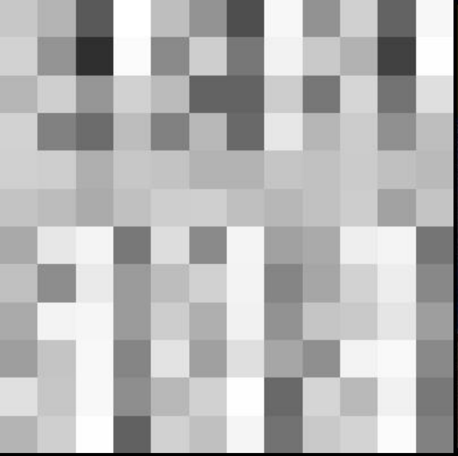
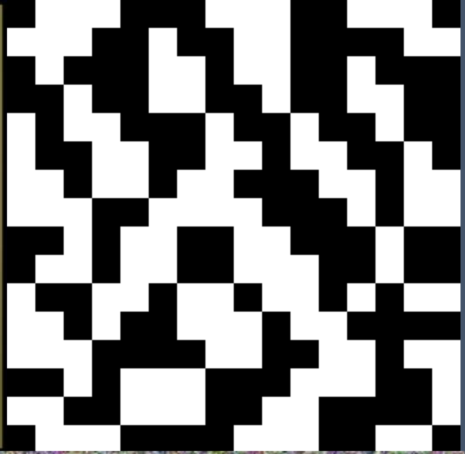





White-Box Attacks on PhotoDNA Perceptual Hash Function

Diane Leblanc-Albarel

diane.leblanc-albarel@esat.kuleuven.be

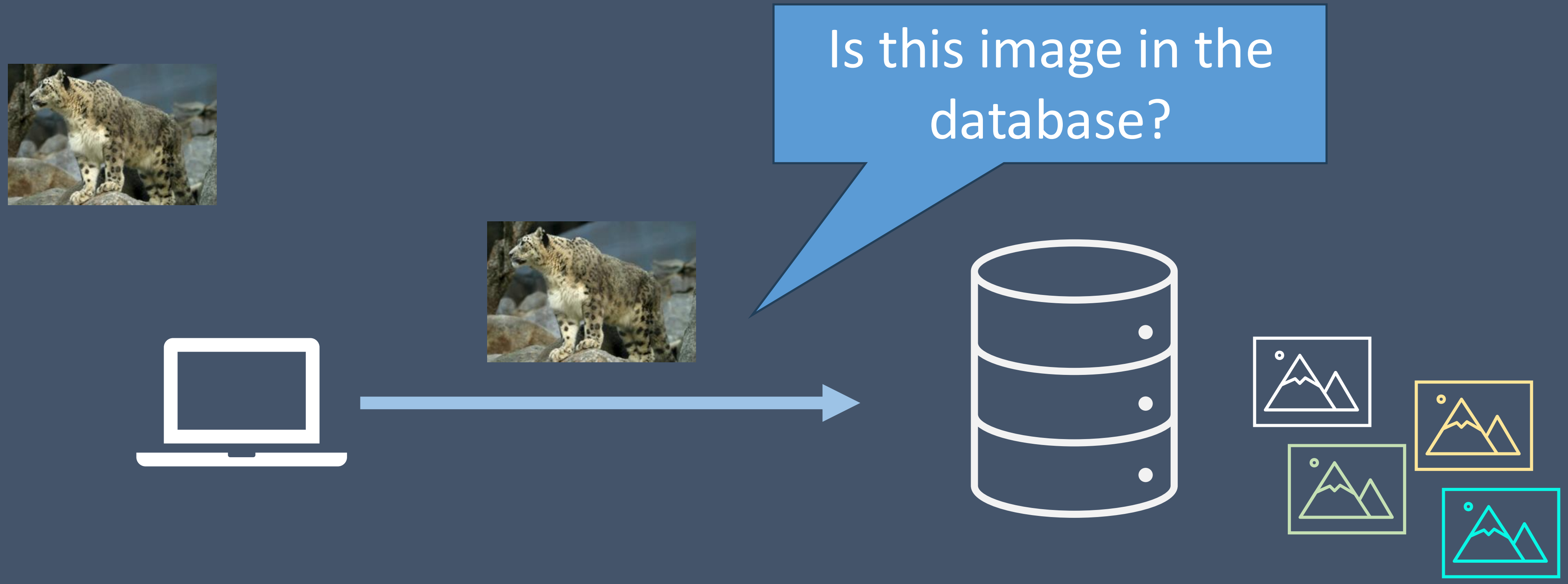
2nd June 2026

Joint work with
Bart Preneel
Maxime Deryck

KU LEUVEN

COSIC

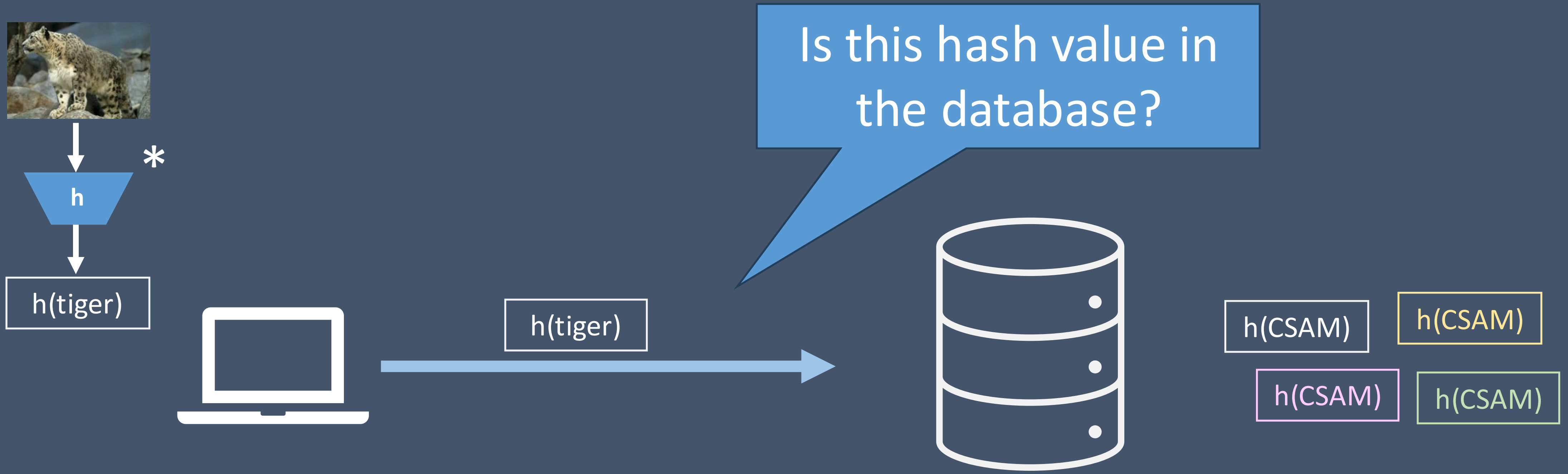
Detection of known illegal images



Database with illegal images

- Huge database
- Illegal/private images in clear

Detection of known illegal images



- Smaller database
- Privacy preserving

Database with hash values of illegal images

* h is a cryptographic hash function (create an unique digital fingerprint of an image)

Problem for images

Small change to the image → completely different hash value

h

h

1, 1, 0 , 1 , 1 , 0 , 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1

1, 0, 1 , 0 , 0 , 0 , 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0

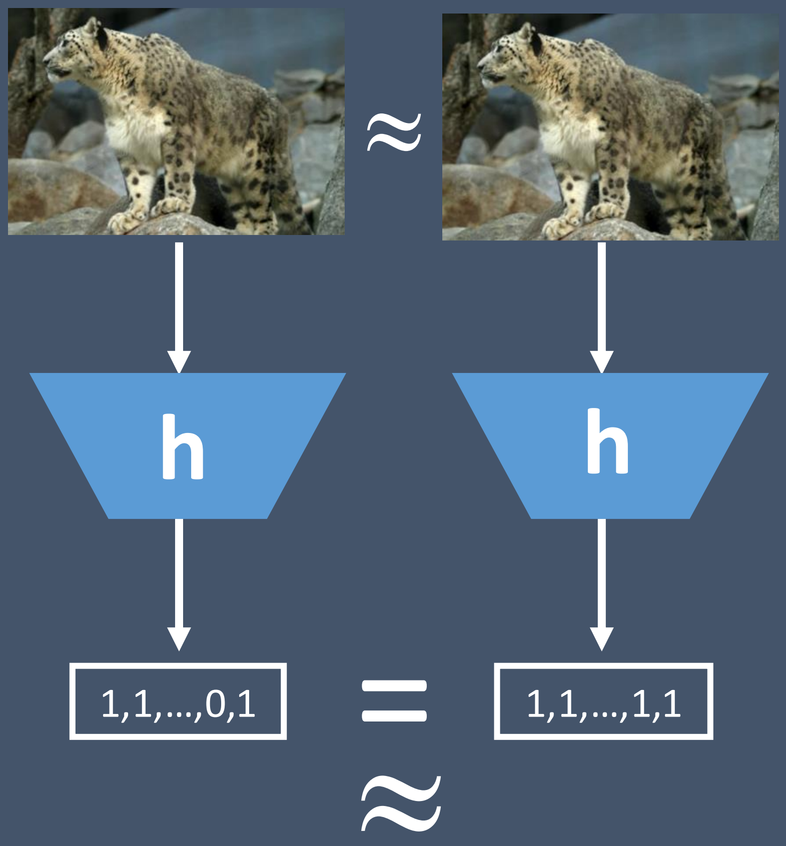
- No longer detected by cryptographic hash functions
- Not visible to the human observer

Solution: perceptual hash function

Hash function for which perceptually similar content gives the same result (or a similar result)

- Color consistency (e.g., brightness)
- Structural similarity (e.g., rotation)
- Content preservation (e.g., compression)
- Noise robustness

When the distance between the hash values of two images is lower than a chosen threshold τ, the images are considered similar



Copyright violation

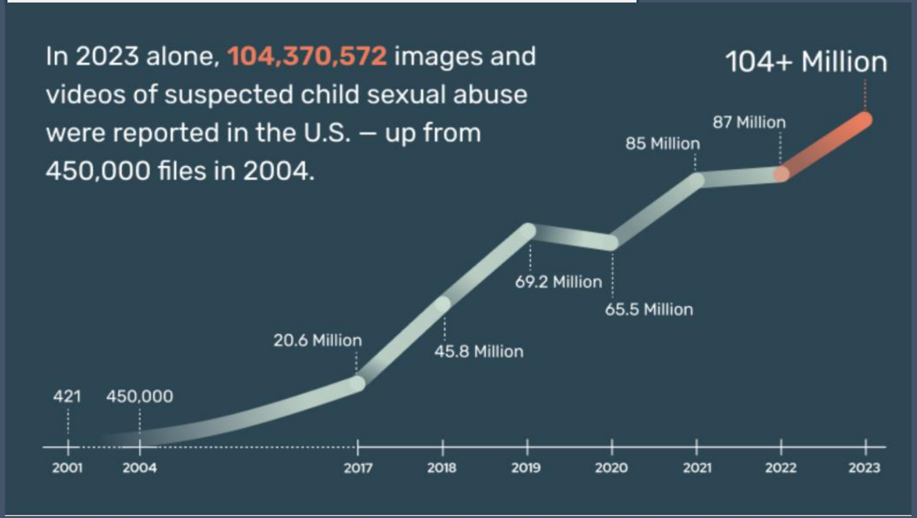
- Meta removes millions of copyright-violating items annually
- Image copyright infringement is massive and structural
- Most infringements are non-malicious and unintentional
- Enforcement data runs into billions of incidents per year
- Automated detection dominates enforcement

CSAM detection

(Child Sexual Abuse Material)

NATIONAL CENTER FOR MISSING & EXPLOITED CHILDREN

THORN



Slide credit: Bart Preneel

PhotoDNA Deployment

- 2009: PhotoDNA is a proprietary function designed by Prof. Farid and Microsoft
- 2014: Wildely deployed by NCMEC and law enforcement for CSAM detection
  - Currently used by more than 1900 service providers
  - Billions of images scanned per day (Instagram, Google, Meta, Microsoft, TikTok etc.)
- NCMEC: PhotoDNA has a false positive rate of less than 1 in 1 trillion (or 10/50* billion?) and a false negative rate of less than 2%

*European Commission, Report from the Commission to the European Parliament and the Council on the implementation of Regulation (EU) 2021/1232 (COM(2023) 797 final, 19 December 2023).

Meta Ireland 2025 transparency report to EU

https://transparency.meta.com/sr/eu-csam-derogation-report-2025/

(4) The number of cases of online child sexual abuse identified, differentiating between online child sexual abuse material and solicitation of children:

Between January 1, 2024 and December 31, 2024, we actioned around 1.5 million pieces of media constituting CSAM that were detected using our media matching technology, in private message threads on Messenger and Instagram Direct which included an EU user.

(5) The number of cases in which a user has lodged a complaint with the internal redress mechanism or with a judicial authority and the outcome of such complaints:

During the same period, users appealed the actions of around 76.9 thousand pieces of their shared media content. Following the appeals process, around 1.8 thousand pieces of content were restored and account actions reversed.

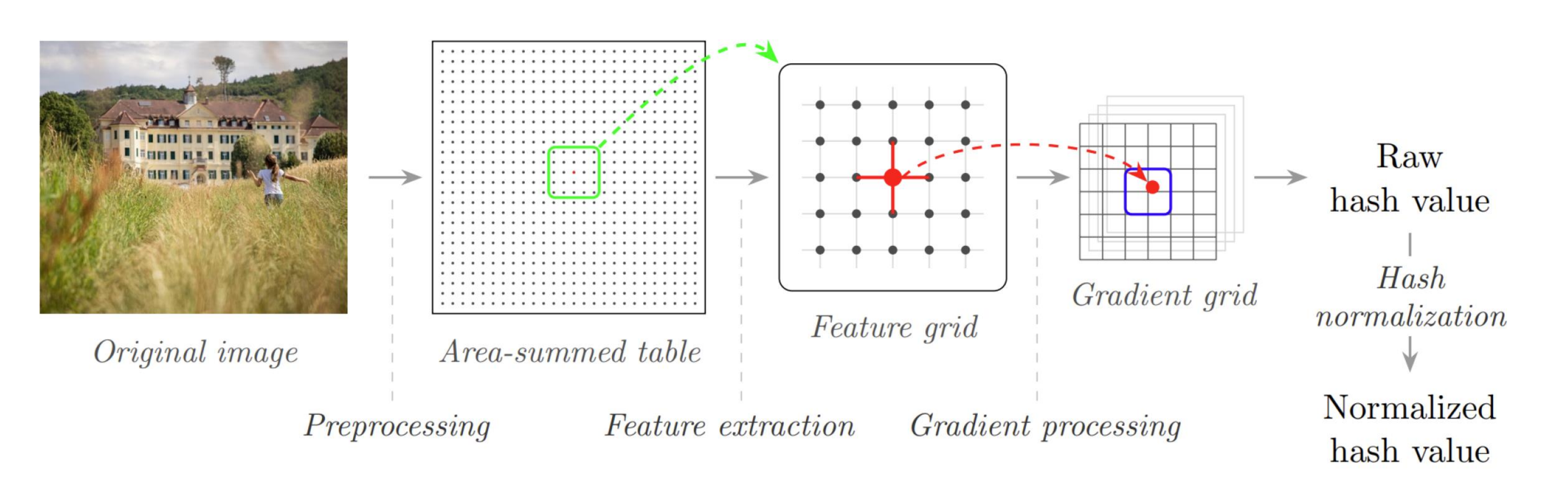
(6) The numbers and ratios of errors of the different technologies used:

Of the around 1.5 million actioned pieces of media, around 1.8 thousand pieces of content were restored and account actions reversed since we identified an error in our detection. The overturn rate is therefore 0.12%.

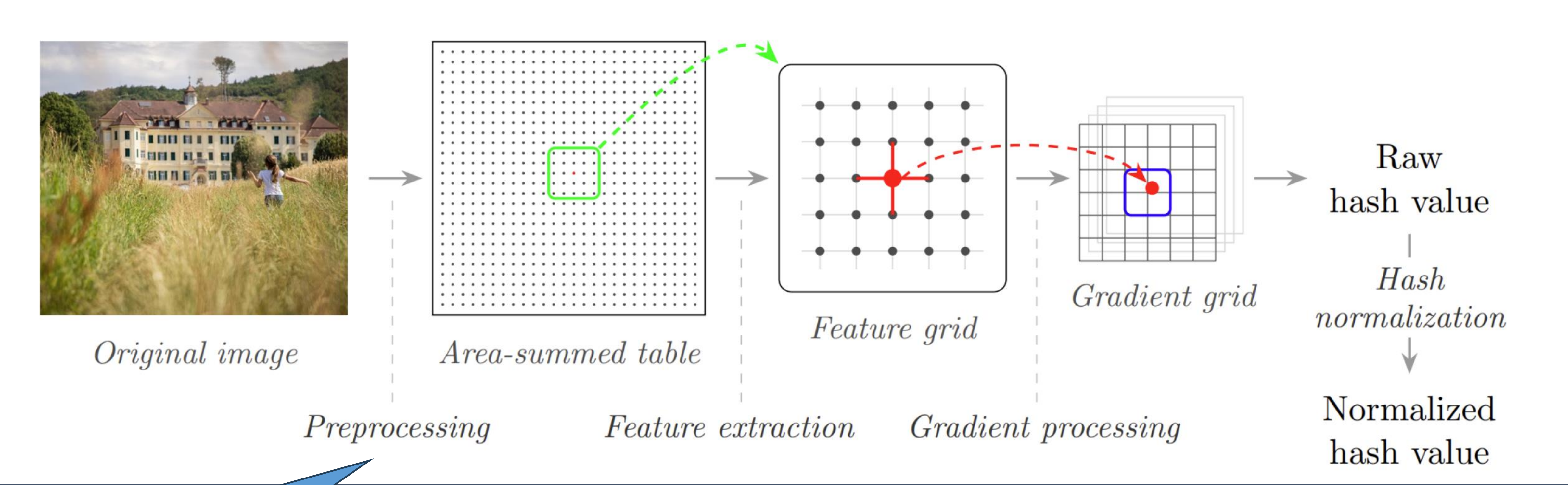
Attacks on PhotoDNA

- 2021: Executable leaked on GitHub
- 2022-2025: Several black-box attacks (using ML)
- 2025 - our results:
  - Identified the complete description “Alleged PhotoDNA”
(verified on 50K examples)
  - Confirmed by Microsoft
  - New attacks and substantial improvement of previous attacks: faster, closer distance, less noisy

PhotoDNA: description



Preprocessing



Dimension check (50x50)
Sum color values
Summation per square

Area-summed table

Original pixel values ($P(x,y)$)

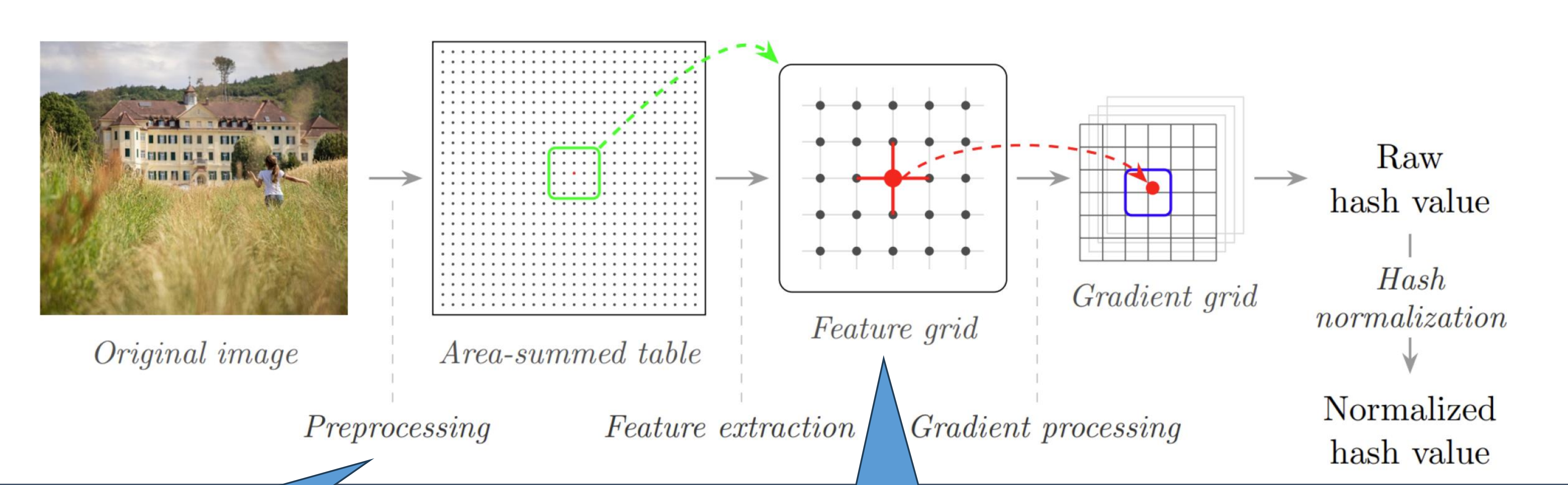
8	3	5	2
1	7	4	6
9	2	3	8
4	6	1	5

Sum: $8+3+1+7=19$

Summed-area table ($I(x,y)$)

8	11	16	18
9	19	28	36
18	30	42	58
22	40	53	74

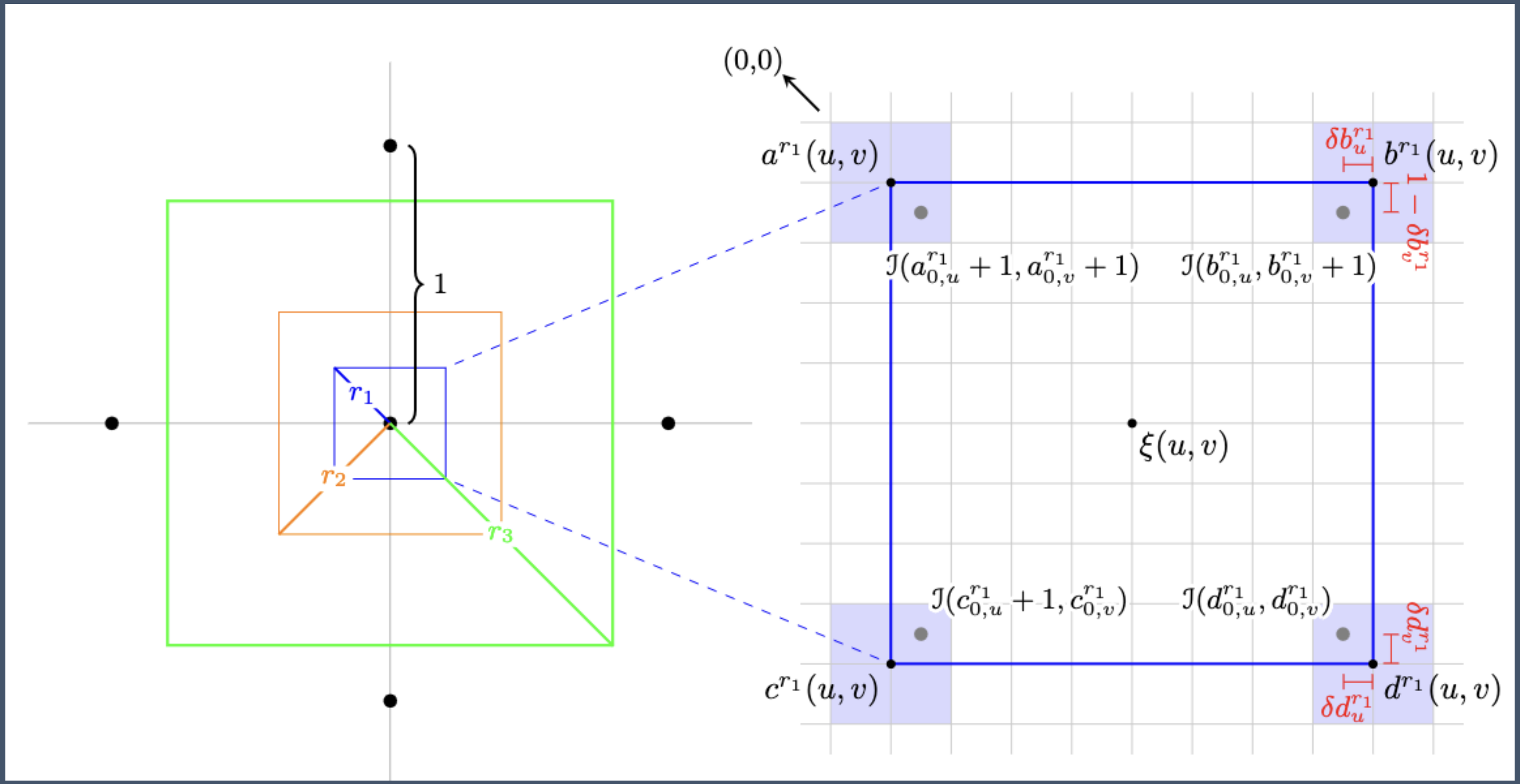
PhotoDNA: description



Dimension check (50x50)
Sum color values
Summation per square

26x26 feature grid capturing brightness and texture weighting of 3 rectangles

Feature Grid Computation

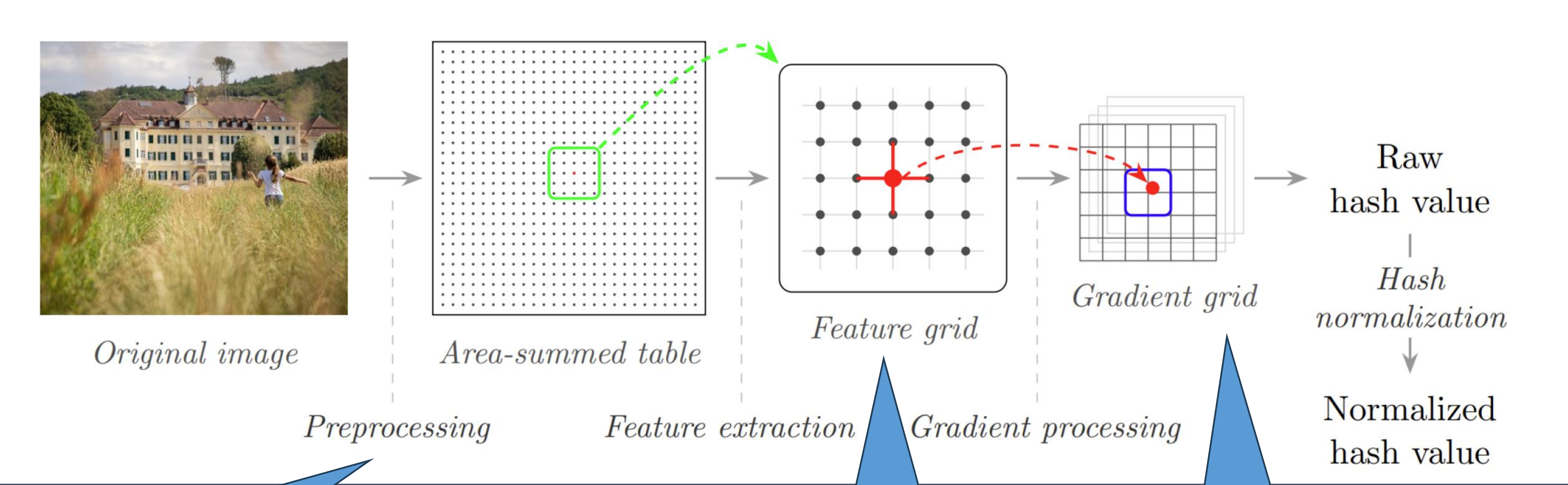


R_1 Fine local details
R_2 Medium-scale structures
R_3 Broder visual context
w_i approx. Gaussian bloor (constant)

$$\begin{aligned} Z^{r_i}(u,v) = &\ \mathcal{I}(z_{c,u}^{r_i}, z_{c,v}^{r_i}) \cdot (1-\delta z_u^{r_i}) \cdot (1-\delta z_v^{r_i}) \\ &+ \mathcal{I}(z_{c,u}^{r_i}+1, z_{c,v}^{r_i}) \cdot \delta z_u^{r_i} \cdot (1-\delta z_v^{r_i}) \\ &+ \mathcal{I}(z_{c,u}^{r_i}, z_{c,v}^{r_i}+1) \cdot (1-\delta z_u^{r_i}) \cdot \delta z_v^{r_i} \\ &+ \mathcal{I}(z_{c,u}^{r_i}+1, z_{c,v}^{r_i}+1) \cdot \delta z_u^{r_i} \cdot \delta z_v^{r_i}\,. \end{aligned}$$

$$F(u,v) = w_1 R_1 + w_2 R_2 + w_3 R_3$$

PhotoDNA: description

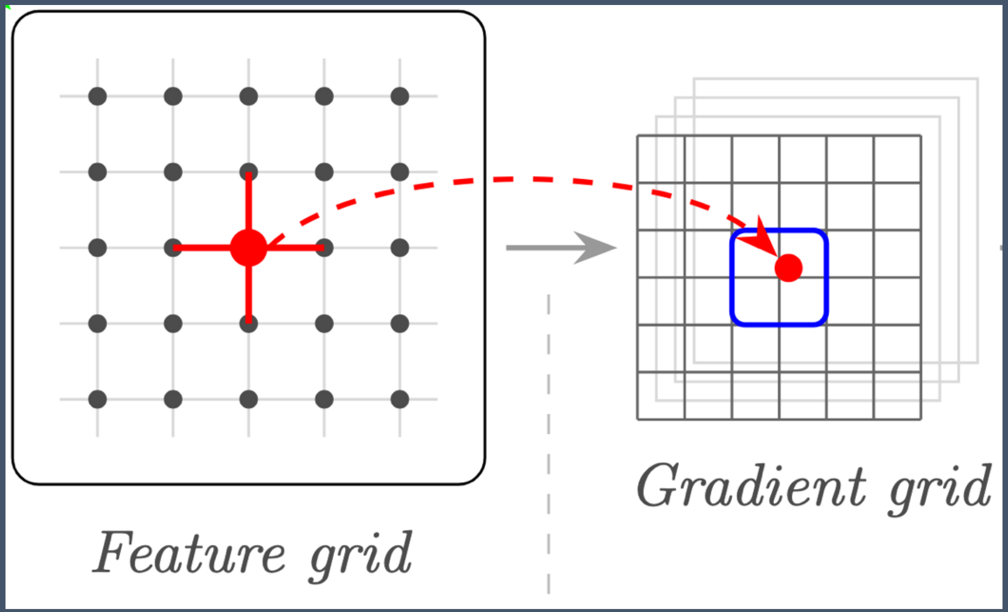


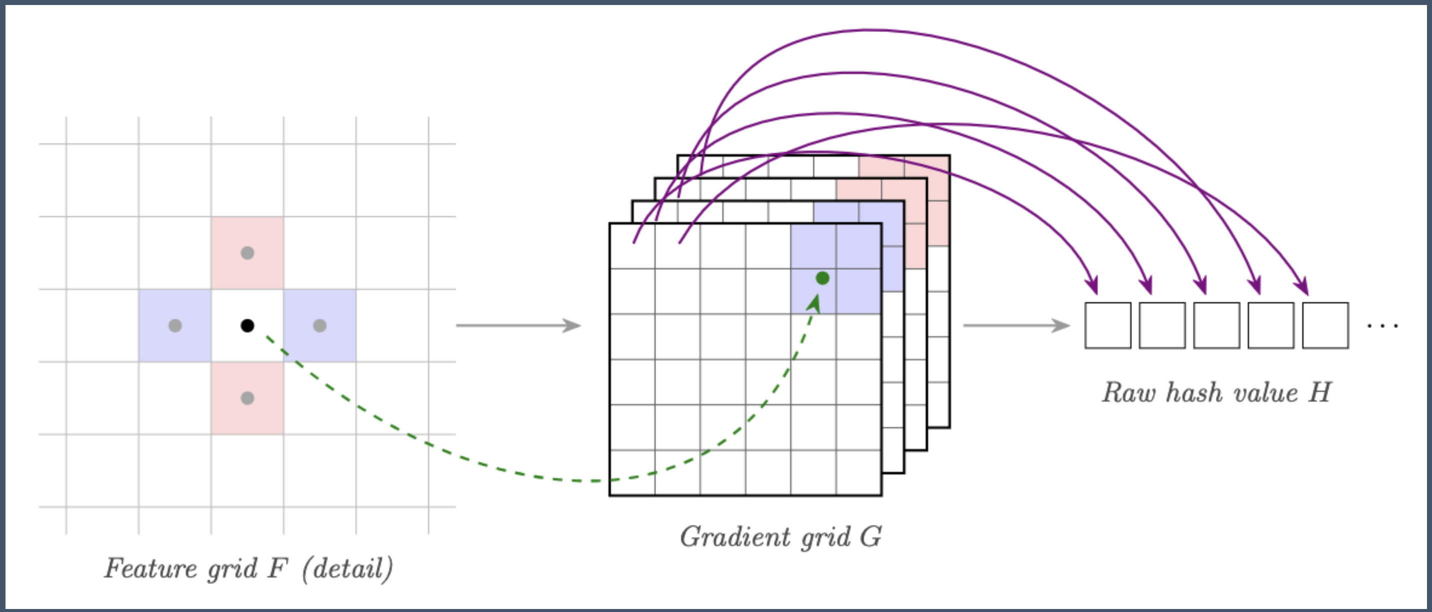
Dimension check (50x50)
Sum color values
Summation per square

26x26 feature grid capturing brightness and texture weighting of 3 rectangles

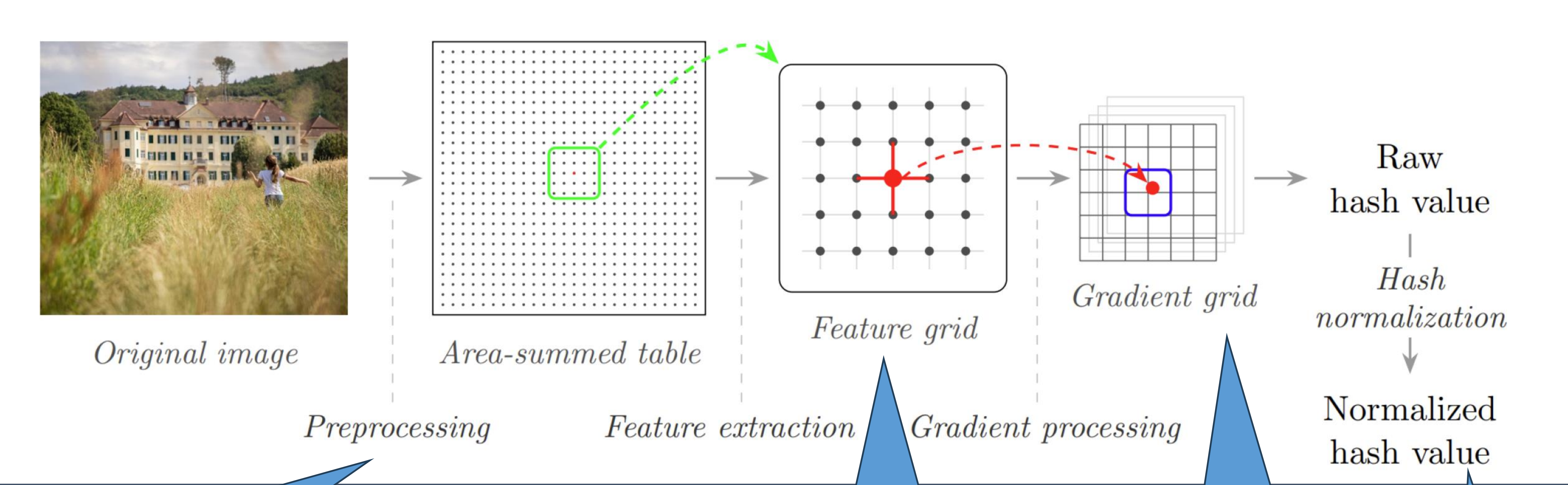
6x6 grid of 4 gradients (144 bytes)

Gradiant Grid Computation





PhotoDNA: description



Dimension check (50x50)
Sum color values
Summation per square

26x26 feature grid capturing brightness and texture weighting of 3 rectangles

6x6 grid of 4 gradients (144 bytes)

144 bytes or 1152 bits

PhotoDNA: properties

Strengths: robust against minor image modifications

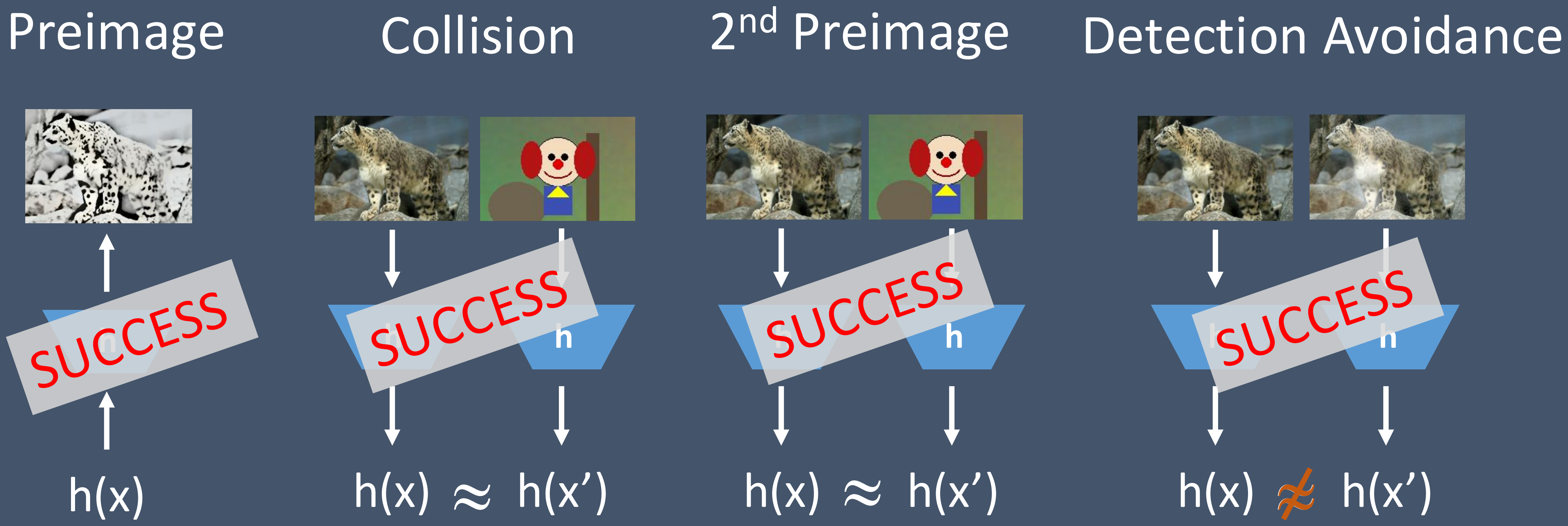
- resolution change (initial scaling)
- small shifts of features (splatting)
- brightness and contrast (gradient + normalization)

Weaknesses

- Lack of cropping procedure
- Sparse sampling (pixels do not have the same influence)
- Almost differentiable and piecewise linear
  - differentiable until last truncation (use straight-through estimator)
  - piecewise linear without normalization step (use raw hash values)
- Gradient-based optimization

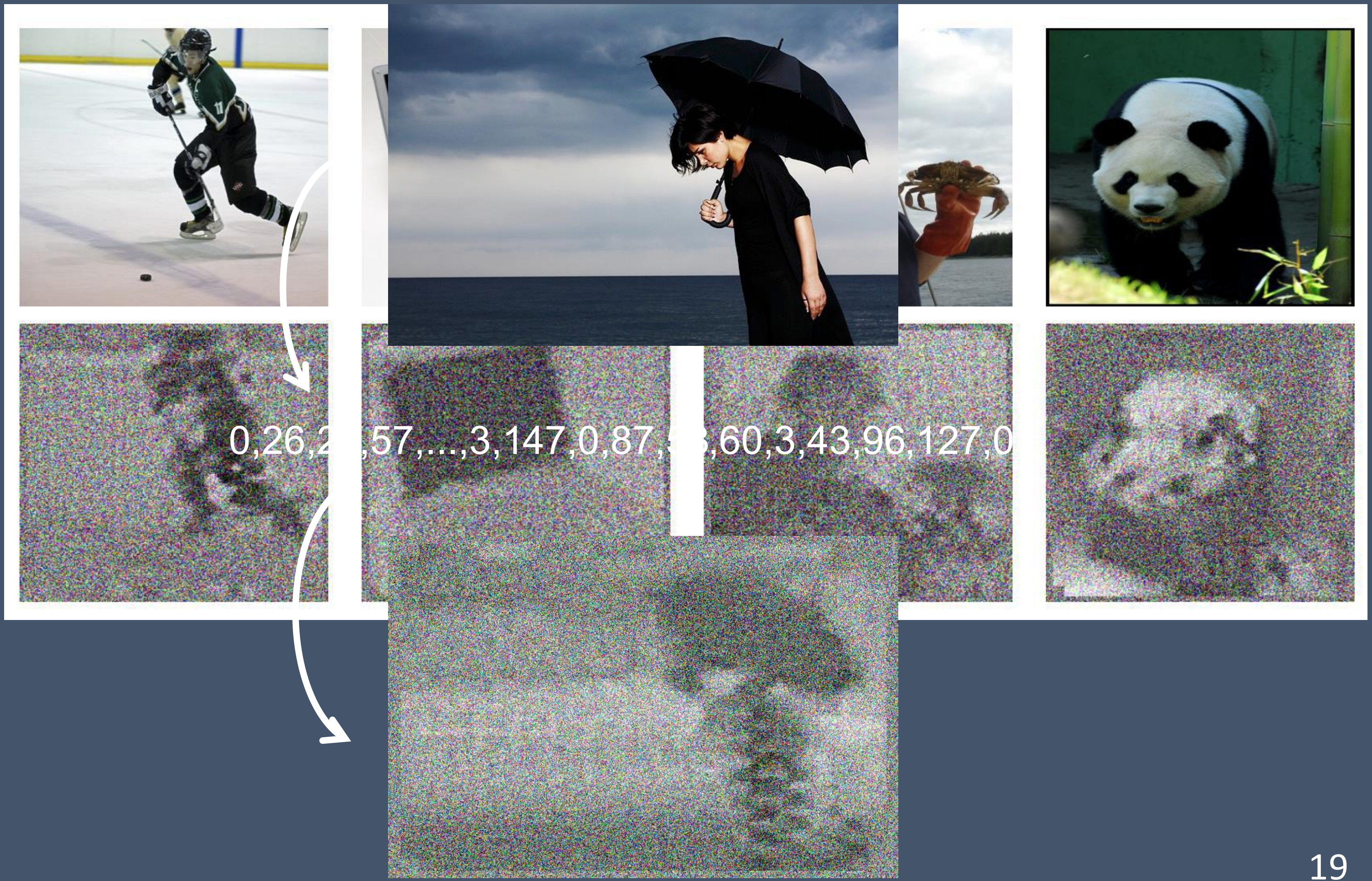
Slide credit: Maxime Deryck

Perceptual Hash Functions Attacks



Preimage attack: information leakage

- Gradient descent on L1 (hash) loss
- Sensitive to starting image: restarts needed
- 99.5% success rate
  - Remaining 5 images succeeded after manually adjusting learning rate
  - 88.0% success after first (re)start
- Mean time: 11.2 s for one restart (max. 2357 iterations)



Preimage Attack

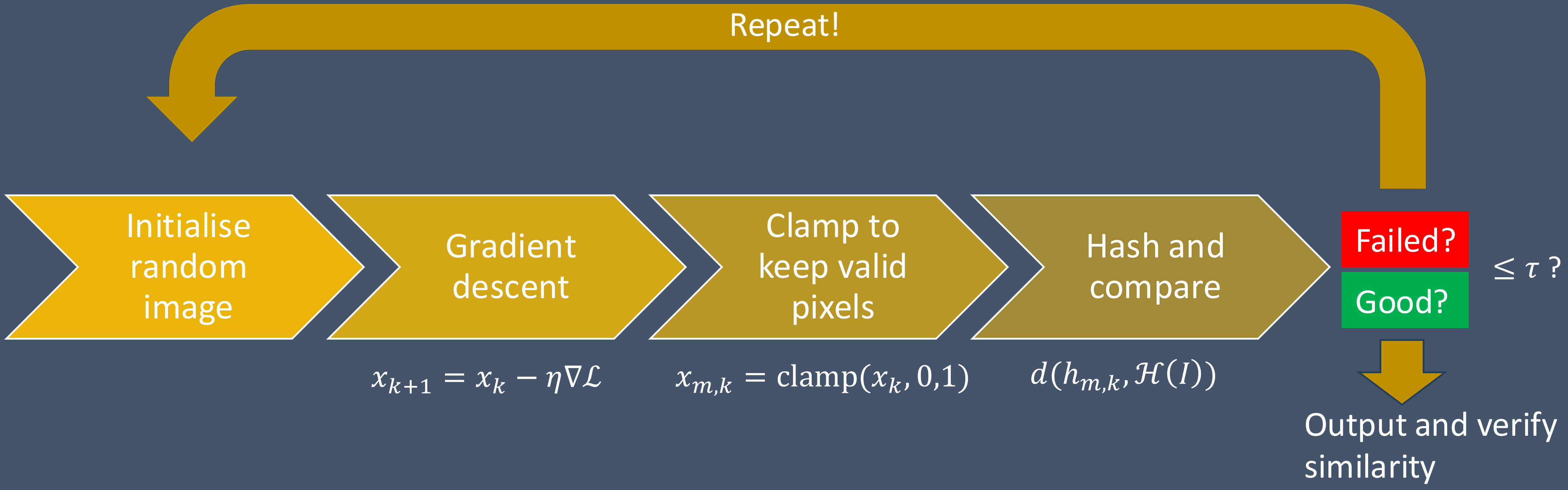
I	Target image
$\mathcal{H}(I)$	Target hash value
x_k	Vector representing I
$x_{m,k}$	Candidate vector at iteration k

$$\boxed{d(h_t, \mathcal{H}(I)) \le \tau}$$

$\mathcal{L}$: Loss function to minimize

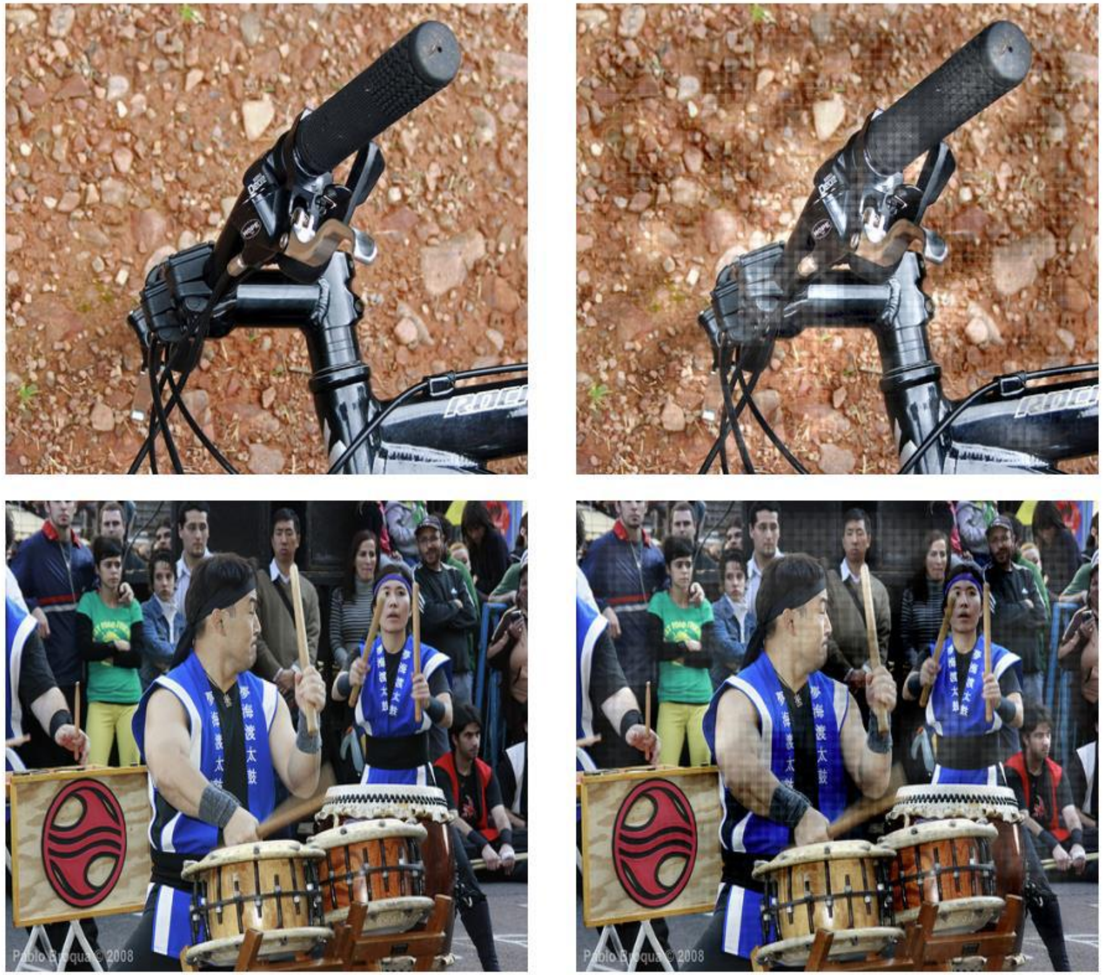
$\nabla\mathcal{L}(x_k)$: Gradient of $\mathcal{L}$ at x_k

$\eta \in \mathbb{R}^+$: Learning rate (step size)



Exact Collision attack

- Gradient descent on composite loss with focus on hash collision
- 100% success rate
- Fast convergence (mean: 4159 iterations)
- Mean time: 215.14 s (97.29 - 420.69 s)
- Can be improved



(a) Original images (b) Exact collision

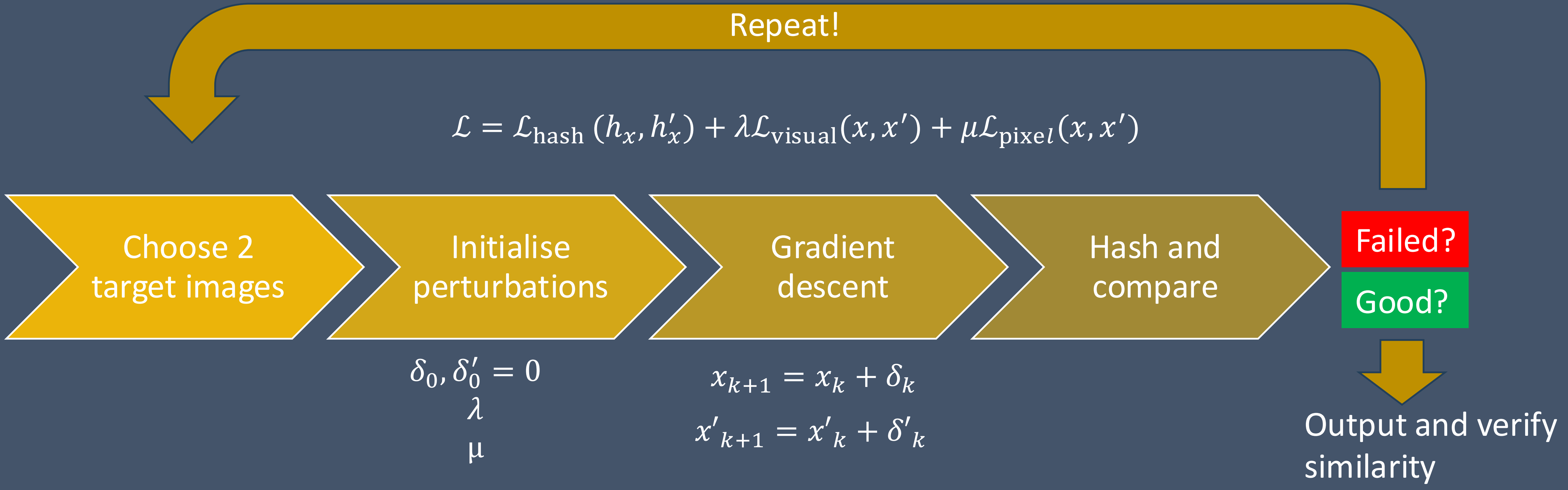
Slide credit: Bart Preneel &Maxime Deryck

Collision Attack

$d(\cdot)$	Distance measure
$\mathcal{H}(I)$	Hash value of target image I
$\mathcal{H}(I')$	Hash value of target image I'
τ	Threshold

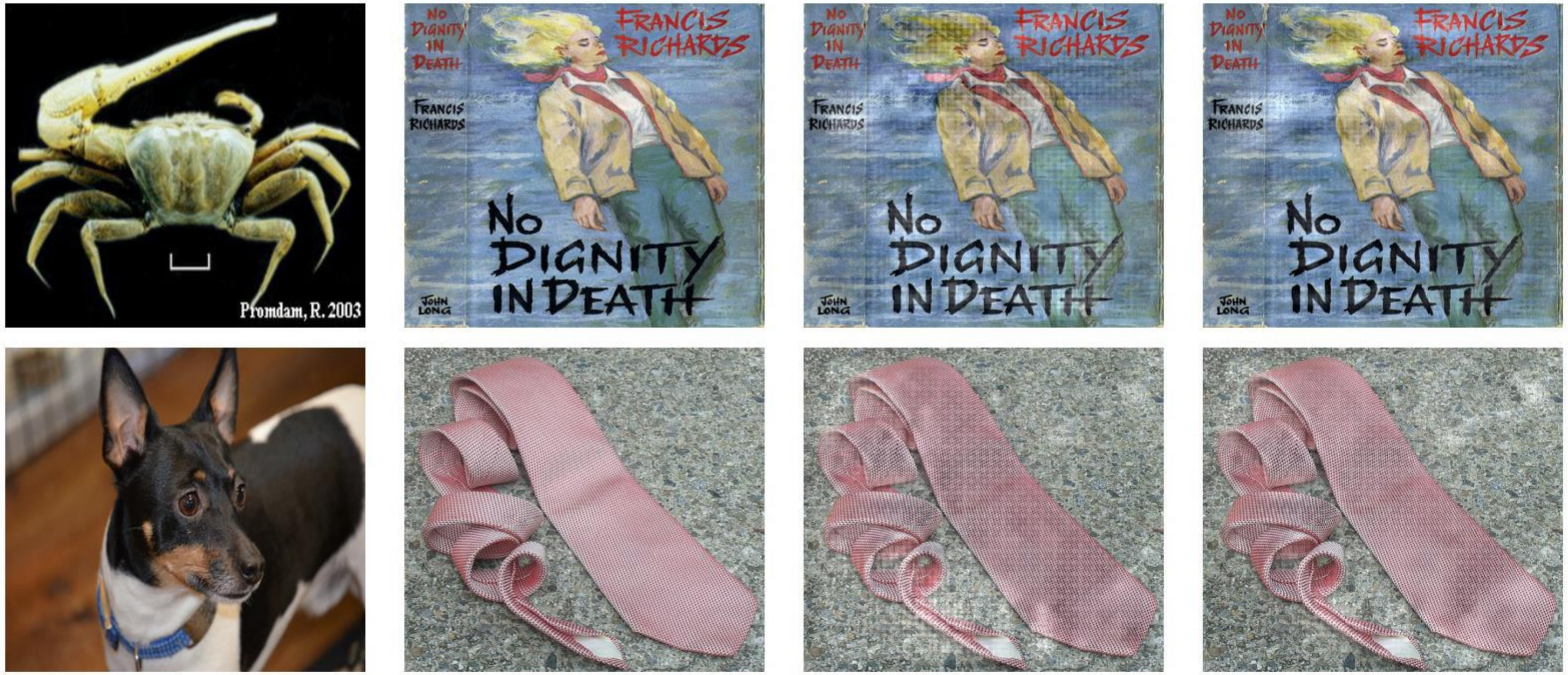
$$\boxed{d(\mathcal{H}(I), \mathcal{H}(I')) \leq \tau}$$

$\mathcal{L}_{\text{hash}}$	Make I''s hash match I's hash
$\mathcal{L}_{\text{visual}}$	Keep I' looking like I
$\mathcal{L}_{\text{pixel}}$	Keep the perturbation small



2nd preimage attack

- Gradient descent on composite loss with focus on hash collision
- Exact collision: 98.6% success rate; mean time 150 sec
- Near-collision: 99.7% success rate; mean time 118 sec



(a) Original image (b) Target image (c) Exact collision (d) Near-collision

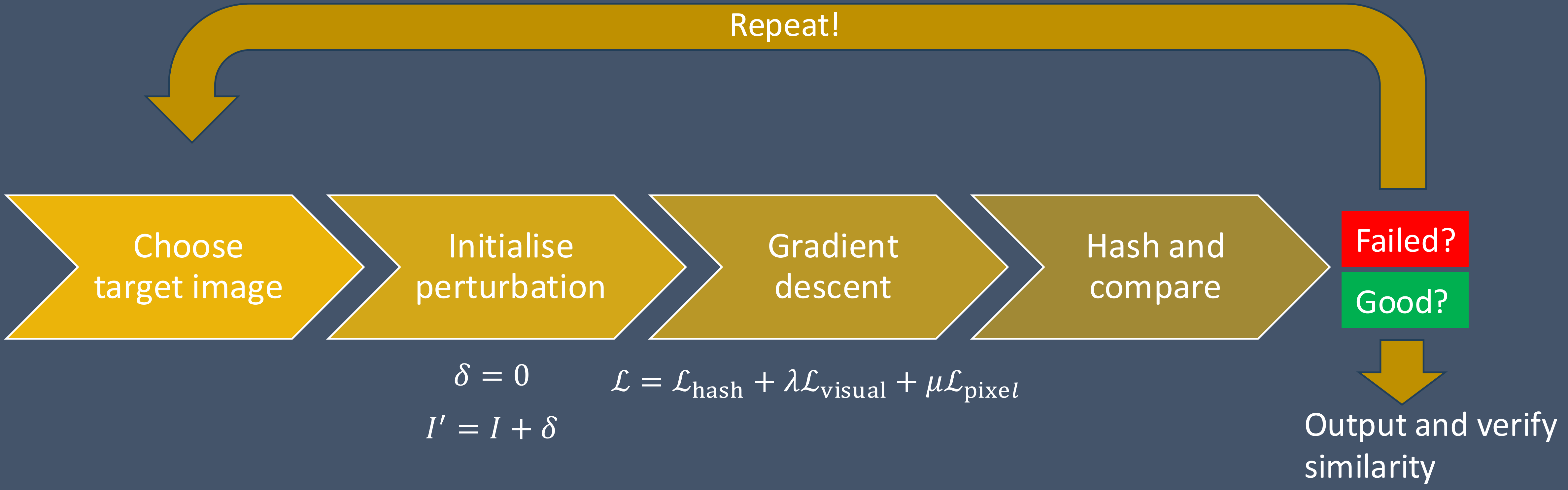
Slide credit: Bart Preneel &Maxime Deryck

Second Preimage Attack

$d(\cdot)$	Distance measure
$\mathcal{H}(I)$	Hash value of existing image
$\mathcal{H}(I')$	Hash value of target image
τ	Threshold

$$\boxed{d(\mathcal{H}(I), \mathcal{H}(I')) \leq \tau}$$

$\mathcal{L}_{\text{hash}}$	Make I''s hash match I's hash
$\mathcal{L}_{\text{visual}}$	Keep I' looking like the target image
$\mathcal{L}_{\text{pixel}}$	Keep the perturbation small



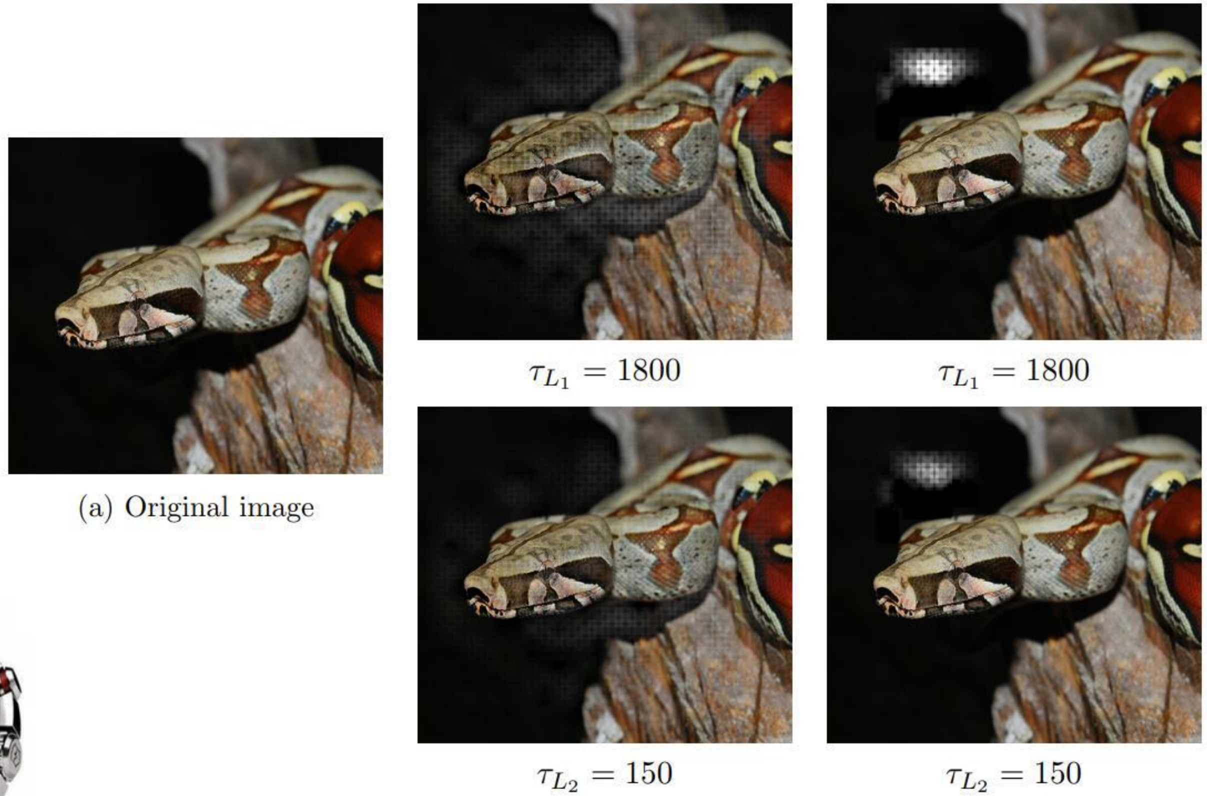
$\delta = 0$

$I' = I + \delta$

$\mathcal{L} = \mathcal{L}_{\text{hash}} + \lambda \mathcal{L}_{\text{visual}} + \mu \mathcal{L}_{\text{pixel}}$

Detection avoidance

- Global vs. localized changes
- 100% vs. 96.4% success rate
  - 2nd run with different orthogonal vector: 100% success rate
- Mean time: 9.94 vs. 78.48 s



(a) Original image (b) Origin direction (c) Orthogonal direction

(a) Original image (b) L_1 distance 836.0 (c) L_1 distance 1866.0

Slide credit: Bart Preneel & Maxime Deryck

Detection Avoidance

$d(\cdot)$	Distance measure
$\mathcal{H}(I)$	Hash value of target image
$\mathcal{H}(I')$	Hash value of modified image
τ	Threshold

$$\boxed{d(\mathcal{H}(I), \mathcal{H}(I')) > \tau}$$

$$\mathcal{L}(\delta_k) = \underbrace{\delta_k^T \delta_k}_{\text{Visual distortion penalty}} - \underbrace{\lambda v_k^T J_k \delta_k}_{\text{Hash movement reward}}$$

Repeat!

Choose target image → Initialise perturbation → Gradient descent → Hash and compare → Failed? / Good?

Initialise perturbation: $\delta = 0$, $I' = I + \delta$

Gradient descent: $x_{k+1} = x_k + \delta_k$

Output and verify similarity

In the press

English

- *Microsoft PhotoDNA is vulnerable to false positives and data leakage*, by Bill Mann — Featured in: Cyber Insider
- *EU Parliament rejects Chat Control message scanning*, by Bill Goodwin — Featured in: Computer Weekly
- *European Parliament Rejects Mass "Chat Control" Surveillance by Single Vote*, by Dan Frieth — Featured in: Reclaim The Net

Dutch

- *Onderzoekers KU Leuven ontdekken ernstige beveiligingslekken in PhotoDNA van Microsoft*, by Belga — Featured in: MSN
- *Leuvense onderzoekers kraken tool van Microsoft om kinderpornografie op te sporen: "Risico op valse beschuldigingen"*, by Roel Damiaans — Featured in: Het Belang van Limburg [Paywall]
- *De tool die al 17 jaar gebruikt wordt om beelden van kindermisbruik op te sporen, blijkt lek: "Beelden kunnen gemanipuleerd worden voor onterechte beschuldigingen"*, by Arthur De Meyer — Featured in: Nieuwsblad, Gazet van Antwerpen
- *Onderzoekers vinden kwetsbaarheden in systeem dat kinderporno moet opsporen*, by Els Bellens — Featured in: Trends DataNews
- *Belgisch onderzoek legt bom onder AI voor bestrijden kindermisbruik: lekken in software die TikTok en Instagram gebruiken*, by Kenneth Dée — Featured in: Het Laatste Nieuws [Paywall]

French

- *Une étude flamande révèle des failles de sécurité dans un logiciel réputé de Microsoft*, by Belga — Featured in: RTBF, La Libre, La Dernière Heure
- *Des chercheurs découvrent des failles dans un système de détection de contenus pédopornographiques*, by Els Bellens — Featured in: Le Vif
- *Une étude belge démontre que les logiciels IA censés protéger les mineurs sur TikTok et Instagram ne fonctionnent pas*, by Kenneth Dée — Featured in: 7sur7
- *Le système mondial de détection d'images pédopornographiques en ligne est boiteux*, by Philippe Laloux — Featured in: Le Soir [Paywall]
- *Le logiciel PhotoDNA de détection des contenus pédosexuels peut incriminer des innocents*, by Jean-Marc Manach — Featured in: Next [Paywall]
- *Pédopornographie en ligne : le scan européen expire le 3 avril 2026*, by Sarah Lebot — Featured in: France Jeunes
- *À 1 vote près, le Parlement européen bloque la surveillance de masse des GAFAM*, by Guillaume Belfiore — Featured in: Clubic
- *Journal de 3 avril à 12h30*, by Thomas Cluzel and Diane Leblanc-Albarel — Featured in: Radio France
- *La Commission européenne s'entête à vouloir espionner nos téléphones*, by Odile Benyahia-Kouider and Christophe Labbé — Featured in: Le Canard Enchaîné [Paywall]

Other

- *Ende der Chatkontrolle: EU-Parlament stoppt Massenscans im Abstimmungskrimi - Weg frei für echten Kinderschutz*, by Patrick Breyer (German) — Featured in: Patrick Breyer
- *Chat Control: niente scansione chat private dal 3 aprile 2026*, by Michele Nasi (Italian) — Featured in: Il Software
- *Parlamento Ue blocca la proroga al controllo dei contenuti privati nelle comunicazioni digitali*, by Michele Ficara Manganelli (Italian) — Featured in: Assodigitale
- *Evropský parlament čelí bezprecedentnímu lobbistickému tlaku: Cílem je zvrátit statečné stanovisko podporující soukromí*, by Hynek Trojánek (Czech) — Featured in: Digitální svobody
- *Parlamentul European respinge prelungirea normelor „Chat Control" privind scanarea mesajelor private*, by Tiberiu Nan (Romanian) — Featured in: Go4IT
- *Parlamentul European respinge scanarea mesajelor Chat Control*, by Bogdan Marchidanu (Romanian) — Featured in: Club It&C
- *"Lỗ hổng" trong hệ thống phát hiện ảnh lạm dụng trẻ em trực tuyến*, by Hương Giang (Vietnamese) — Featured in: Vietnam Plus

Status: https://pseudodna.eu/

- PhotoDNA does not achieve its design goals:
  - 1 in 50 billion false positives
  - 1% false negatives
  - irreversibility
- M. Deryck, D. Leblanc-Albarel, B. Preneel, White-Box Attacks on PhotoDNA Perceptual Hash Function, IACR eprint 2026/486 – under review

Slide credit: Bart Preneel

Bonus results: White-box attacks on PDQ

- 2019: https://github.com/facebook/ThreatExchange/tree/main/pdq

- Normalization step: processes the luminance data
- Downsample to 64x64: two-pass Jarosz filter
- Compute sum of absolute values of horizontal and vertical gradients
- Apply 16x16 DCT (Discrete Cosine Transform)
- Threshold for 256 DCT values results in 0 or 1

White-box attacks on PDQ

- Hamming distance (threshold 90 out of 256)
- Earlier attacks known (require about 3 hours for a second preimage on 72-core Intel Xeon CPU E5-2695 v4 without a dedicated GPU)
- Earlier attacks used unrealist thresholds for some attacks
- Substantial improvement: new attacks, faster, closer distance, less noisy

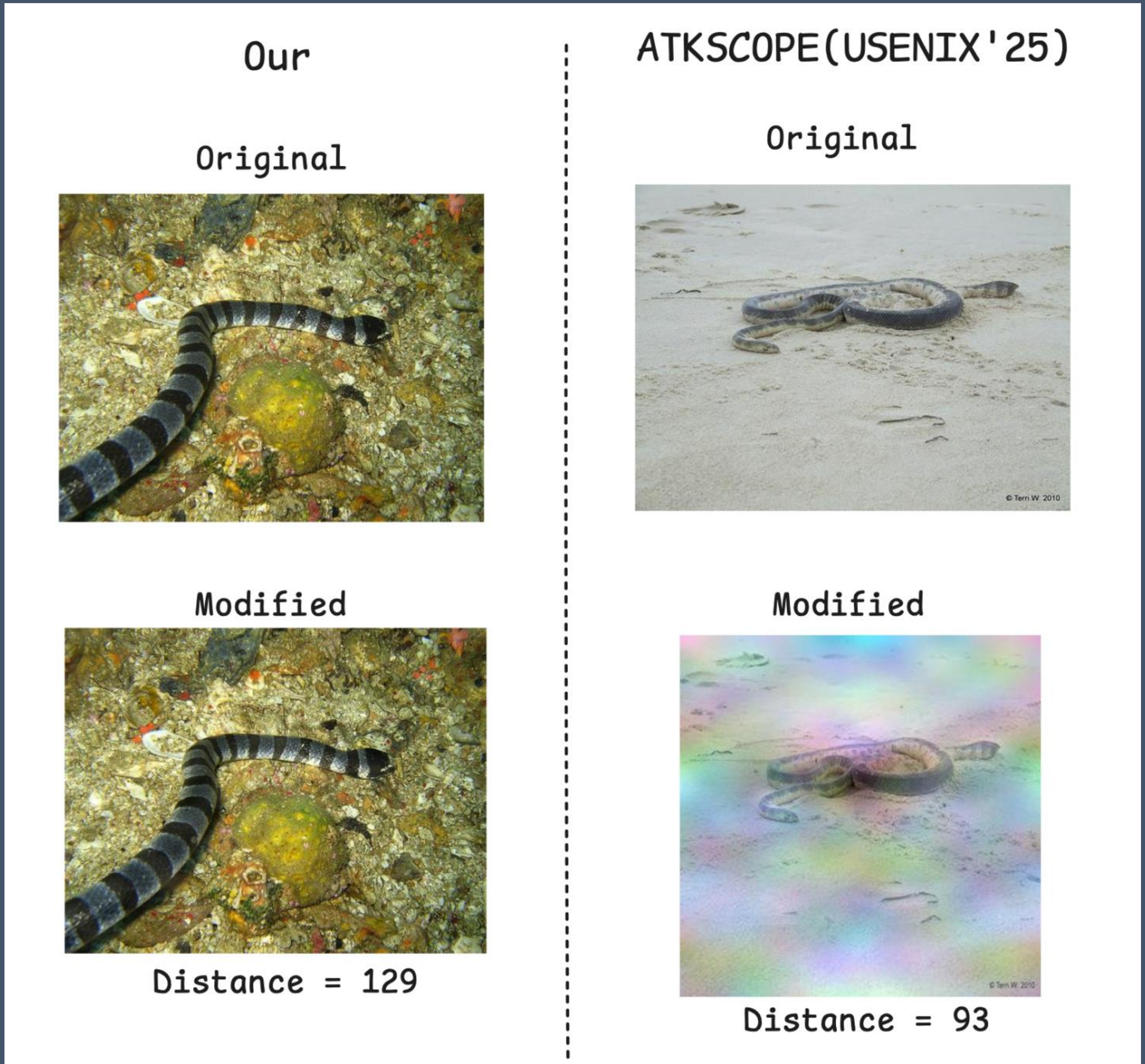
S. Jain, A.-M. Cretu, Y.-Al. de Montjoye. Adversarial detection avoidance attacks: Evaluating the robustness of perceptual hashing-based client-side scanning. USENIX Security Symposium, 2022

J. Prokos, N. Fendley, M. Green, R. Schuster, E. Tromer, T.M. Jois, Y. Cao: Squint Hard Enough: Attacking Perceptual Hashing with Adversarial Machine Learning. USENIX Security Symposium 2023: 211-228

Y. Zhang, Y. Sun, S. Qi, Z. Hua, W. Wen, Y. Fang: ATKSCOPES: Multiresolution Adversarial Perturbation as a Unified Attack on Perceptual Hashing and Beyond, Usenix Security Symposium, 2025

Detection avoidance

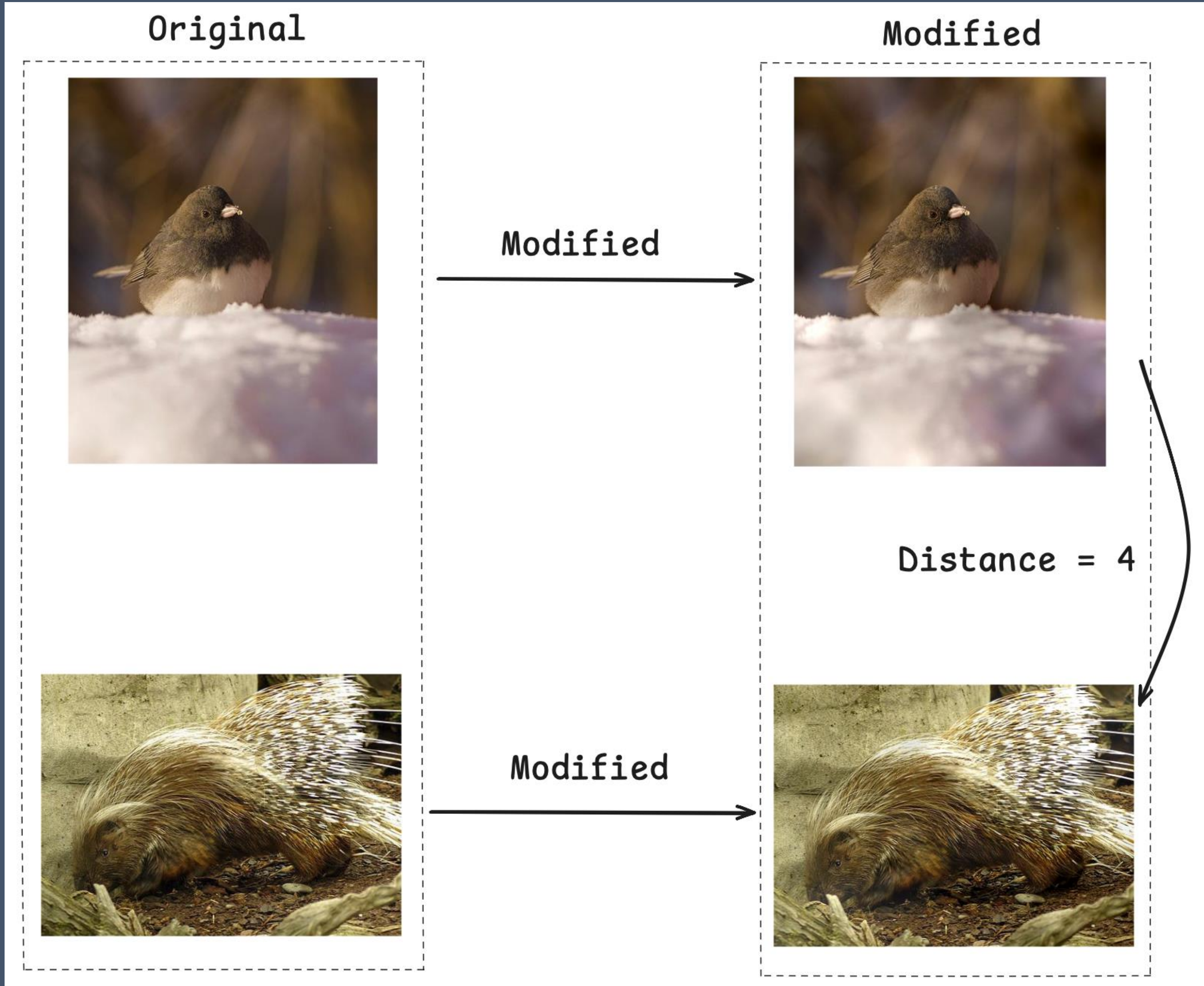
Metric	OUR	ATKSCOPE (USENIX'25)
Threshold	92	92
Avg. Distance	129.96	92.64
Avg. LPIPS	0.022203	0.45
Success Rate	100%	100%
Avg. Runtime	$\sim 2.414 sec/image$	Not given
Avg. Rounds	1	69.86



Slide credit: Madhav Verma

Near-collision

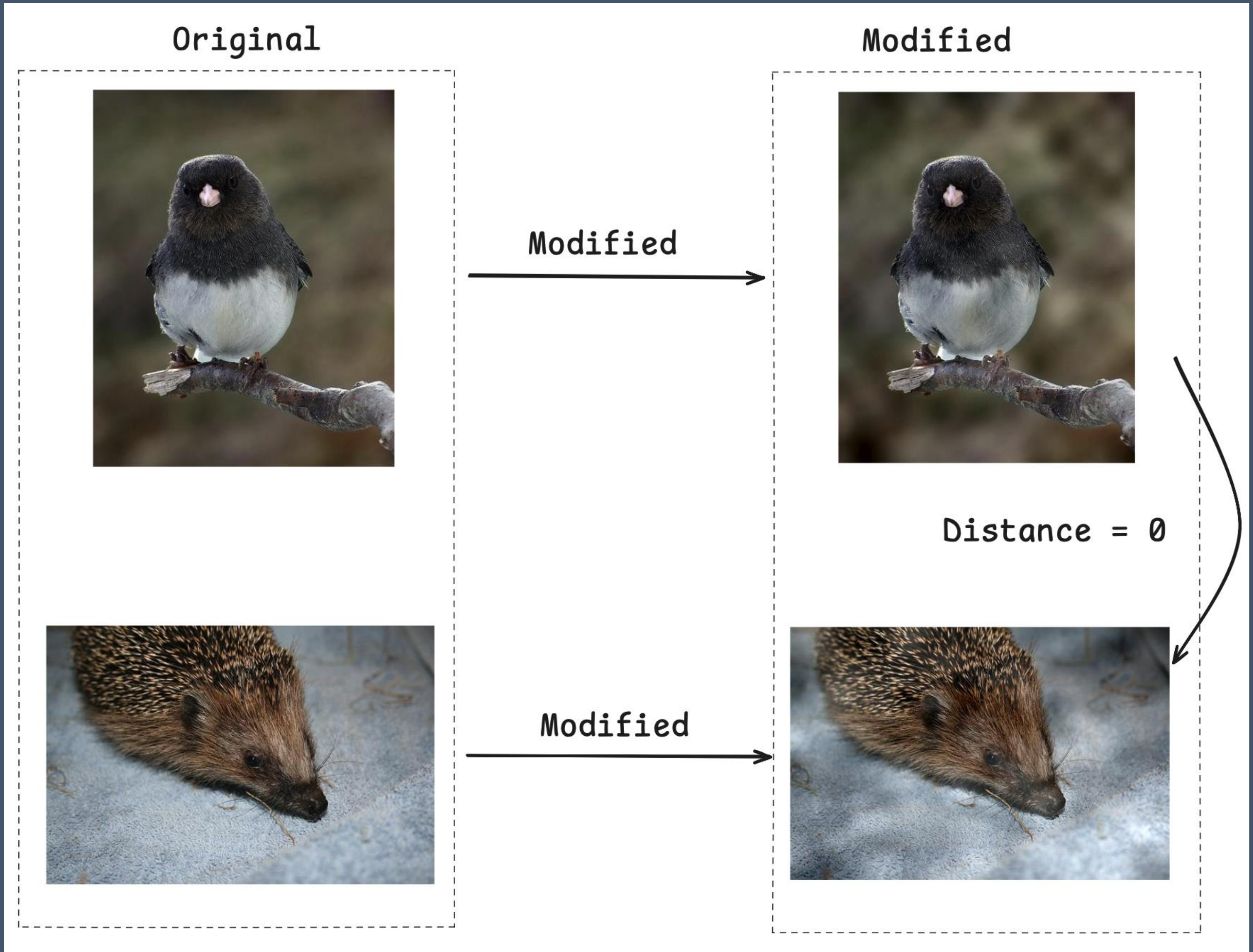
Metric	OUR
Threshold	31
Avg. Distance	8.36
Avg. LPIPS (X')	0.0396
Avg. LPIPS (Y')	0.0401
Success Rate	100%
Avg. Runtime	$\sim 1.1 sec/pair$
Avg. Rounds	1



Slide credit: Madhav Verma

Exact collision

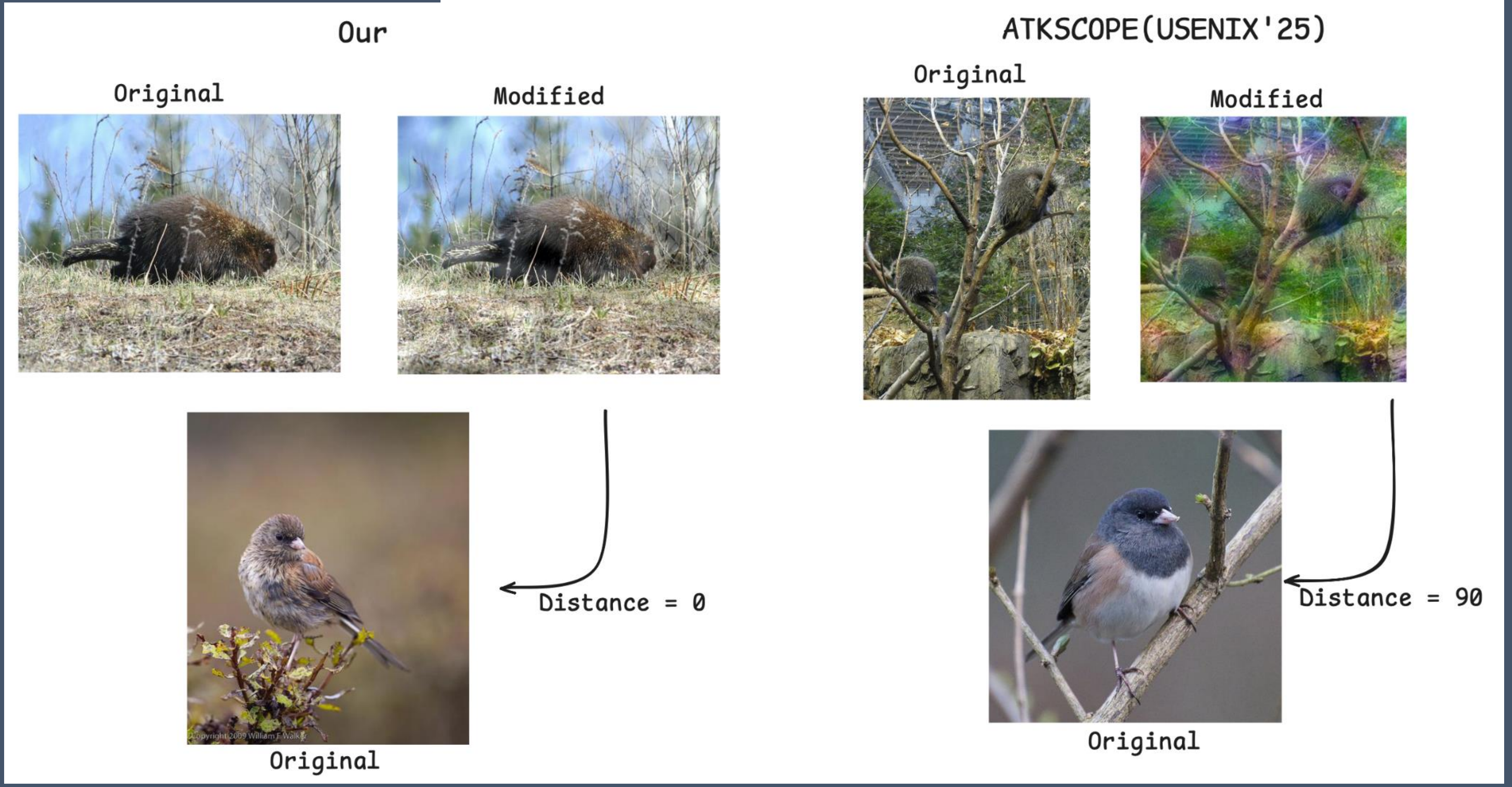
Metric	OUR
Threshold	31
Avg. Distance	0
Avg. LPIPS (X')	0.0867
Avg. LPIPS (Y')	0.0804
Success Rate	100%
Avg. Runtime	$\sim 4.1 sec/pair$
Avg. Rounds	3.27



Slide credit: Madhav Verma

Second preimage

Metric	OUR	ATKSCOPE (USENIX'25)
Threshold	31	92
Avg. Distance	0	90.80
Avg. LPIPS	0.1037	0.42
Success Rate	100%	100%
Avg. Runtime	$\sim 1.976 sec/image$	$\sim 4.2 min/image$
Avg. Rounds	3.5	58.60



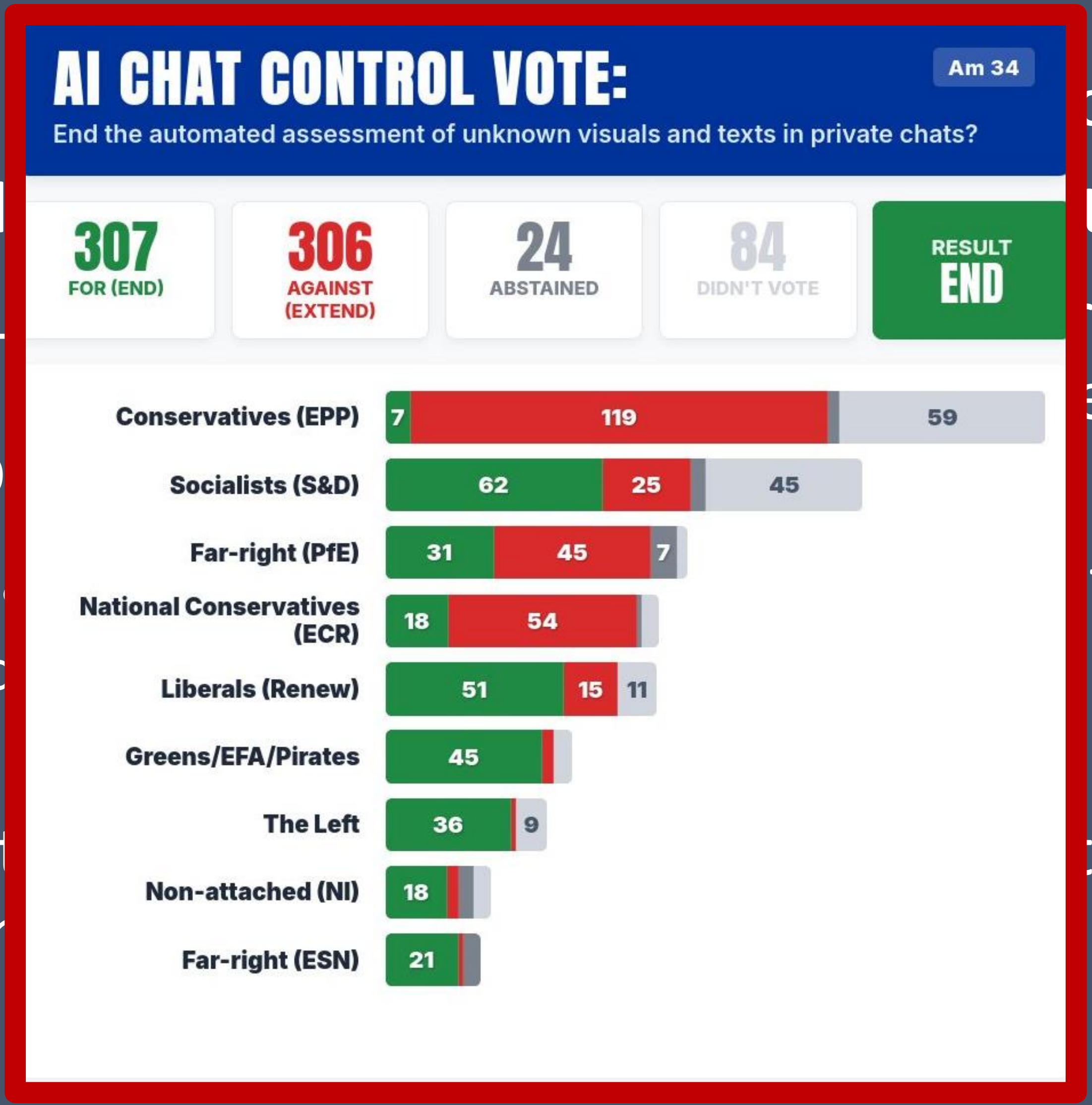
Slide credit: Madhav Verma

Preimages in < 1 second

Slide credit: Madhav Verma

CSAM Detection in Europe

- 2009: ePrivacy dire cation not allowed
- 2014: Massive depl tection
- 2021 ChatControl 1 CSAM detection volountary scan of e.g., by Instagram, Facebook, Gmail, D
- ChatControl 2.0: M te communication?
  - Regurlarly change
  - Up to date : Still in
- 26 March 2026: Vo tion exeption (ChatControl 1.0) th



Victory?

Chat Control 1.0

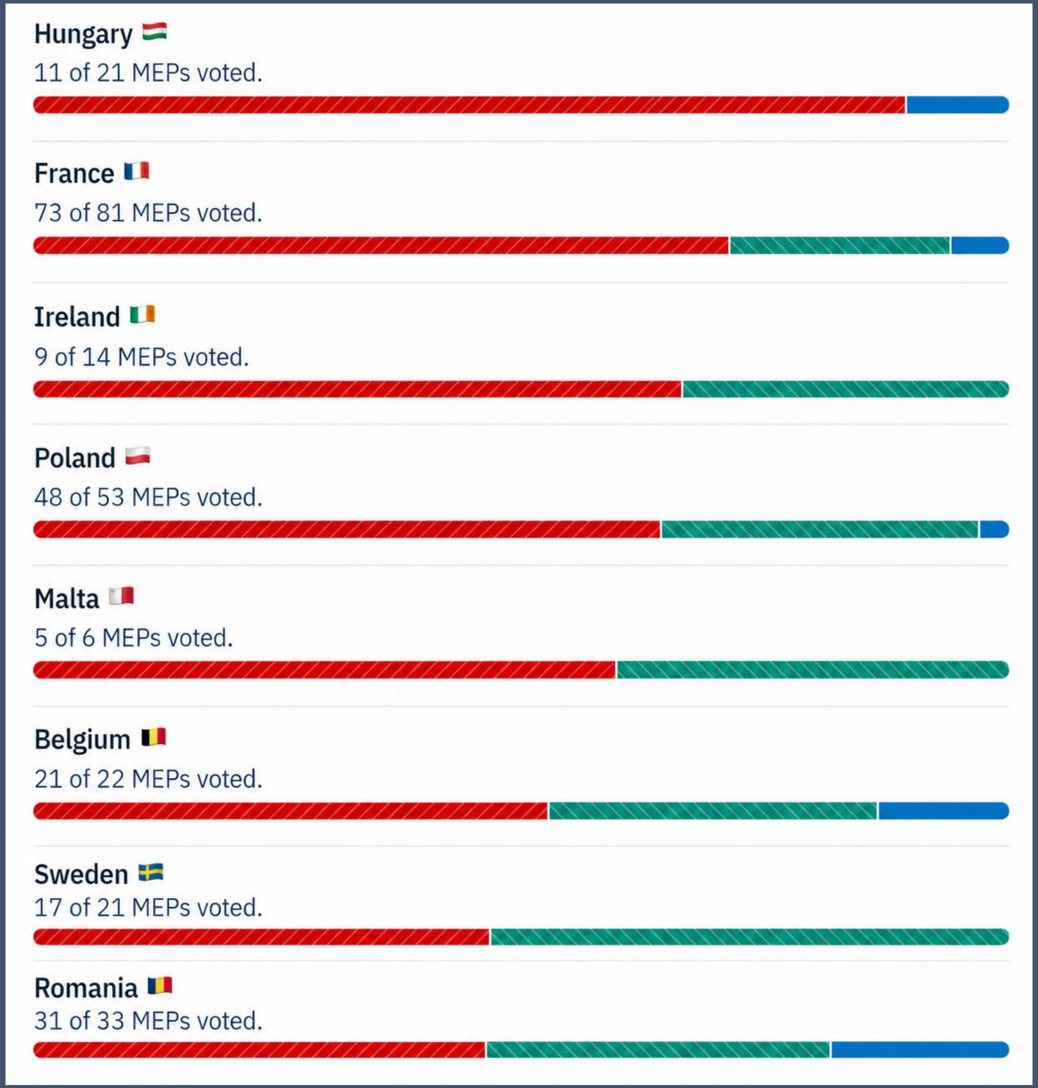
Currently a time-limited interim **regulation** is in place *allowing* providers to scan communications indiscriminately if they choose so (so-called "Chat Control 1.0"). So far, only unencrypted US communication services such as Gmail, Facebook/Instagram Messenger, Skype, Snapchat, iCloud Mail, or Xbox make use of this (overview **here**). In 2024, the EU adopted an extension of the voluntary Chat Control 1.0 regulation by two years, until April 3, 2026 – see timeline and documents. In March 2026, the European Parliament rejected a second extension of the regulation. It expired on 3 April 2026. However Google, Meta, Microsoft and Snap **announced** that they will continue indiscriminate and warrentless scanning of private messages. Users can **complain to their data protection authority** about this violation of the secrecy of communications (Article 5 Directive 2002/58). An abuse survivor is **taking legal action** against Chat Control 1.0.

Chat Control 2.0

On 11 May 2022 **the European Commission presented a proposal which would have made chat control searching *mandatory* for all e-mail and messenger providers** and would even have applied to so far securely end-to-end encrypted communication services. Prior to the proposal a public consultation had revealed that **a majority of respondents, both citizens and stakeholders, opposed imposing an obligation to use chat control.** Over 80% of respondents opposed its application to end-to-end encrypted communications.
Parliament has positioned itself almost unanimously against indiscriminate chat control. Following protests, mandatory chat control was removed also from the EU governments' draft law late 2025. Nevertheless, the government's draft regulation remains **deeply flawed**.
The final content of the Chat Control 2.0 regulation is currently being negotiated in a trilogue (**timeline here**). What's at stake: Permanent indiscriminate mass surveillance of messaging and chats at the discretion of providers, ineffective network blocking, mandatory age verification resulting in the end of anonymous communication, appstore censorship and excluding minors from the digital world.

And now?

- Security by obscurity has failed
- Client-side scanning require publicly known function and database
- Perceptual hash functions not ready for prime-time



Talk to policymakers! Sign open letters!

- Vie privée ou sécurité : jusqu'où faut-il aller ? (AOC média)
- La Commission européenne s'entête à vouloir espionner nos téléphones (Le Canard enchainé)
- Journal 12H30, 3 Avril 2026 (France Culture)

Contact/Webpage

diane.leblanc-albarel@esat.kuleuven.be

https://dianeleblancalbarel.github.io/