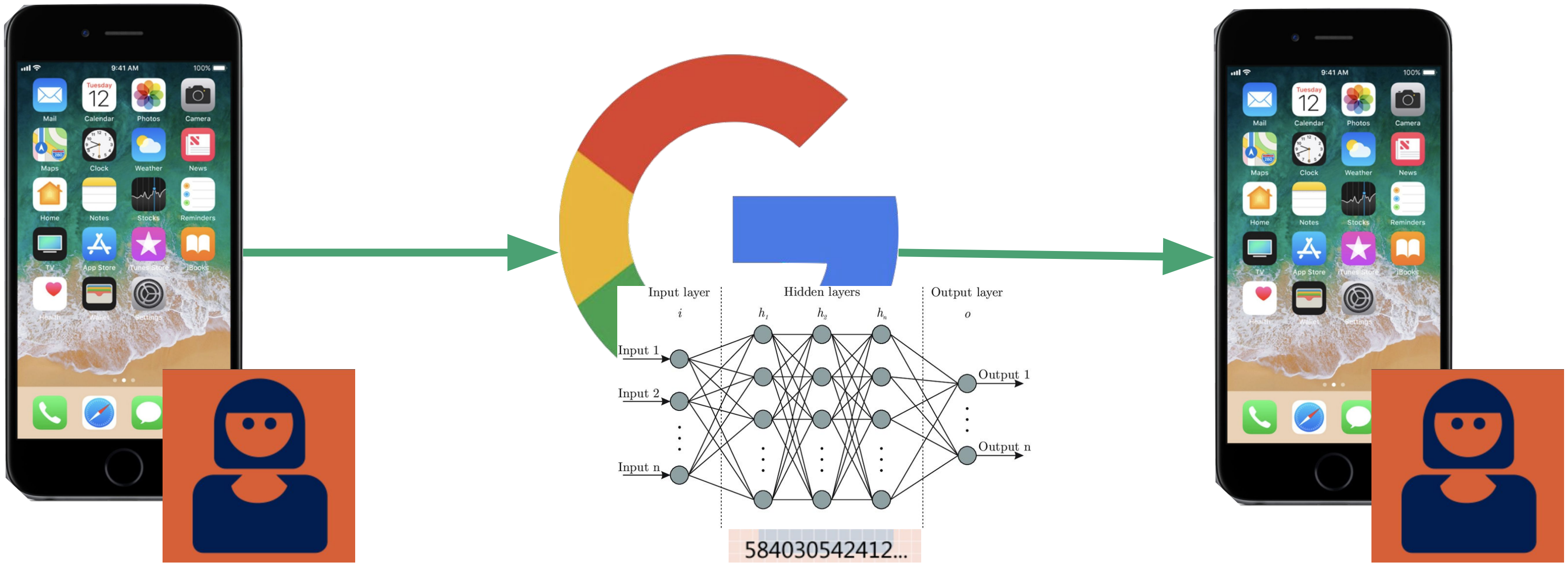
How Are PHFs Used?



If the image does not match, FB allows the image to be sent

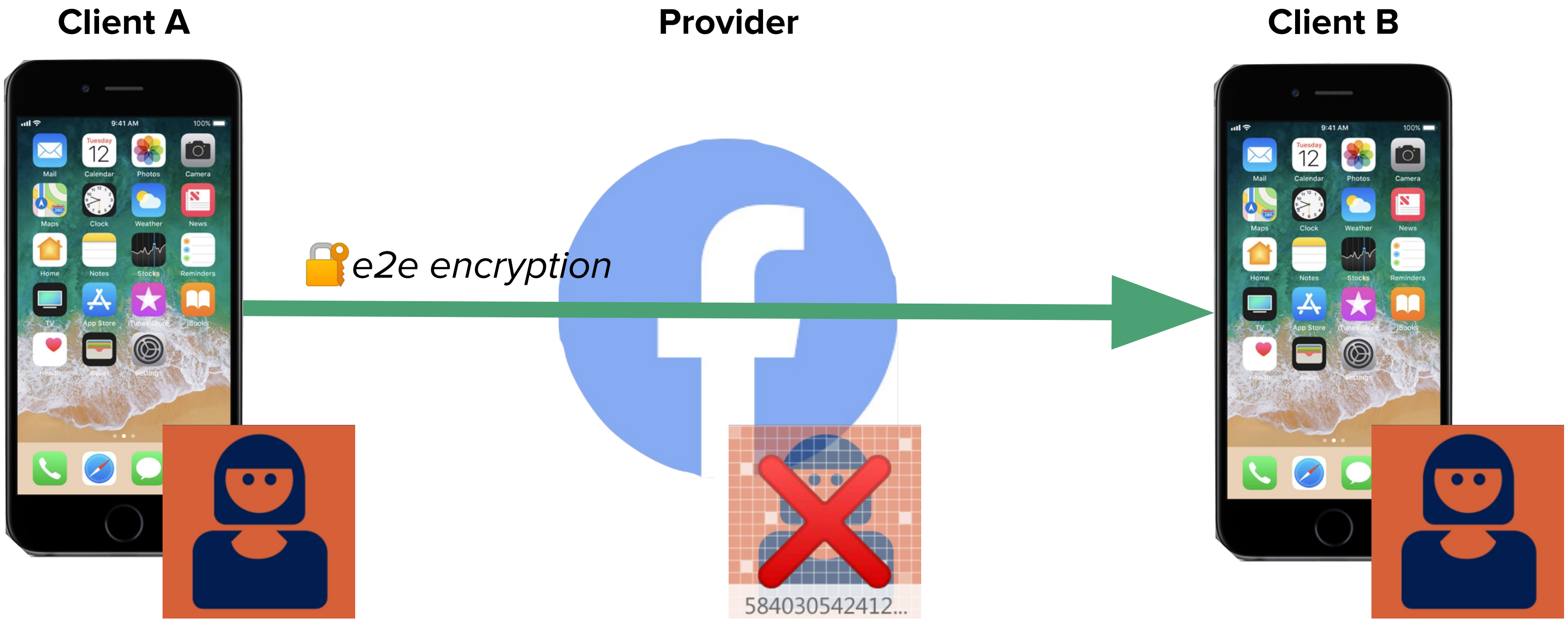
How Are PHFs Used?

Client A **Provider** **Client B**



Filter **<u>unkown</u>** illicit content using neural networks

End to End Encryption



2019: UK/US/AU (Barr) letter to Facebook

Dear Mr. Zuckerberg,

OPEN LETTER: FACEBOOK'S "PRIVACY FIRST" PROPOSALS

We are writing to request that Facebook does not proceed with its plan to implement end-to-end encryption across its messaging services without ensuring that there is no reduction to user safety and without including a means for lawful access to the content of communications to protect our citizens.

- Embed the safety of the public in system designs, thereby enabling you to continue to act against illegal content effectively with no reduction to safety, and facilitating the prosecution of offenders and safeguarding of victims;

We are committed to working with you to focus on reasonable proposals that will allow Facebook and our governments to protect your users and the public, while protecting their privacy. Our technical experts are confident that we can do so while defending cyber security and supporting technological innovation. We will take an open and balanced approach in line with the joint statement of principles signed by the governments of the US, UK, Australia, New Zealand, and Canada in August 2018[1] and the subsequent communique agreed in July this year[2].

Yours sincerely,

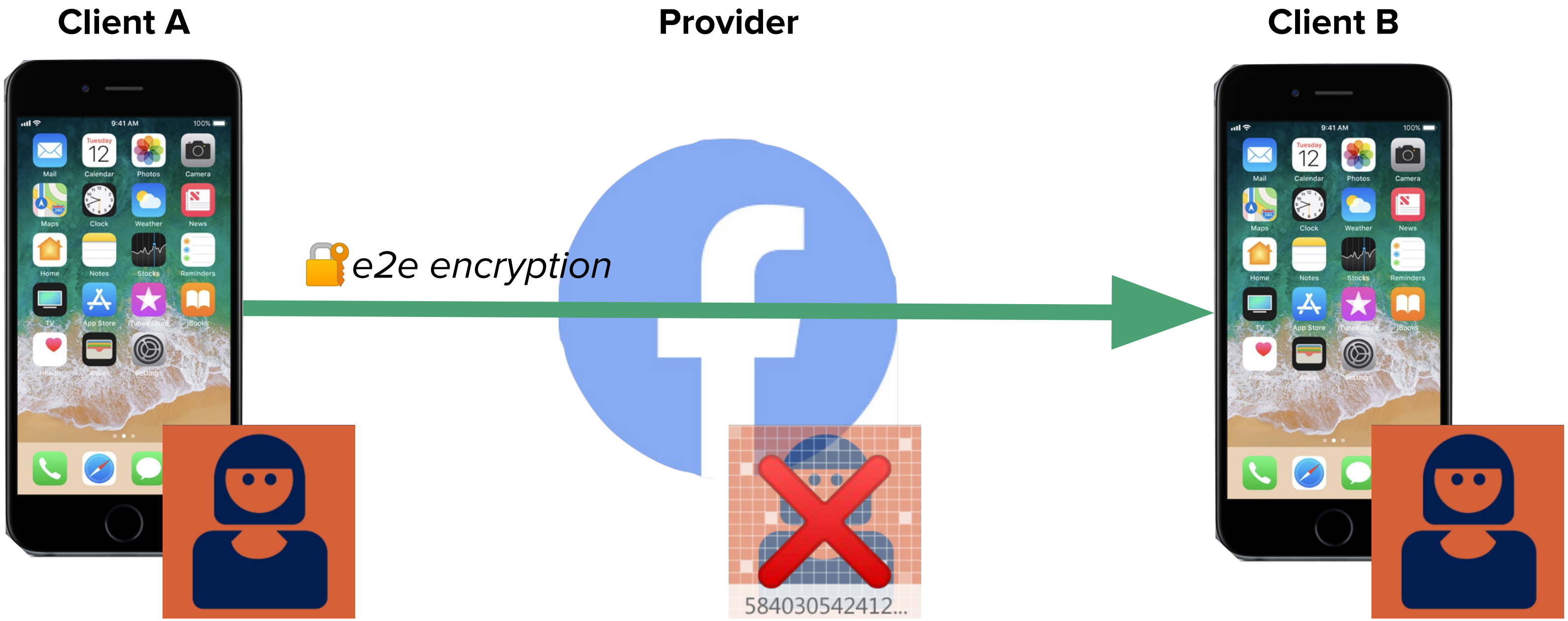
Rt Hon Priti Patel MP
United Kingdom Secretary of State for the Home Department

William P. Barr
United States Attorney General

Kevin K. McAleenan
United States Secretary of Homeland Security (Acting)

Hon Peter Dutton MP
Australian Minister for Home Affairs

End to End Encryption



Potential solution: move content filtering into the local client?

End to End Encryption - Client Side Scanning

When a detection happens, block & transmit image to server

Issues with Client Side Scanning

- Exposes hash database (or neural network weights) to attackers

Potential impacts

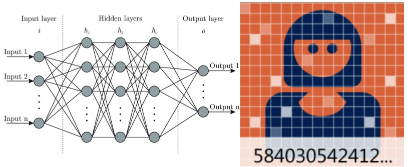
- Collision generation
  - Generating non-CSAM (Child Sexual Assault Material) media that triggers CSAM detection
- Detection avoidance
  - Altering CSAM media so it does not trigger CSAM detection
- Extract existing CSAM from database or generate novel (ML modeling)

Alternative Solutions - 2PC

Client A

Provider

two party computation



Use cryptography to split computation <u>privately</u>

Client has image, provider has algorithm/database/network

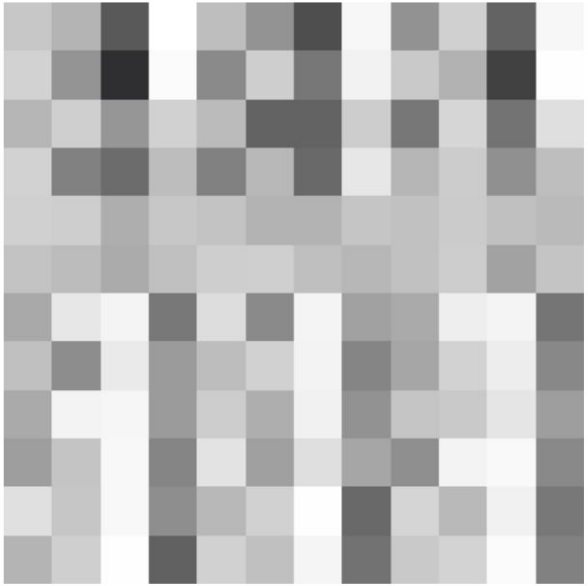
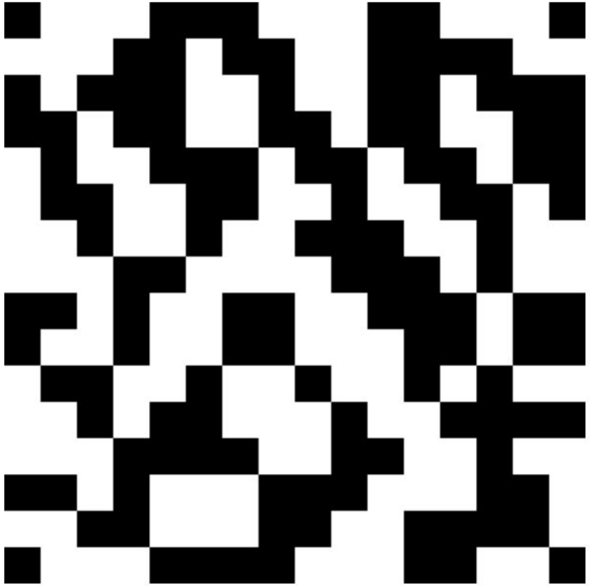
Issues with PHFs - NeuralHash

- Developed by Apple
- Standard Neural Network
  - Fully differentiable
- Trivial Collisions


`59a34eabe31910abfb06f308`
Collision Generated by
https://github.com/anishathalye/neural-hash-collider

Alternative PHFs

Alternative PHFs - PhotoDNA & PDQ

Input Image	PhotoDNA Digest (Microsoft)	PDQ Digest (Facebook)
		
	144 Bytes `0x04045e0005...`	256 Bits `0x1501505454...`

Alternative PHFs - PhotoDNA & PDQ

Input Image	PhotoDNA Digest (Microsoft)	PDQ Digest (Facebook)
	04045e0005...	1501505454...
	04045c0005...	1501505054...

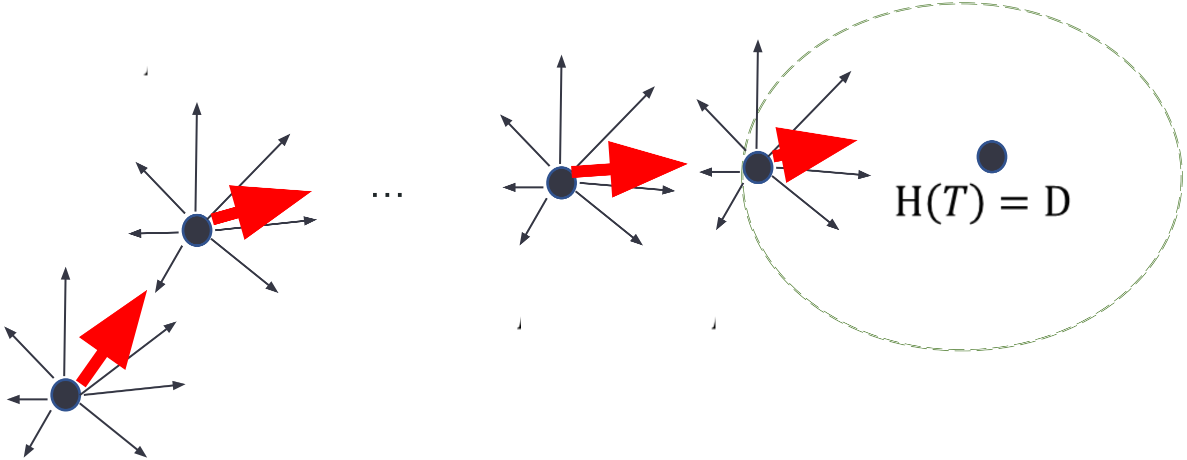
Attacking PhotoDNA & PDQ

Targeted Second-Preimage Attack

...

Images

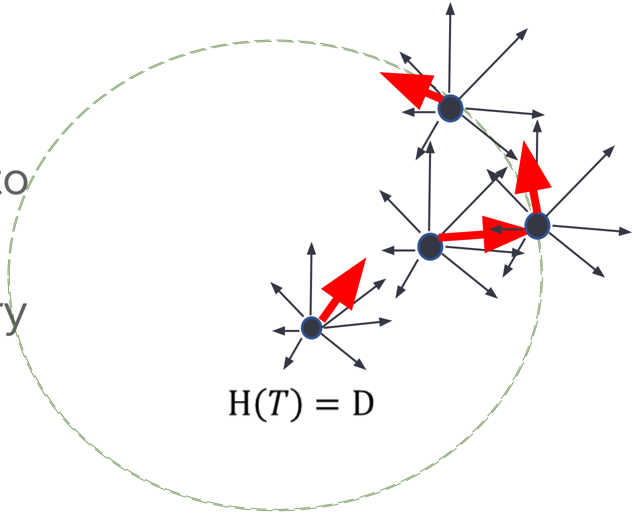
Hash Space



Detection Avoidance Attack

- Semantically equivalent Images which hash above threshold
  - Baseline Experiments
  - FP-rates
- Based on HopSkipJump Attack

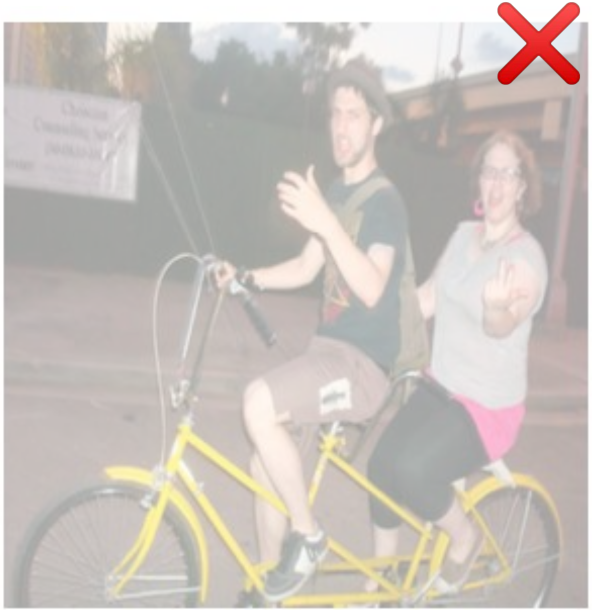
  Jianbo Chen et. al (2020)
- Generate random perturbations at boundary to compute gradient
- Move along gradient to find decision boundary
- Take a step towards target and repeat



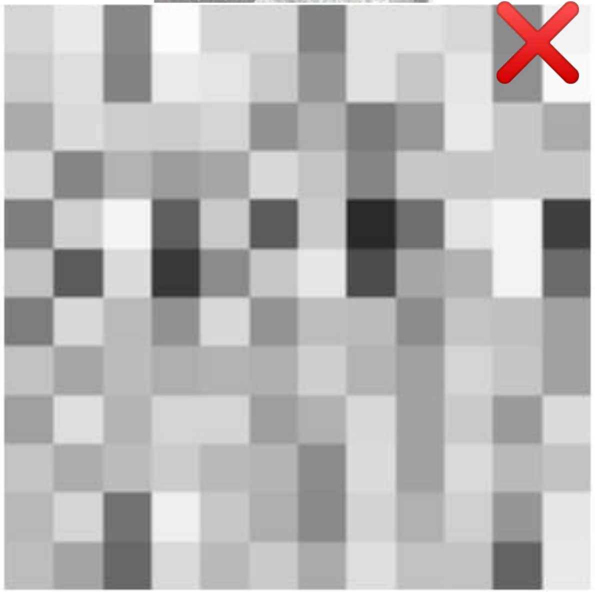
Results

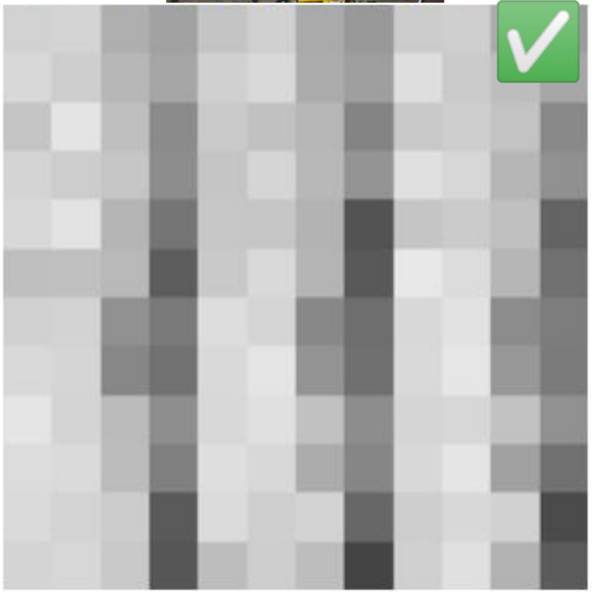
Targeted Second-Preimage Attack

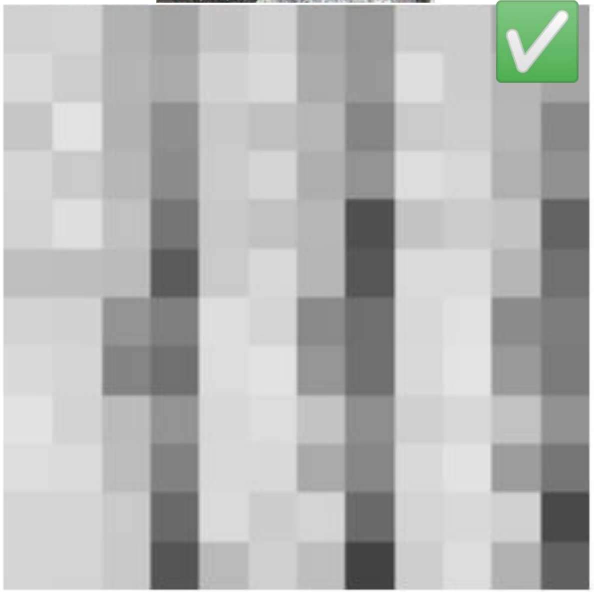
Targeted Second-Preimage Attack



Tar  ond-Preima

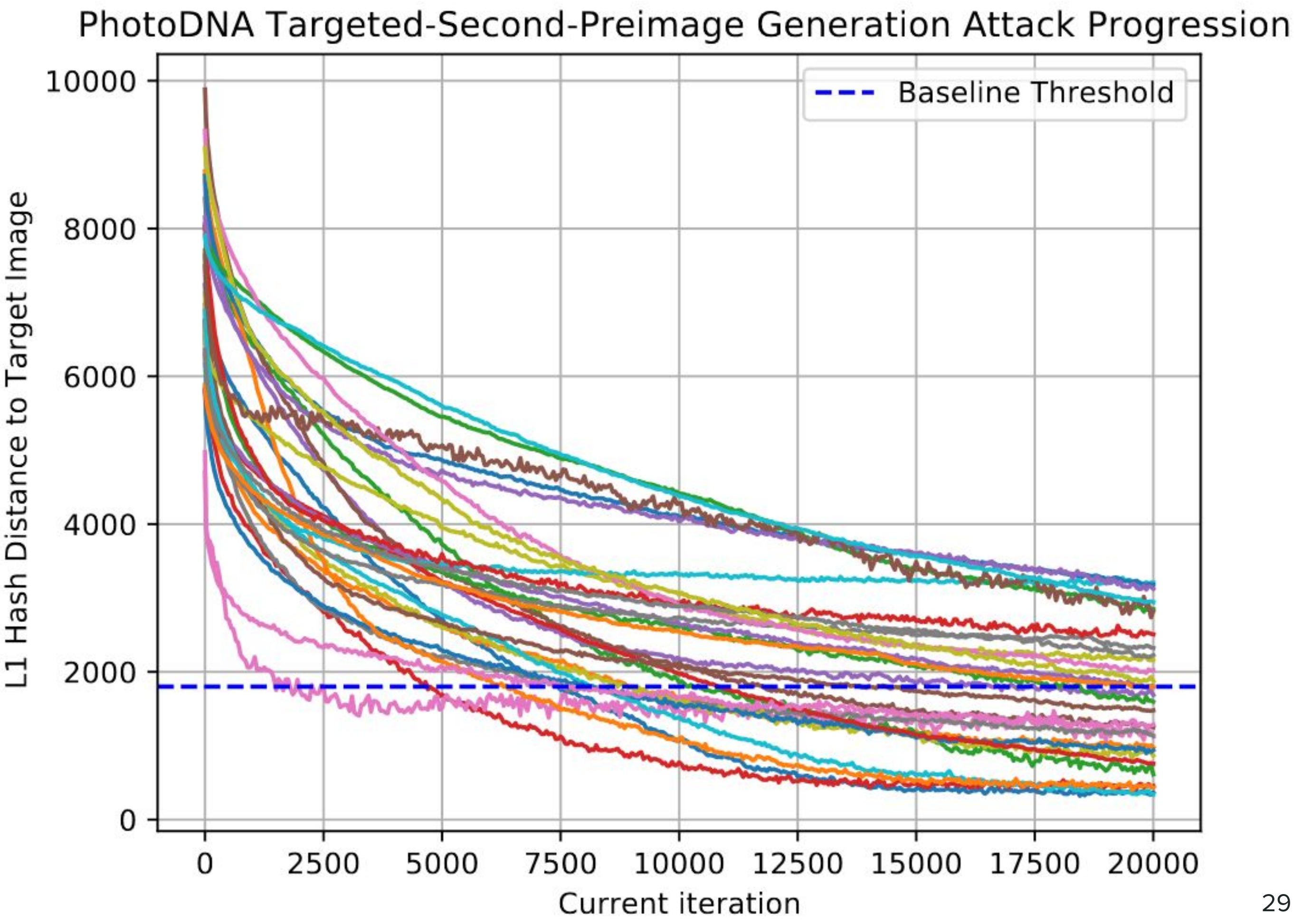






PhotoDNA

- ImageNet Pairs
- 17/30 Reached Baseline

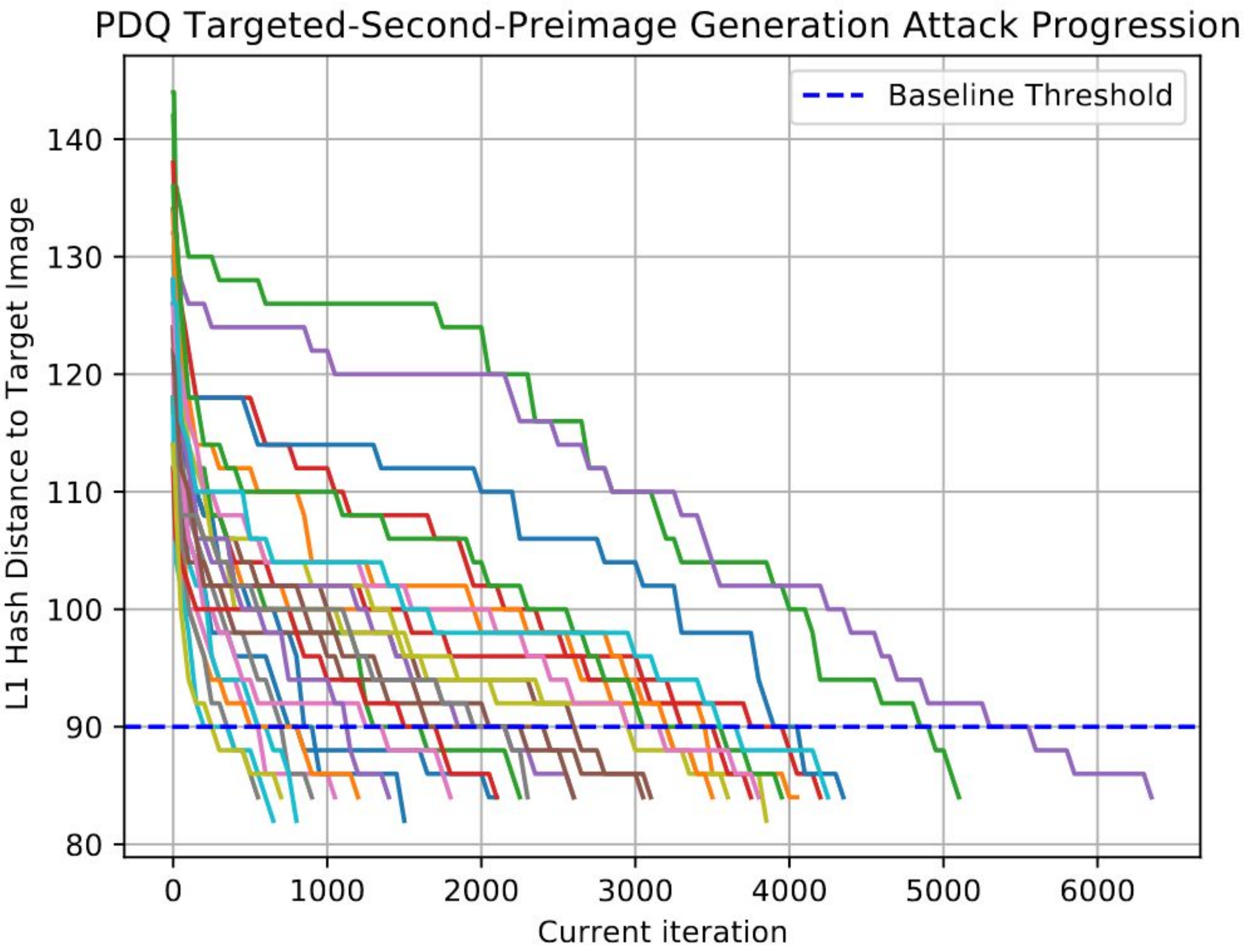


Targeted Second-Preimage Attack (PhotoDNA)

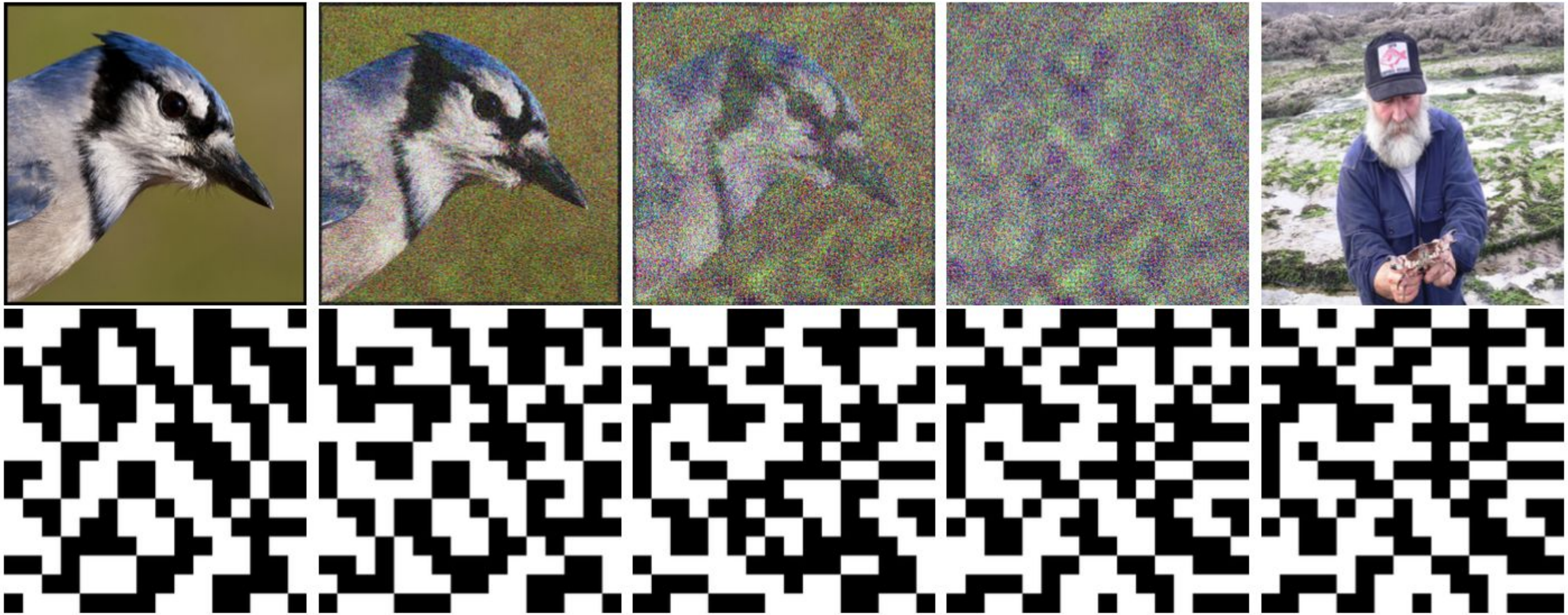
(a) Start: 6162 (b) Step 8200: 1797 (c) Step 12000: 963 (d) Step 20000: 342 (e) Target Image

PDQ

- ImageNet Pairs
- All 30 Reached Baseline



Targeted Second-Preimage Attack (PDQ)



(a) Start: 120 (b) Step 300: 88 (c) Step 800: 38 (d) Step 1600: 0 (e) Target Image

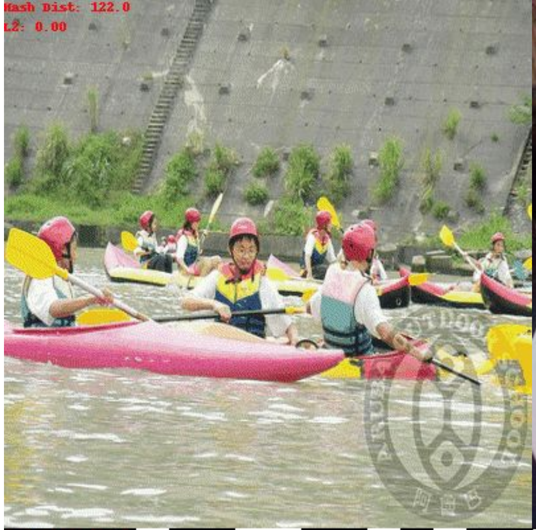
Source

Target

Source PhotoDNA Hash

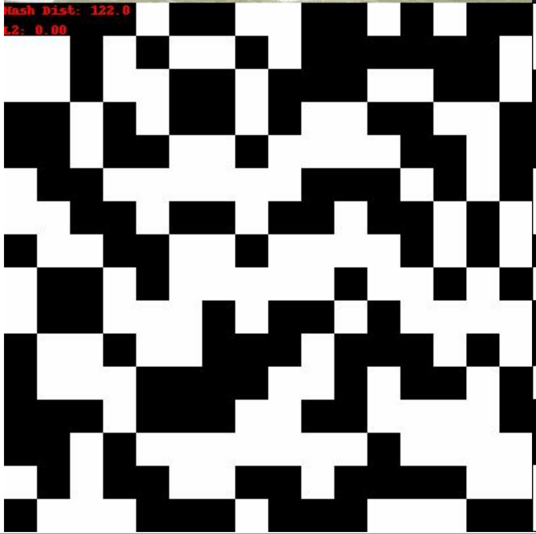
Target PhotoDNA Hash

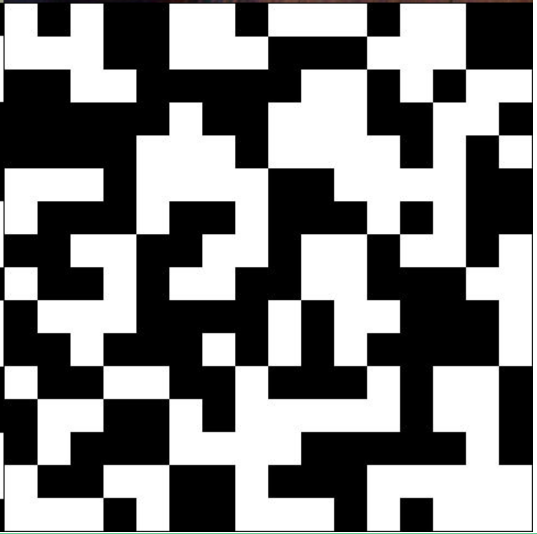
Source



Target

Source
PDQ
Hash





Target
PDQ
Hash

Detection Avoidance Attack (PhotoDNA)

(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

[Prokos et al. Squint hard enough: Evaluating perceptual hashing with machine learning (2021).]

Conclusion

- PHF susceptible to adversarial ML
- Still need content monitoring
- Breaks end-to-end encryption

ARTIFACT EVALUATED
usenix ASSOCIATION
FUNCTIONAL

ARTIFACT EVALUATED
usenix ASSOCIATION
REPRODUCED

https://www.perceptualhashing.lol/

Thank You! Questions?

Appendix

Threat Models

Targeted-Collision Surveillance Attacks

- Semantically non-equivalent match collision

1. Post innocuous images which hash to illicit images
    a. Nefarious service provider or insider threat can track deanonymized users
2. Introduce innocuous digest into E2EE-PHM database
    a. Send illicit image to NCMEC to add to database which matches to desired tracking image

Framing and Censorship

- Introduce innocuous hash to illicit database causing target user to be flagged
- Similarly introduce illicit image which hashes to censored image to database

Detection Avoidance

- Local DB checks
- Generate arbitrary images which evade detection
  - Disseminate throughout network

User Data Leakage

- Edge-hashing E2EE-PHM
- Preimage attribute recovery (classification)
- Preimage reconstruction (pix2pix)

Illicit-Content Data Leaks

- User gains access to DB
    - Detect attributes or reconstruct images

Background

What is a hash?

Term coined in the 1960s[1]

Properties of an effective hash[2]:

1. Distinct
2. Resilient
3. Deterministic
4. Efficient
5. Non-reversible

- Can't (and sometimes shouldn't) be all!

1. Hellerman, Herbert. 1967. *Digital computer system principles*. McGraw-Hill Companies.
2. Farid, Hany. "An Overview of Perceptual Hashing." *Journal of Online Trust and Safety* 1.1 (2021).

What about SHA?

Hash Function (checksums...)

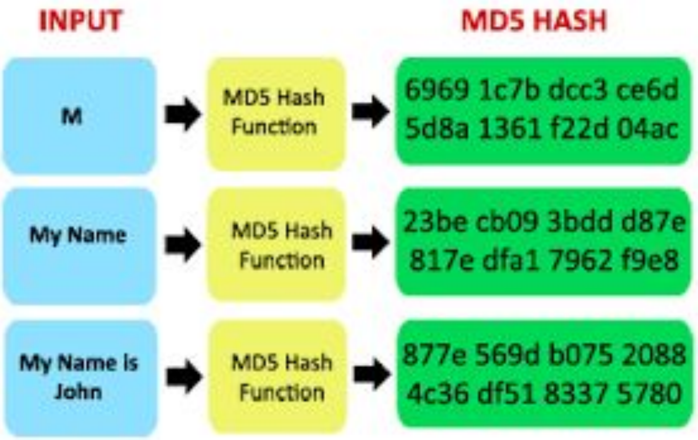
- Any function to map N-size to fixed size values
  - Error detection, lossy compression, etc

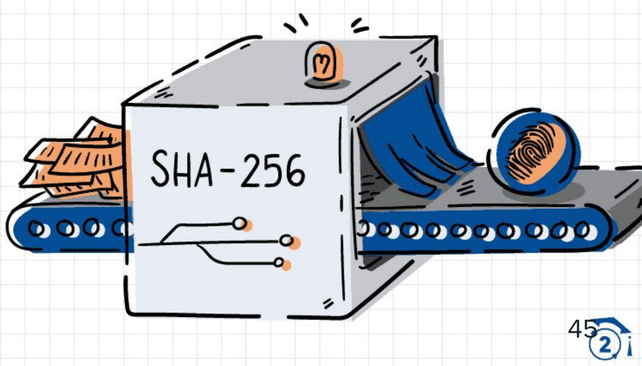
Cryptographic Hash Function (SHA...)

- One-way function which is **infeasible to invert**
  - Data authentication/integrity

Perceptual Hash Function

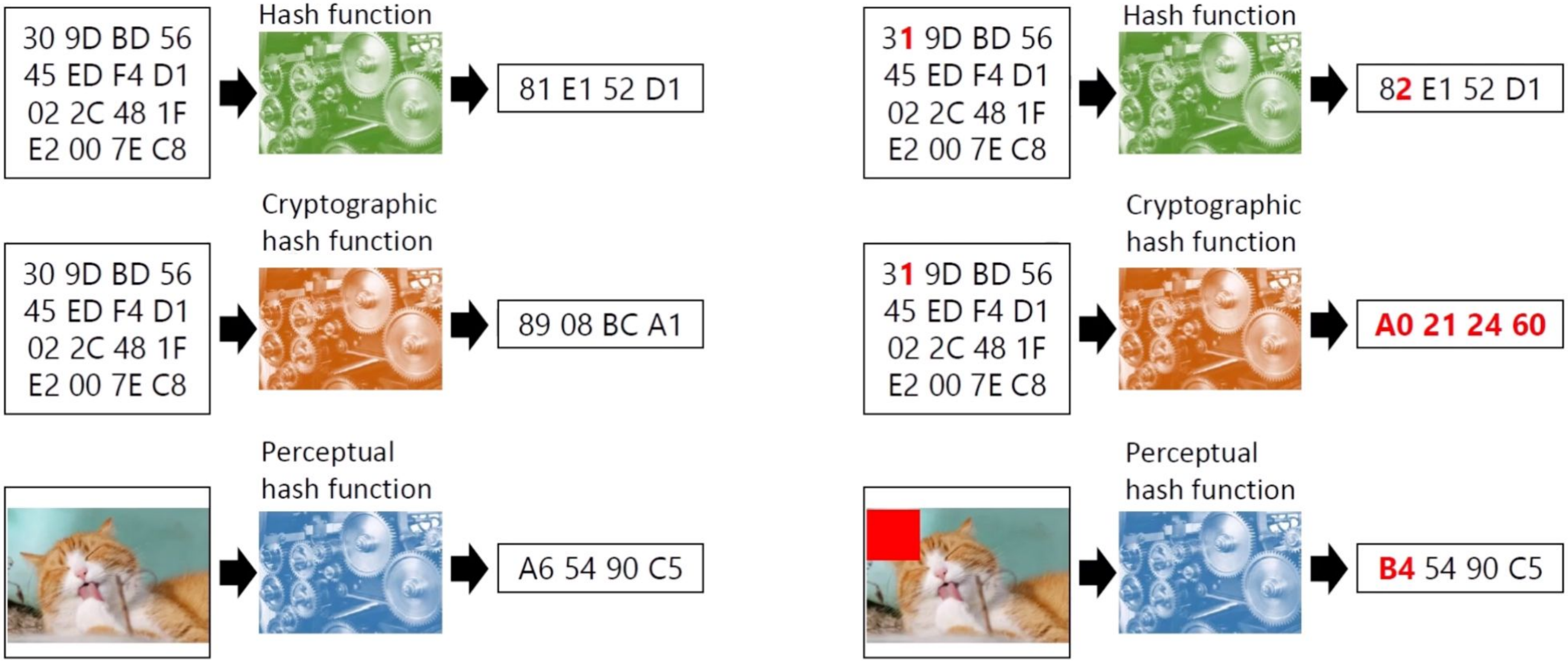
- Locality-sensitive
  - Image matching





1. Cryptographic Hash vs Perceptual Hash (2 of 2)

How changes to the input data affect the hash value



https://www.youtube.com/watch?v=lJ-QjDCaz-o

What is a Perceptual Hash Function

Hash Function

- Arbitrary input ➔ fixed-size values

(*Secure*) Cryptographic Hash Function

- One-way **non-invertible**; low-probability of collisions

Perceptual Hash Function

- Locality-sensitive; embeds multimedia semantics; fuzzy

Illicit Image Monitoring

- Prevent the spread of known illicit images
  - Impossible in fully end-to-end encrypted setting
- Safeguards without fear of corporate or government interference

What do we need?

- Feature-based privacy-preserving transforms
  - ~95% accuracy and a false pos on order of 1 in ten million†

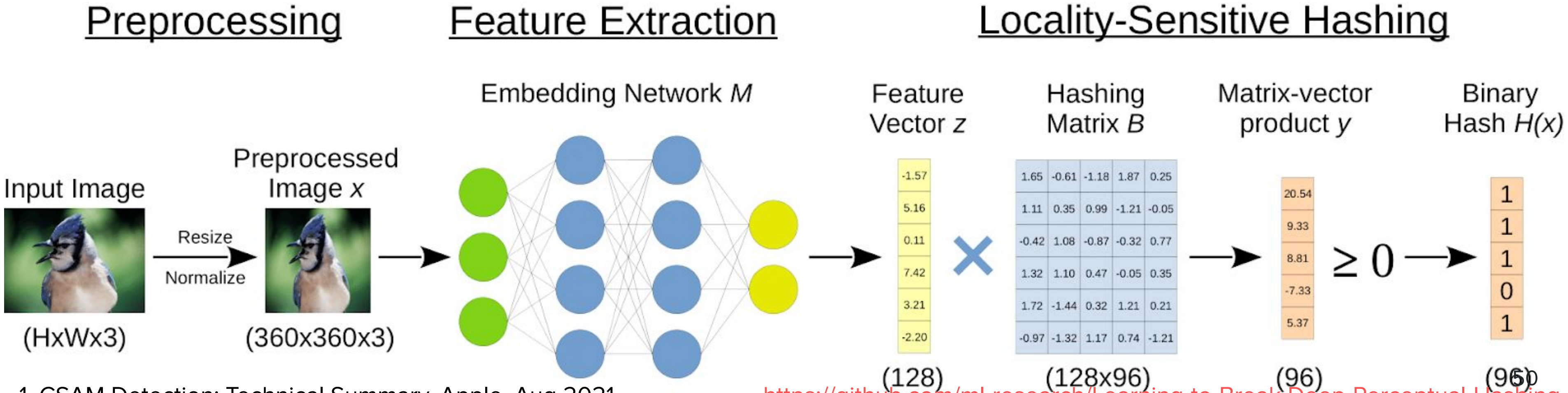
†Priyanka Singh and H. Farid. Robust homomorphic image hashing. In CVPR Workshops, 2019.

Existing Solutions

NeuralHash

- Developed by Apple
- Standard DNN
  - Fully Differentiable
- Trivial Collisions
  - Requires many assumptions within matching scheme[1]

59a34eabe31910abfb06f308
Collision Generated by https://github.com/anishathalye/neural-hash-collider

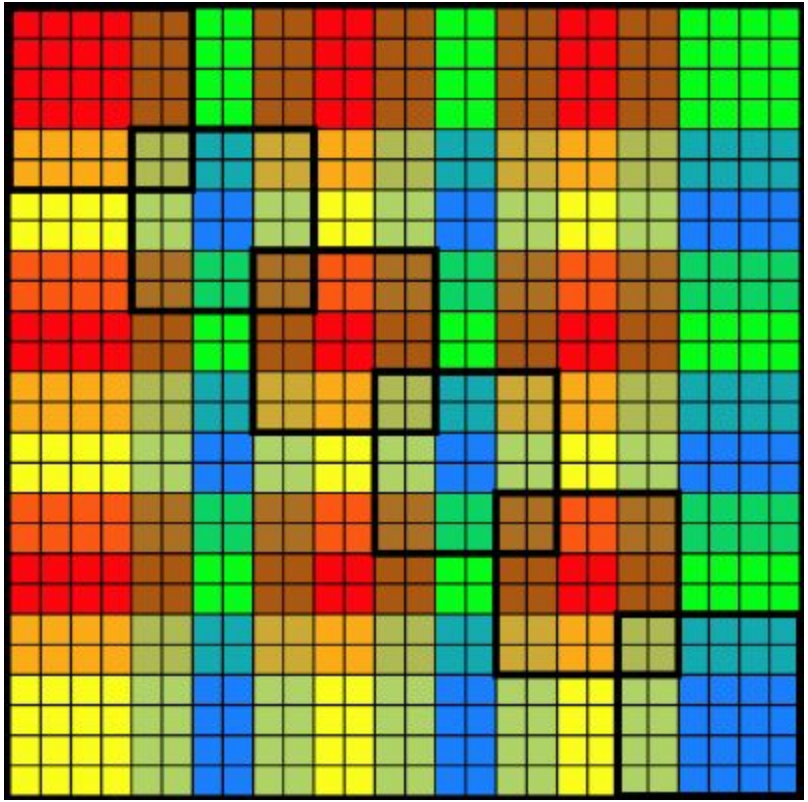


1. CSAM Detection: Technical Summary, Apple, Aug 2021

https://github.com/ml-research/Learning-to-Break-Deep-Perceptual-Hashing

PhotoDNA Construction

- Normalization
- Sobel Gradients
- Partitioning
- Concatenation
- L1-norm difference & MSE



Neal Krawetz. PhotoDNA and its limitations, 2021

PDQ Construction

- Two-pass Jarosz Filters
- L1 Norm of Quantized Gradients -> Rescale
- 2D Discrete Cosine Transform
    - Quantize around median

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[Facebook. ThreatExchange GitHub repository.](#)

Deployed Services

- YouTube Content ID (2007)
  - Copyright infringement
- PhotoDNA in Bing & SkyDrive (2009)
  - Followed by Twitter (2011) & Google (2016)
- PDQ & TMK+PDQF on Facebook (‘19)
- NeuralHash delayed due to security

Does it work?

- 1,348 ISIS videos matched from 229 known (‘18)

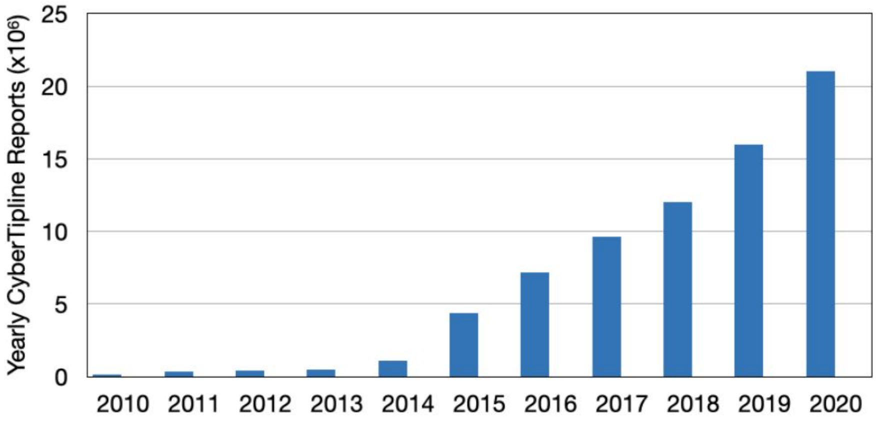


Figure 5: Yearly CSAM reports to NCMEC’s CyberTipline. From 2010 to 2020, the number of yearly reports jumped from slightly more than 100,000 to over 20,000,000.

Farid, Hany. "An Overview of Perceptual Hashing." *Journal of Online Trust and Safety* 1.1 (2021).

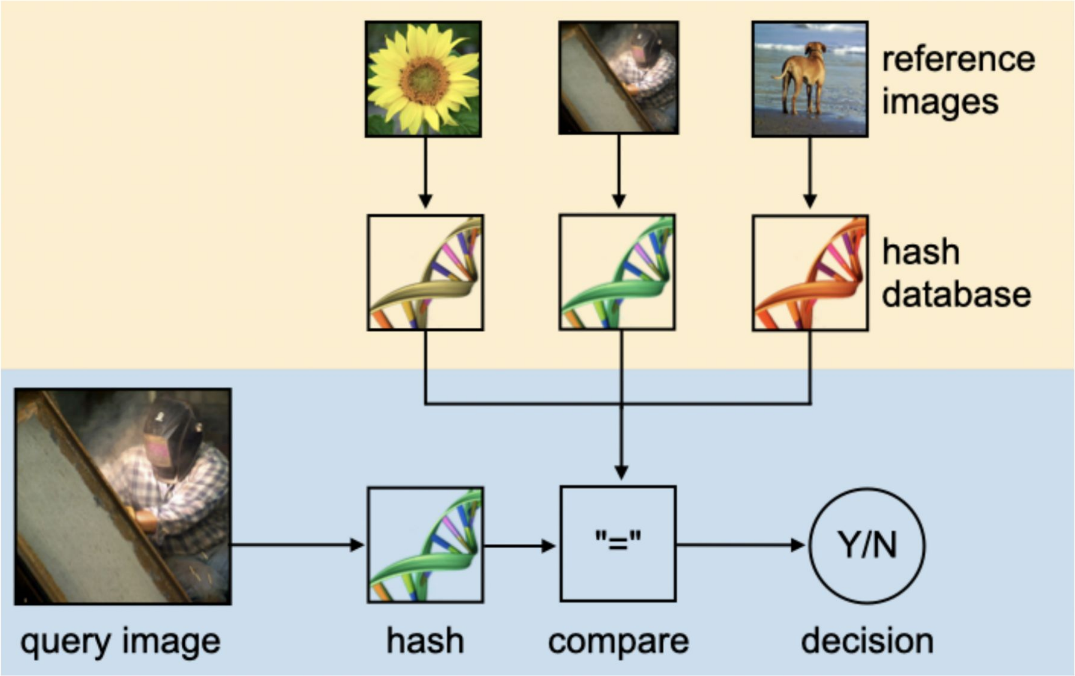
Perceptual Hash Matching (PHM) Scheme

Perceptual Hash Function produces digest

Computed digest compared against pre-computed illicit digest database

Several designs

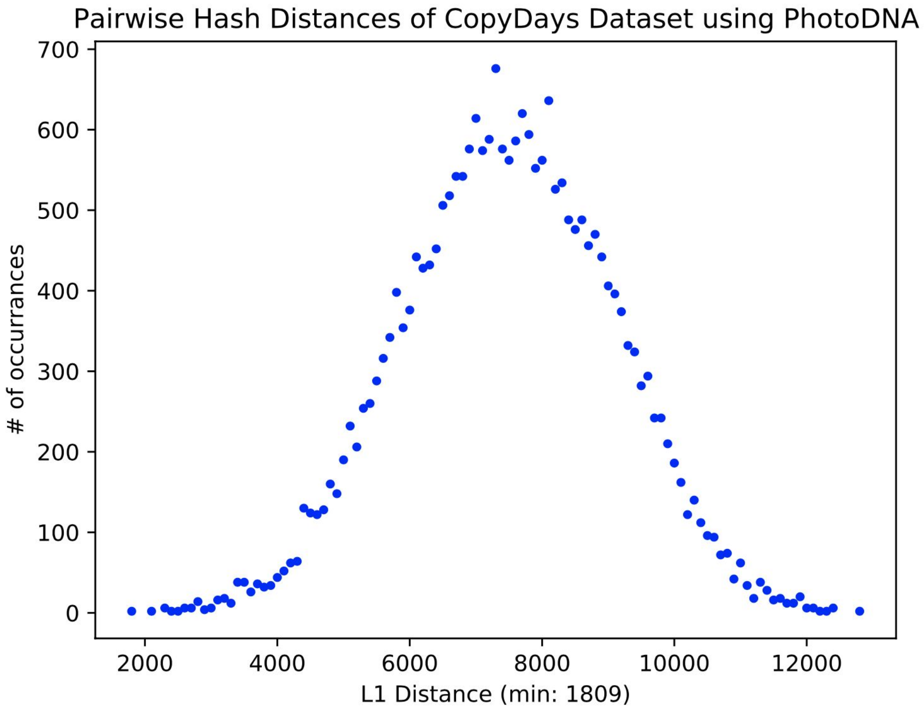
- Client-Side
- Private Set Intersection
- Edge Hashing (common)

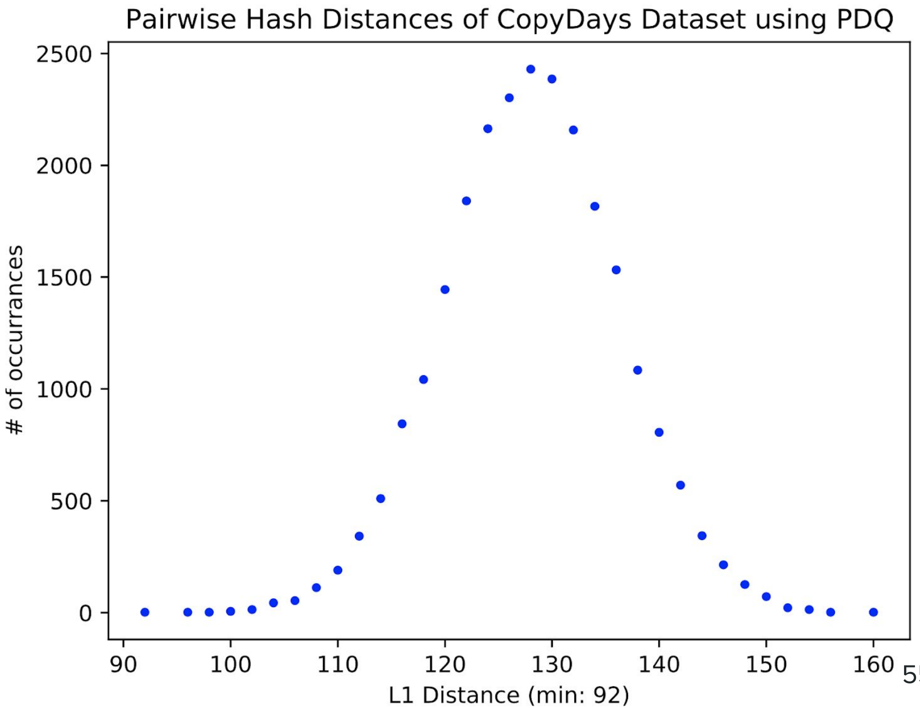


Farid, Hany. "An Overview of Perceptual Hashing." *Journal of Online Trust and Safety* 1.1 (2021).

Pairwise Hash Matching Distance Computation

- 157 Perceptually Distinct Images



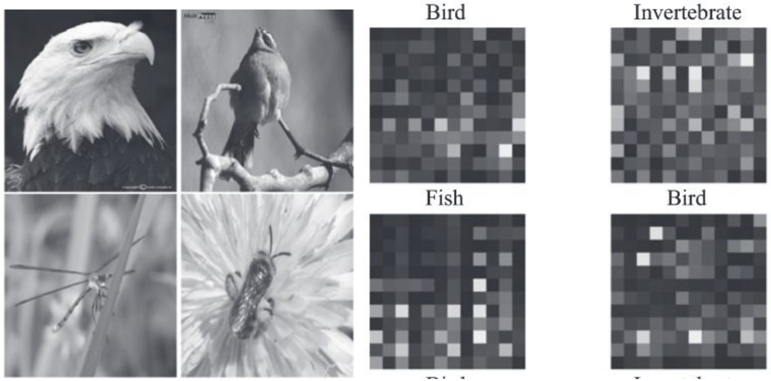


Prior Investigations

Privacy verification of PhotoDNA based on ML

Nadeem, Franqueira, Zhai (Aug 2019)

- Microsoft provided dataset (ImageNET)
- Trained for classification
  - Used CNN with 3 conv layers
- Claim to show resistance to machine-learning-based classification attacks

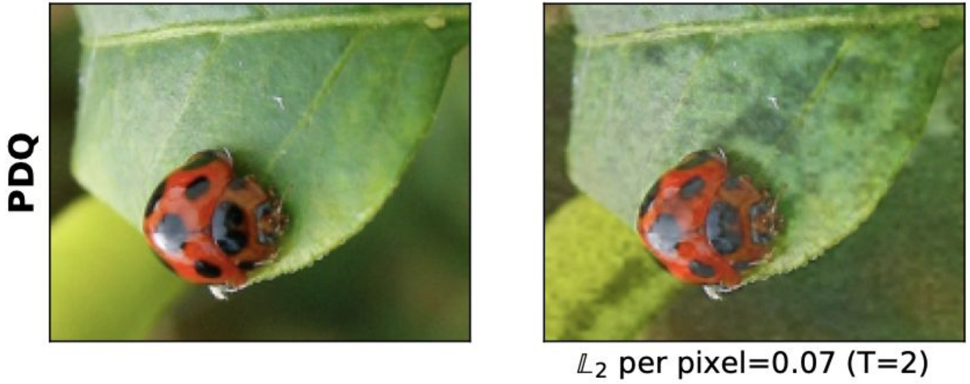


Classifier type	Classifier	Accuracy
Distance based	*K*NN	47.50
Tree based	Decision tree (DT)	42.32
	Random forest (RF)	**57.20**
Function based	SVM	34.23
	ANN	40.47
	CNN	**53.40**

Adversarial Detection Avoidance Attacks

Subham Jain et al. (2022)

- Evaluation of DCT based algorithms
- Able to generate images which avoid matching

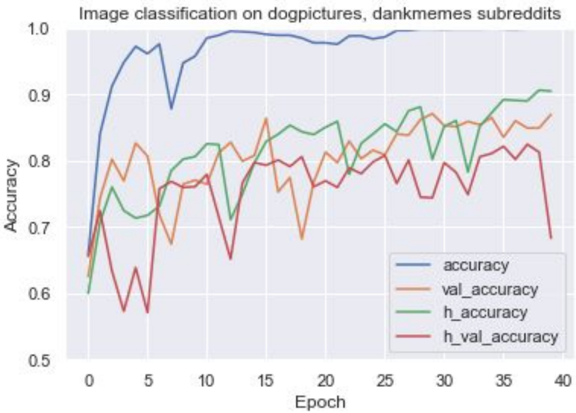


Initial Investigations

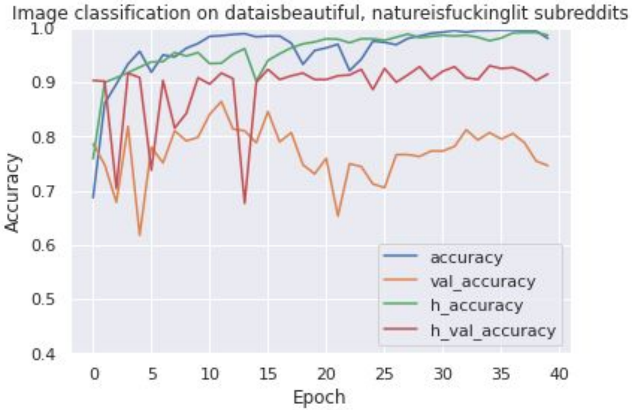
Not part of USENIX ‘23 submission

Binary Classification of Subreddits (Hash vs Orig)

- 300x300 resolution, 5x5 blocks, same CNN structure
- 3,385 images from DogPictures and DankMemes
- 1,971 images from DataIsBeautiful and NatureIsF******Lit



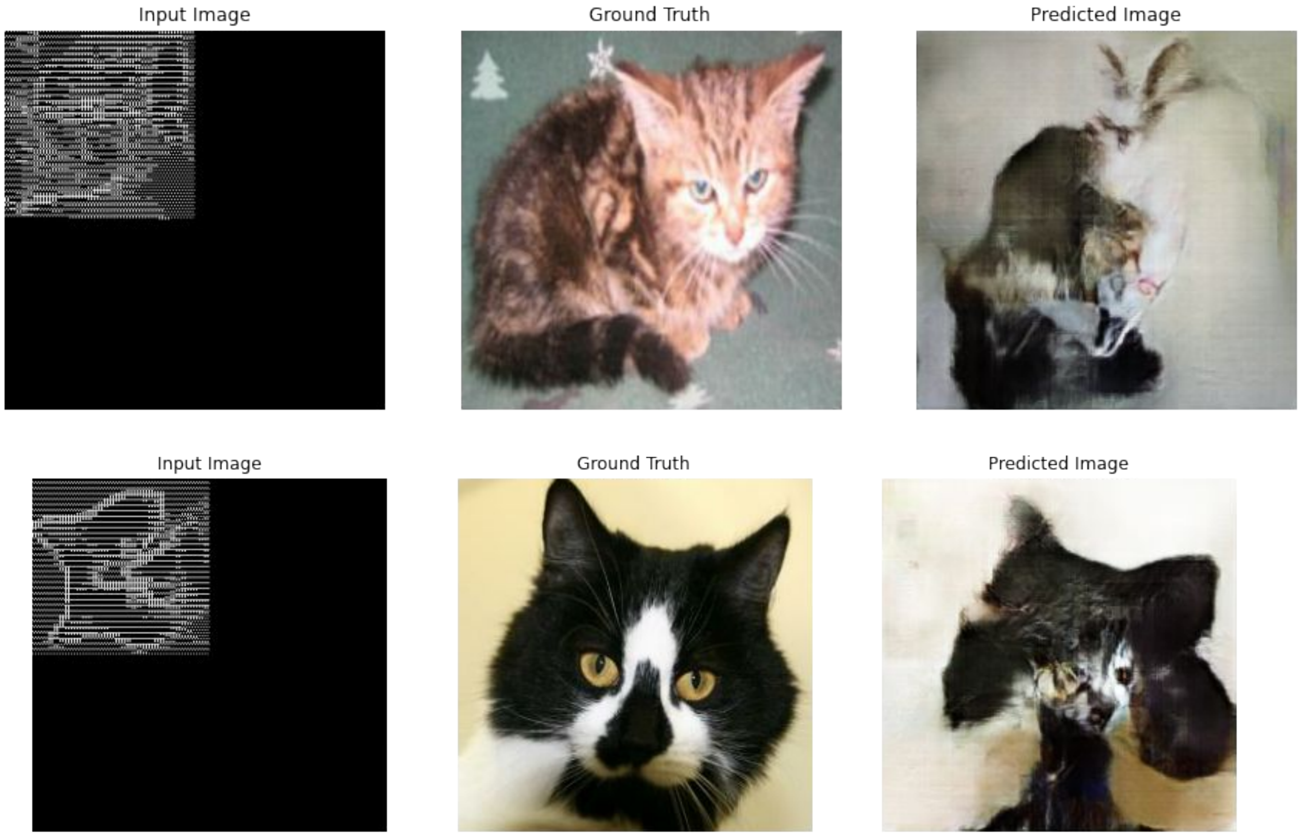
Dogs vs Memes
No Hashing: loss: 2.8479 accuracy: 0.8701
Hashing: loss: 0.7700 accuracy: 0.6831

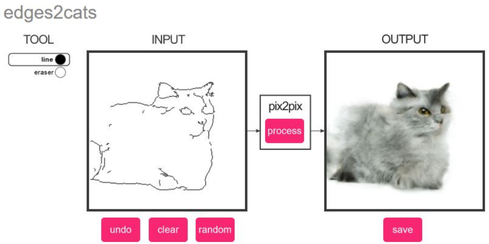


Data vs Nature
No Hashing: loss: 4.2674 accuracy: 0.7462
Hashing: loss: 0.6451 accuracy: 0.9154

pix2pix

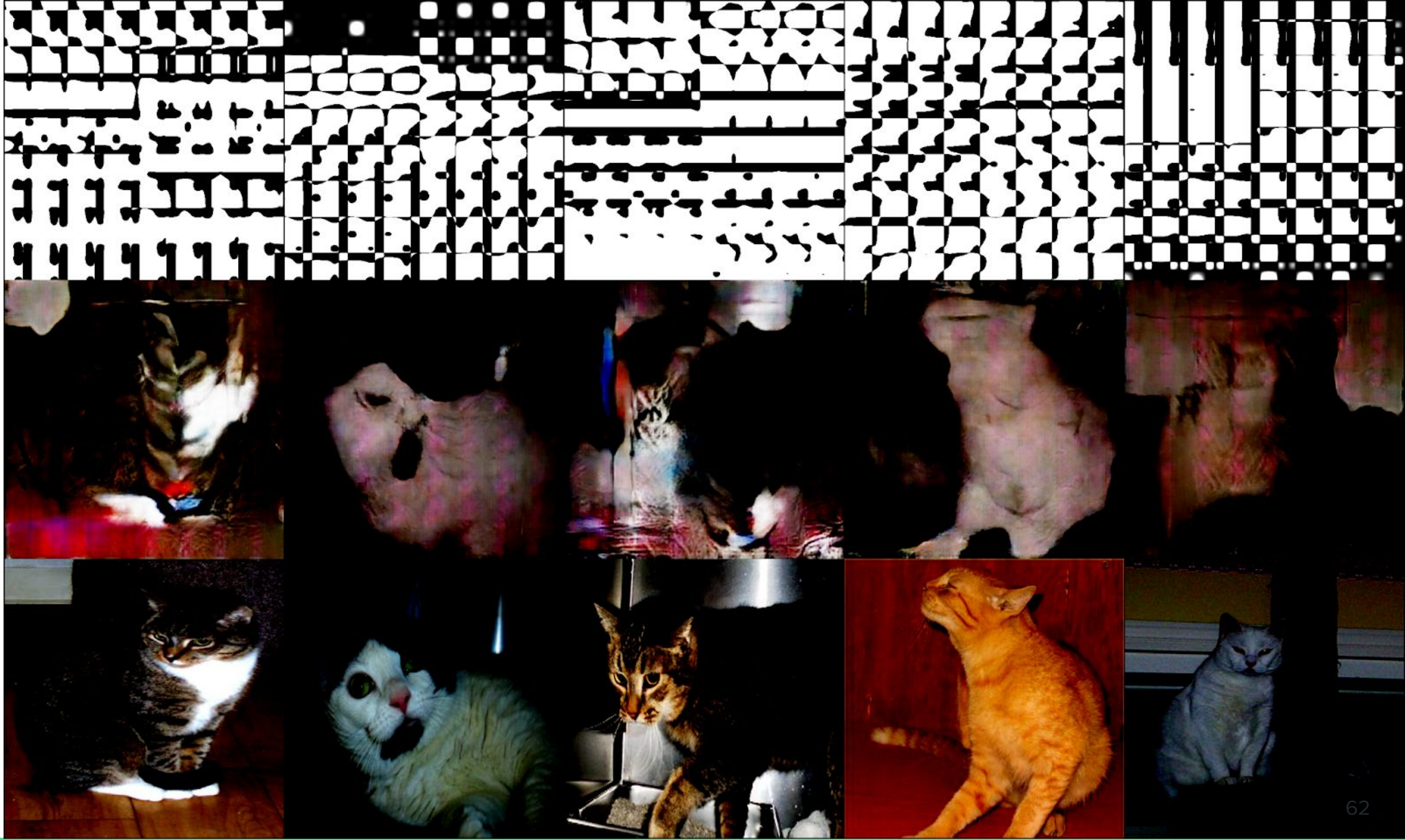
- Conditional GANs
- Default L1 Loss
- 32x32 Blocks



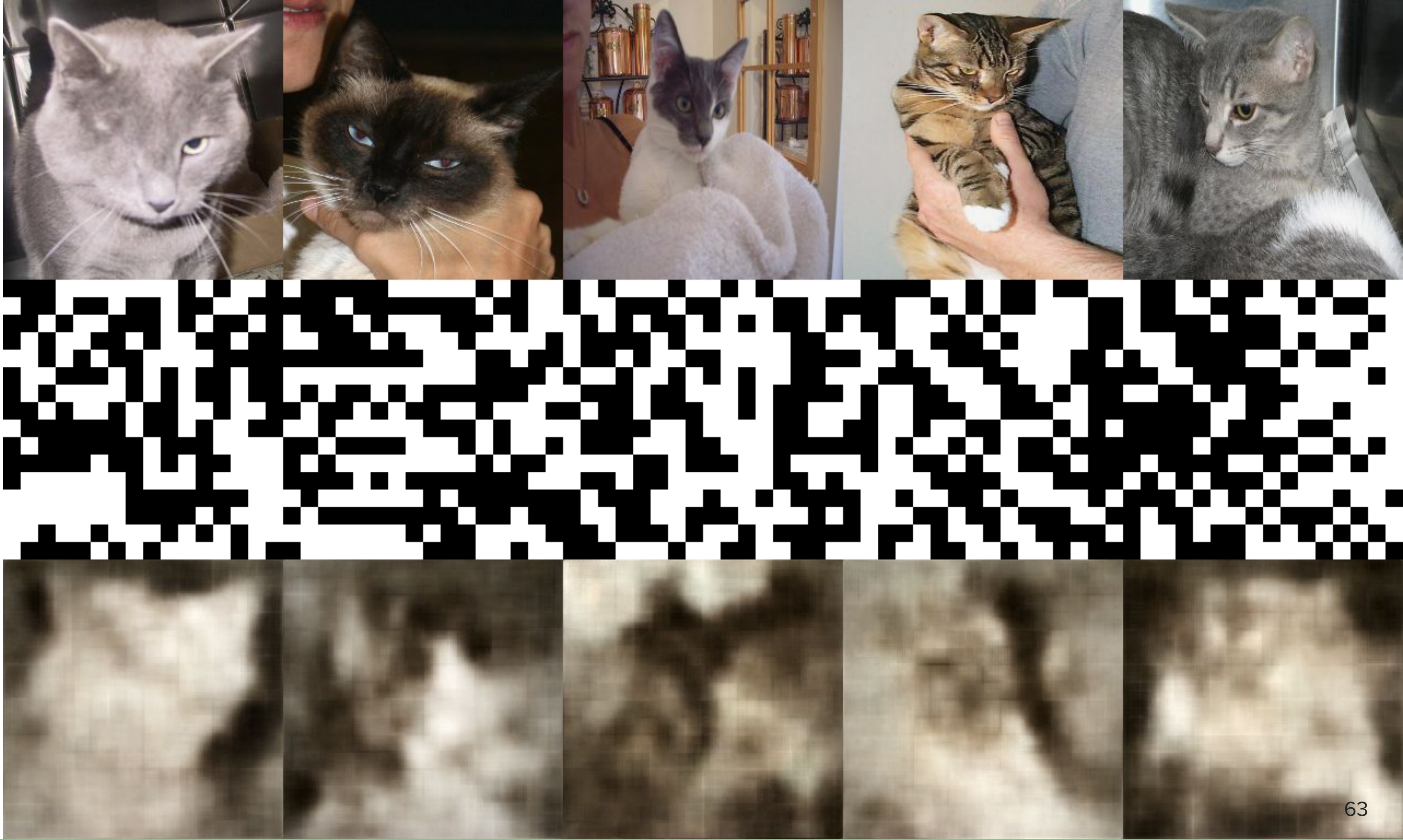


https://affinelayer.com/pixsrv/

PDNA (MOD)



PDQ



References

Farid, Hany. "An Overview of Perceptual Hashing." *Journal of Online Trust and Safety* 1.1 (2021).

Hellerman, Herbert. 1967. *Digital computer system principles*. McGraw-Hill Companies.

Perceptual Hashing To Compare Images Explained, https://youtu.be/lJ-QjDCaz-o

CSAM Detection: Technical Summary, Apple, Aug 2021. https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf

Neal Krawetz. PhotoDNA and its limitations, 2021

Facebook. ThreatExchange GitHub repository.

Learning to Break Deep Perceptual Hashing. ML-Research GitHub repository.

Nadeem, Franqueira, Zhai (Aug 2019)

Subham Jain et al. (2022)

Image-to-Image Demo, pix2pix, https://affinelayer.com/pixsrv/.

Prokos et al. Squint hard enough: Evaluating perceptual hashing with machine learning (2021).