Watermarking Before and After Deep Learning

Mauro Barni,. Benedetta Tondi
University of Siena, Siena, Italy
Visual Information Processing and Protection (VIPP) Group

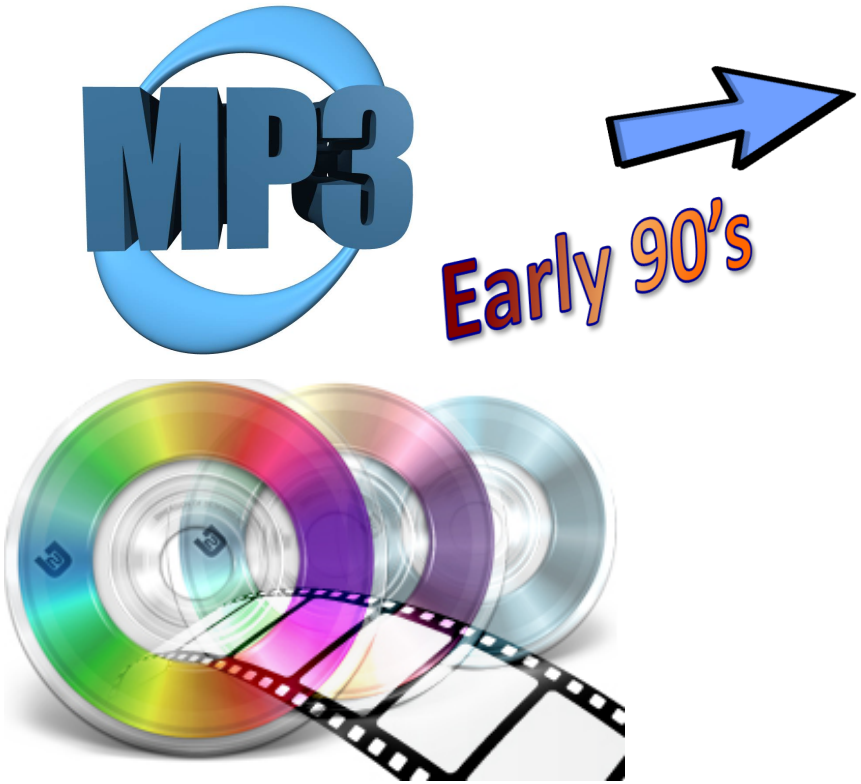


Outline

- Classical watermarking: history and foundations
  - Lessons learnt (and then forgotten) between 1995 and 2010
- Watermarking in the AI age
  - Old and new challenges

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

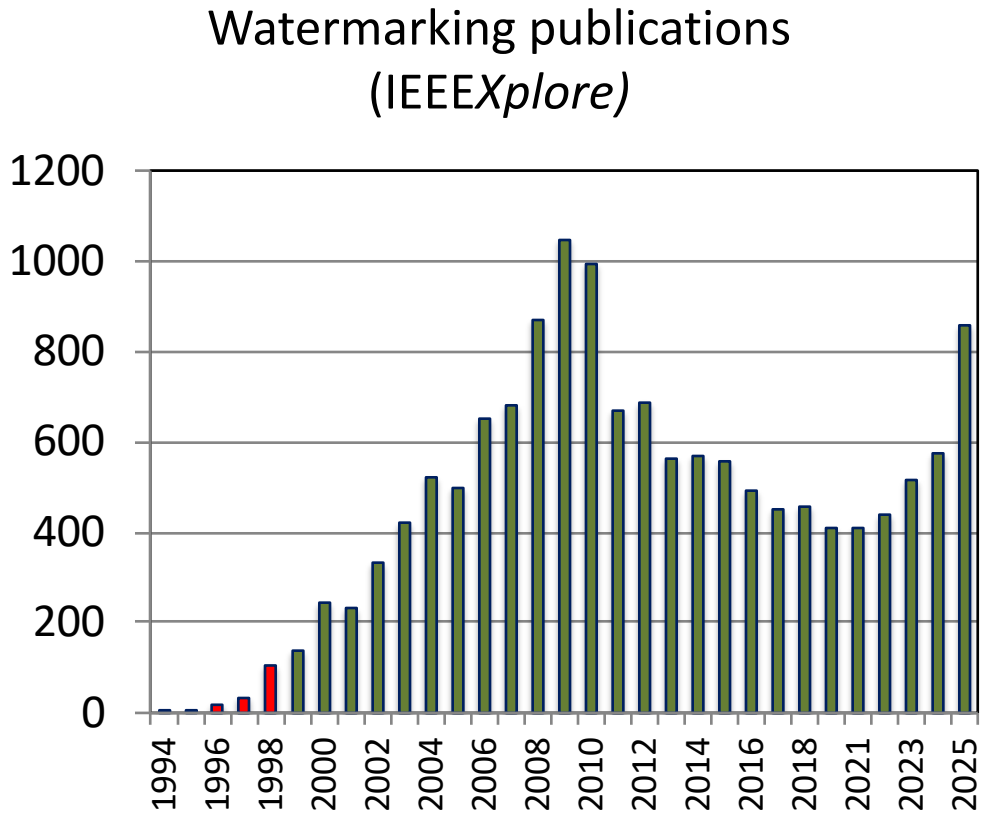
The early days: copy control



- Copyright concerns: revenues from digital music and digital video at risk - after the digital revolution (middle of nineties)
- DRM (digital risk management): a set of technology to prevent non-authorized viewing, illegal copying, editing, distributing of copyrighted material – based on cryptography
- Role of watermarking: an additional layer of security surviving D2A and A2D conversion

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

The watermarking history



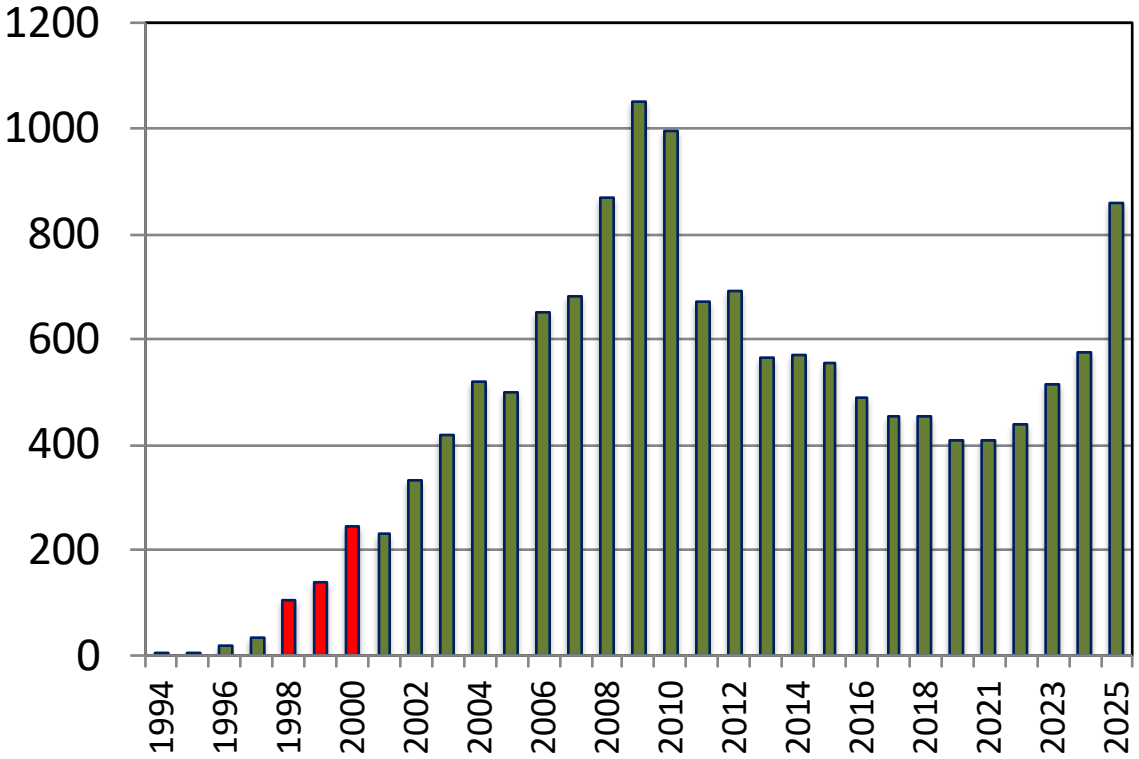
('94 - '98) First solutions

- First watermarking systems were **rather naïf**
  - Reference theory was missing
  - Terminology was confused
  - Application-based requirements were not fully understood
  - Security and robustness confused

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

The spread spectrum paradigm (’98 - ’00)

Watermarking publications (IEEE*Xplore)*





Cox, I. J., Kilian, J., Leighton, F. T., & Shamoon, T. (1997). Secure spread spectrum watermarking for multimedia. IEEE Trans. on Image Processing, 6(12), 1673-1687



Hartung, F., & Girod, B. (1998). Watermarking of uncompressed and compressed video. Signal processing, 66(3), 283-301

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

A theory for multi-bit watermarking

- SS communication provided the reference theory for multi-bit watermarking
- Analogy with secret communications in very noisy, hostile environment

$$y_n = x_n + \gamma b w_n \quad \textbf{(Encoding)}$$

x_n = sequence with host coefficients

w_n = pseudo-random spreading sequence (secret key)

γ = watermark energy

$b \in \{-1, 1\}$ = embedded bit

$$\hat{b} = \text{sign}(\sum_i y_i w_i) \quad \textbf{(Decoding)}$$

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

A theory for 0-bit watermarking

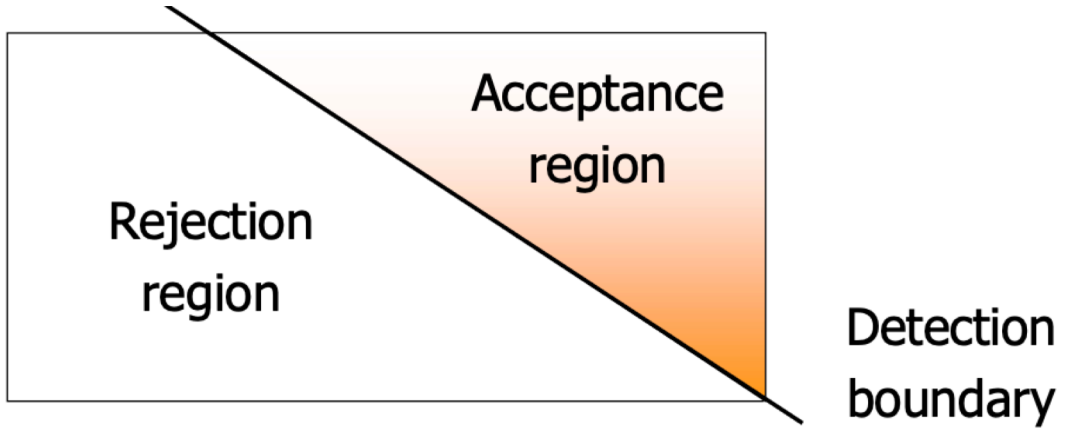
- Hypothesis testing provided the reference theory for watermark detection

$$y_n = x_n + \gamma w_n$$

$$z = \frac{1}{m} \sum_i y_i w_i$$

$$\text{Hp0:}\ \ z \in \mathcal{N}(0, \sigma^2/m)$$
$$\text{Hp1:}\ \ z \in \mathcal{N}(\gamma, \sigma^2/m)$$

$$z \gtrless \lambda$$ watermark present/absent



Other lessons learnt in late nineties

- **Importance of perceptual modeling**

  The **human visual system (HVS)** does not perceive all distortions equally. Watermark strength can be adapted based on: **texture, edges**, **luminance, frequency sensitivity**

  - Podilchuk, C. I., & Zeng, W. (1998). Image-adaptive watermarking using visual models. IEEE Journ. Selected Areas in Communications, 16 (4), 525-539.
  - Barni, M., Bartolini, F., & Piva, A. (2001). Improved wavelet-based watermarking through pixel-wise masking. IEEE Trans. on Image Processing, 10 (5), 783-791

- **Application-based requirements**
  - Each application has its own requirements
  - System level solutions are needed

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Other lessons learnt in late nineties

- **The importance of benchmarking**
  - The necessity of comparing different systems on a fair ground became evident

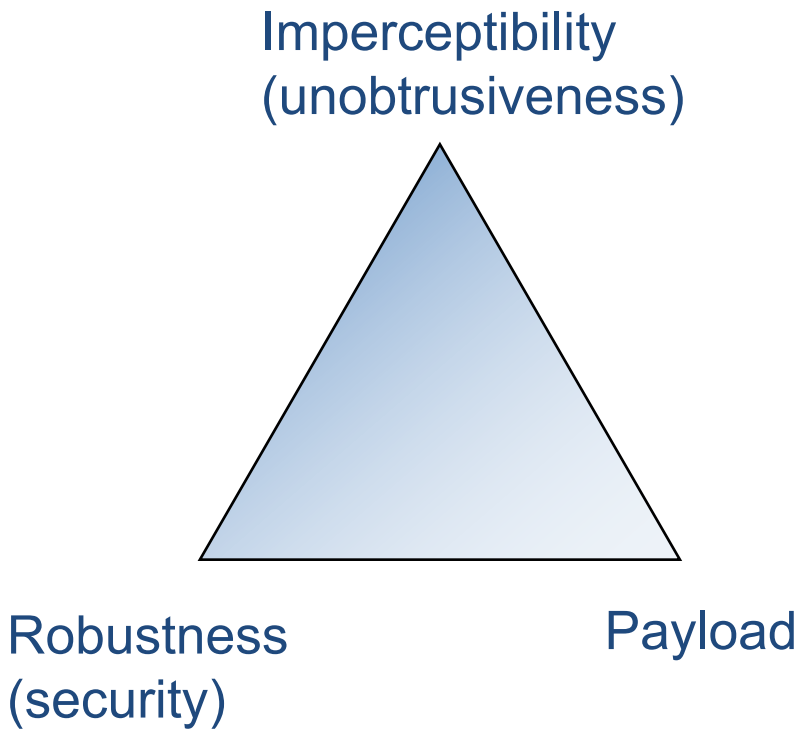


- Petitcolas, F. A., Anderson, R. J., & Kuhn, M. G. (1998, January). Attacks on copyright marking systems. In Information Hiding (pp. 218-238). Springer Berlin Heidelberg.
- Kutter, M., & Petitcolas, F. A. (1999). A fair benchmark for image watermarking systems. Electronic Imaging, 99, 219-239.

  - *Stirmark software developed and released in 1997*

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

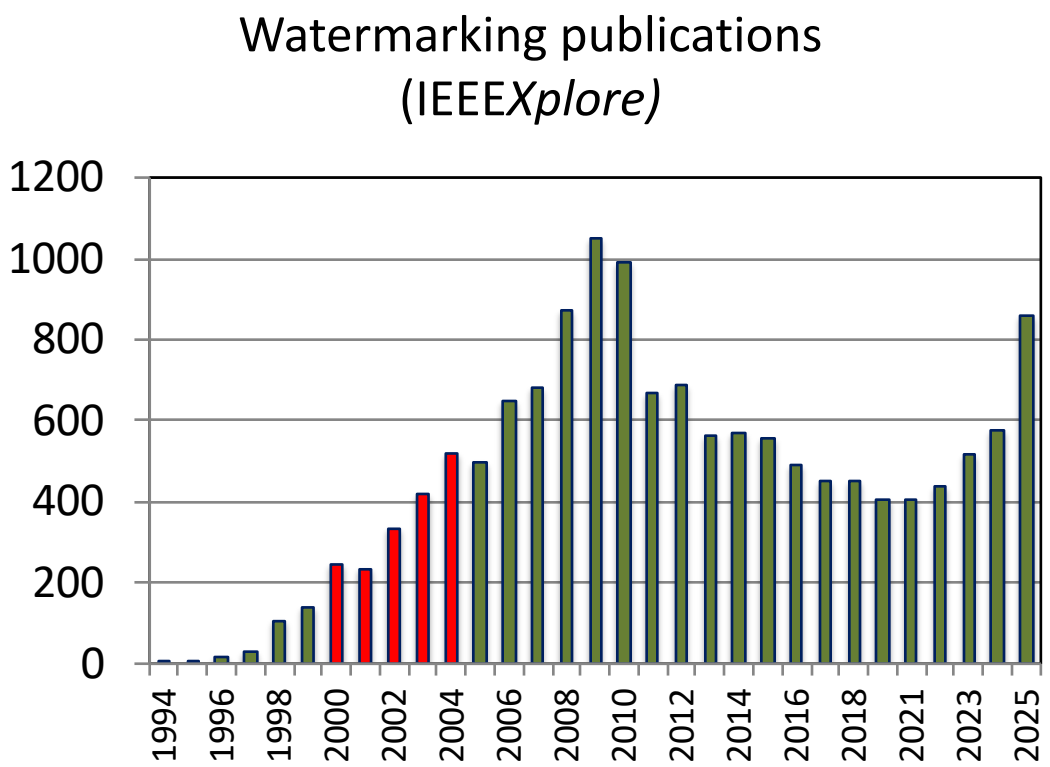
The watermarking trade-off triangle



The basic trade-offs to be considered in *any* watermarking system, were summarized in the watermarking trade-off triangle

Still mentioned nowadays (AI watermarking), but not widely considered

The informed watermarking age (00-04)



- A theoretical revolution
- New approach to watermarking (besides SS)
- Deep links with IT theory
- *Communication over a channel with side information at the encoder*

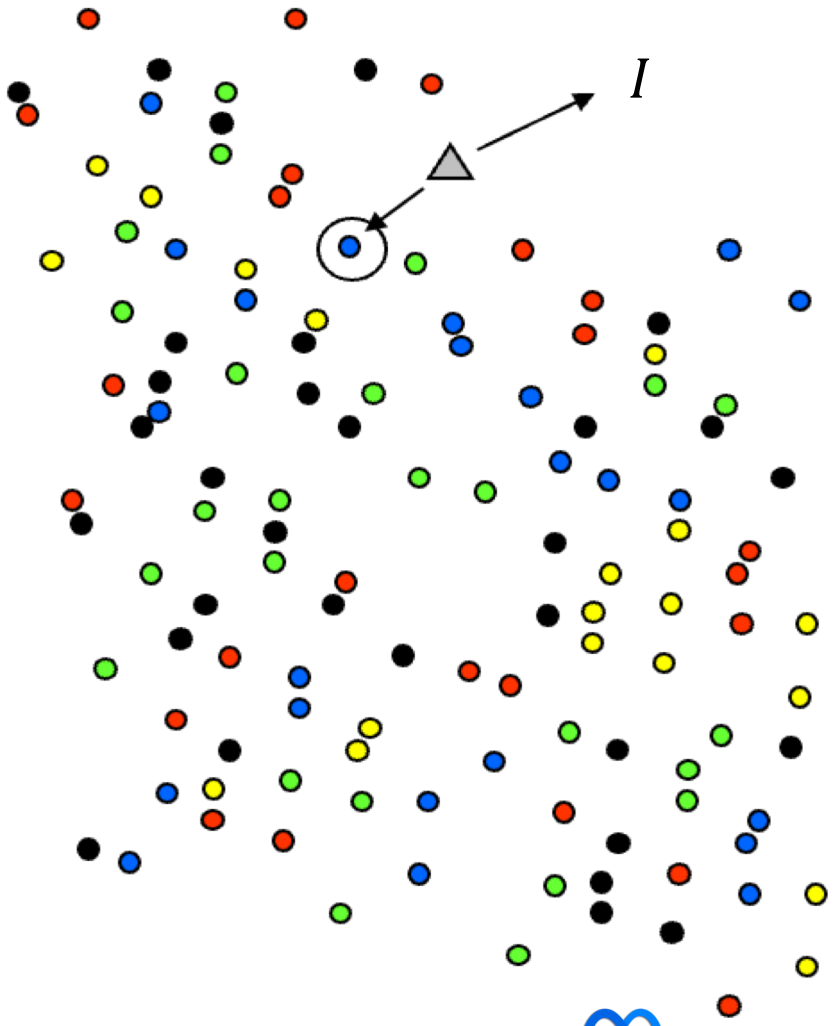


J. Cox, M. L. Miller and A.L. McKellips. "Watermarking as communications with side information." Proceedings of the IEEE vol. 87 n. 7 (1999), pp. 1127-1141.



B. Chen and G. W. Wornell. "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding." IEEE Trans. Information Theory, IEEE vol. 47 n. 4 (2001), pp. 1423-1443.

QIM and side informed watermarking



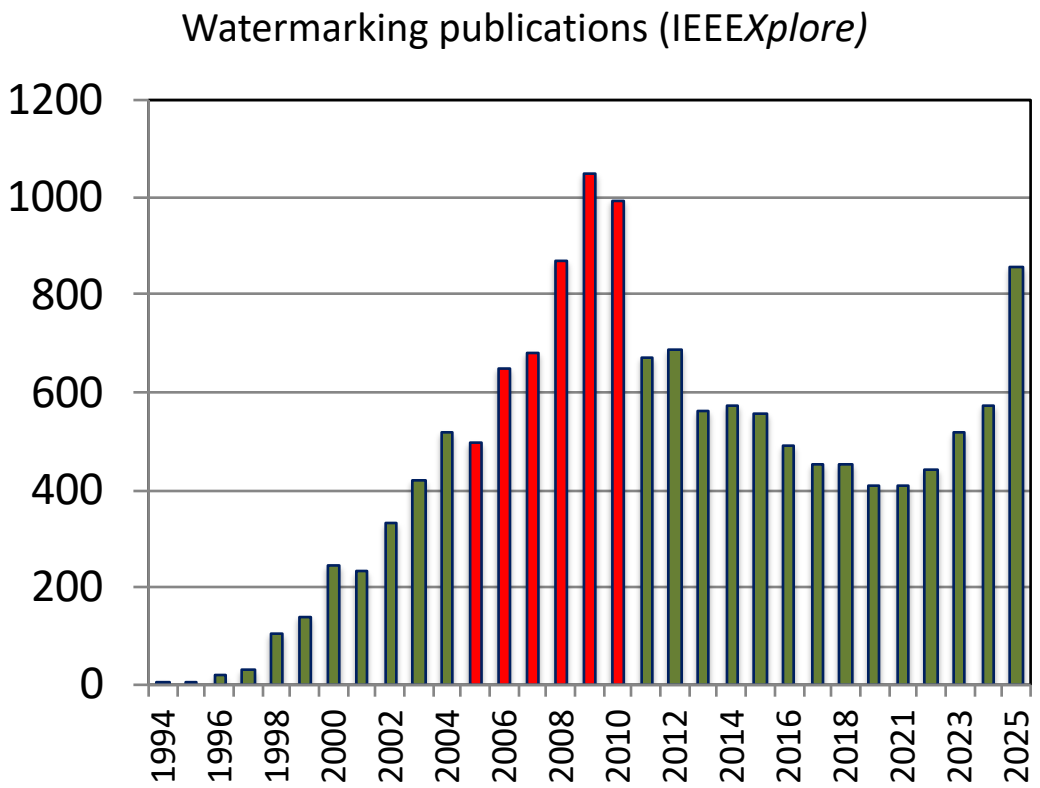
QIM (Quantization Index Modulation) watermarking

- Let I be the to-be-marked image (in some domain)
- We consider different quantizers (colors)
- Every message is associated to a different quantizer (colour)
- Say blue. We look for the blu quantized value closest to I
- We transmit it or a linear combination with I
- Decoding works by looking at the colour of the closest point

- A clear tradeoff among the requirements

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Maturity and open problems (2005-10)



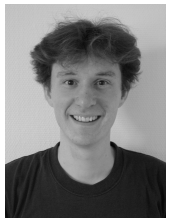


Cox, I., Miller, M., Bloom, J., Fridrich, J., & Kalker, T. (2007). *Digital watermarking and steganography*. Morgan Kaufmann.

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Watermarking security

- Start understanding the difference between robustness and security
- The presence of an enemy who wants to disrupt the watermark was considered
- Security defined in terms of **Mutual Information between observed data and watermark key** (perfect security: $I(K;Y) = 0$)



- F. Cayre, C. Fontaine, T. Furon. "Watermarking security: Theory and practice", IEEE Trans. Signal Processing, vol. 53, no.10 (2005)



- P. Comesana, L. Pérez-Freire, F. Pérez-González, "Fundamentals of data hiding security and their application to spread-spectrum analysis." Information Hiding Workshop, 2005.

- Very powerful attacks were developed: Sensitivity (oracle) attacks, Collusion attacks

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Open problems

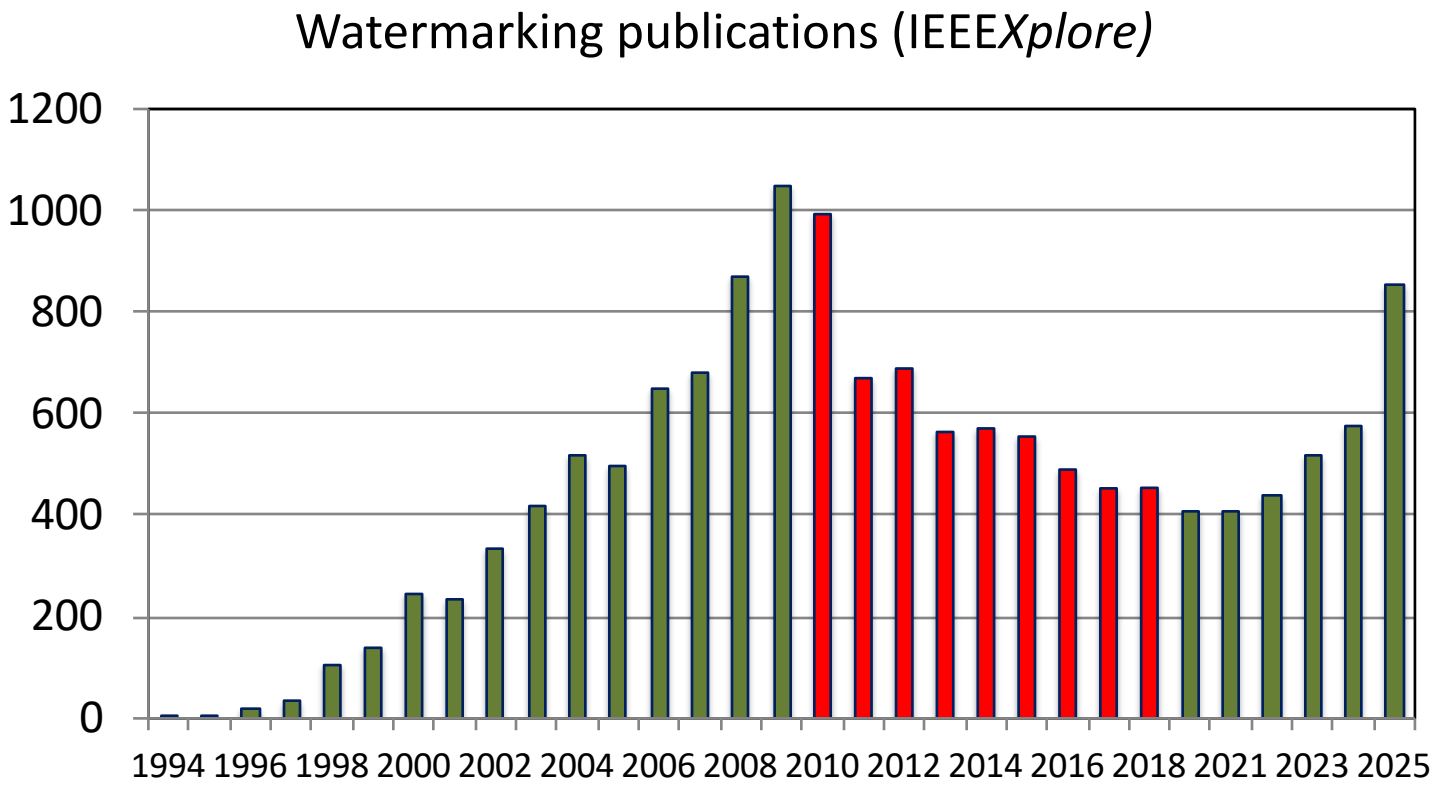
After more than a decade of research some some important problems still had no easy solution

- Robustness against **desynchronization attacks** (geometric attacks for images) was not easy to achieve

- Lack of good and easy to use **perceptual models**

- **Security** (whatever it means) was very difficult to reach
  - Sensitivity attacks –> doesn't they resemble adversarial examples !!!
  - Collusion attacks
  - **Role of key and key secrecy was widely recognized**

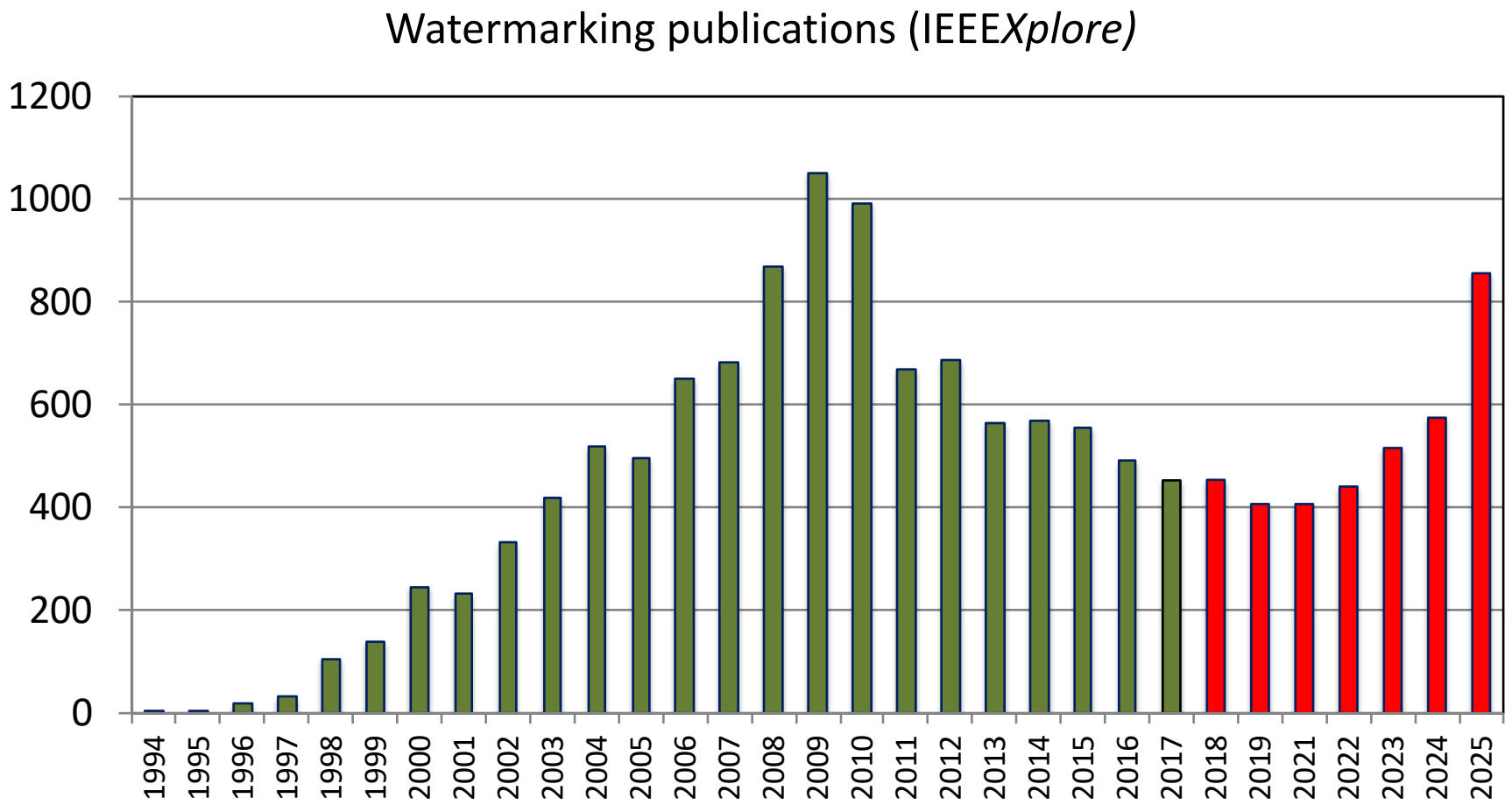
Declining interest (….a dormant phase)



- Focus on less relevant problems/minor improvements

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

The AI age



Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

From DNN Watermarking to Watermarking of GenAI content

- Once again, copyright protection was the motivation for the renaissance of watermarking - watermarking **of DNNs**

- A shift towards **DNN-based** watermarking (for GenAI detection and attribution)

Generation-time

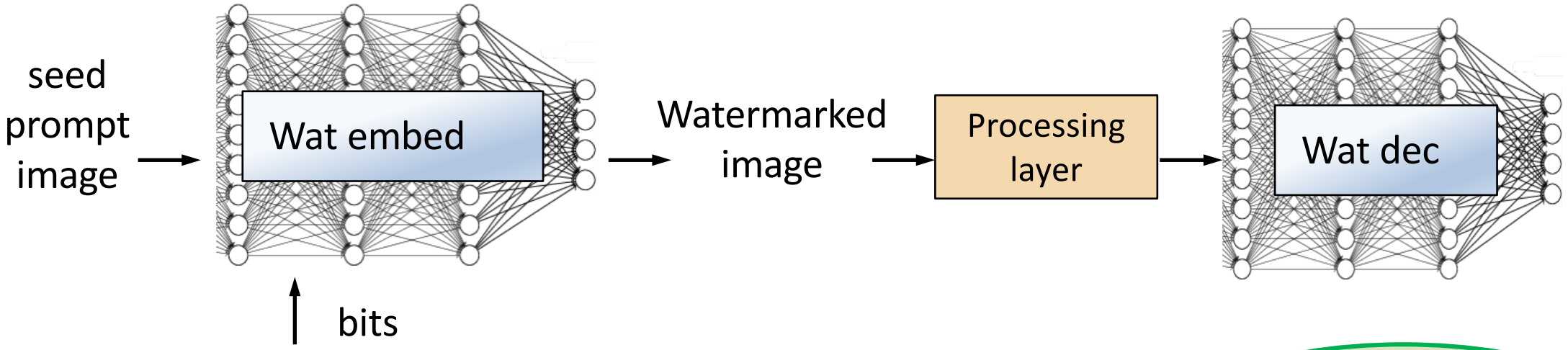
- Stable Signature
- Tree-Ring Watermark

Post-generation - equivalent to classical image watermarking

- StegaStamp, HiDDeN, CIN, MBRS, TrustMark, VINE

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

The AI route to image watermarking



Trained with ad-hoc loss: $\mathcal{L}_{tot} = \mathcal{L}_q + \lambda \mathcal{L}_{wat}$

It replaces SS and QIM watermarking

Data-driven embedding and decoding strategies

Robustness is obtained by introducing a processing layer before decoding

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Some questions

- How do AI-based methods compare to traditional schemes
  - *In terms of payload, robustness, quality …*
- Does the communication with side channel information make any sense with AI-based watermarking?
- Can AI be used to solve some of the long-standing problems in watermarking?
  - *Watermark desynchronization* – can the noise layer help?
  - *Perceptual models* – perceptual features learned directly from data
  - *Security* - what about it?

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Back to the triangle trade-off

Spread-spectrum watermarking: $y_n = x_n + \gamma b w_n$

γ directly trades-off robustness and image quality, length of spreading sequence trades-off robustness and payload (and indirectly quality). ***Statistical guarantees of accuracy***

QIM watermarking: $y_n = Q_{0/1}(x_n, \Delta)$

Δ trade-offs robustness and image quality, number of quantizers and length of sequence trade-offs robustness and payload (and indirectly quality). ***Provably correct decoding****.*

AI watermarking: $\mathcal{L}_{tot} = \mathcal{L}_q + \lambda \mathcal{L}_{wat}$

λ trade-offs (empirically) accuracy and quality, payload has an indirect effect on both quality and accuracy, robustness controlled by noise layer. ***Empirical guarantees of accuracy. No security***

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Payload and size limitations

- The payload is often limited (*Gaussian Shading* can reach a few hundred bits)
  - Most DNN-based watermarking schemes often have **small and fixed input size**
  - Images from AI generators (StableDiff, Flux, Freepik…) have typically very large size (1024x1024 and above!)
  - Some methods use a **downsample-embed-upsample** strategy (TrustMark)- embedding performed on downsampled images to fit the network input
    - reduced space for embedding !!
    - Poor perceptual shaping

The security problem

- Security and robustness are often confused (as it was until 2005)
- Public (or accessible) watermark decoders - e.g., *TrustMark* - are an asset for adversaries
  - Complete access to decoders -> *white-box adversarial attacks* are possible, or *overwriting attacks*
  - API (black-box) access to decoders -> *query-based attacks* can be launched – sensitivity attacks !

Attacking a watermark detector is a very easy job

- Case of CIN decoder - Algorithm for multi-label targeted attack [+]



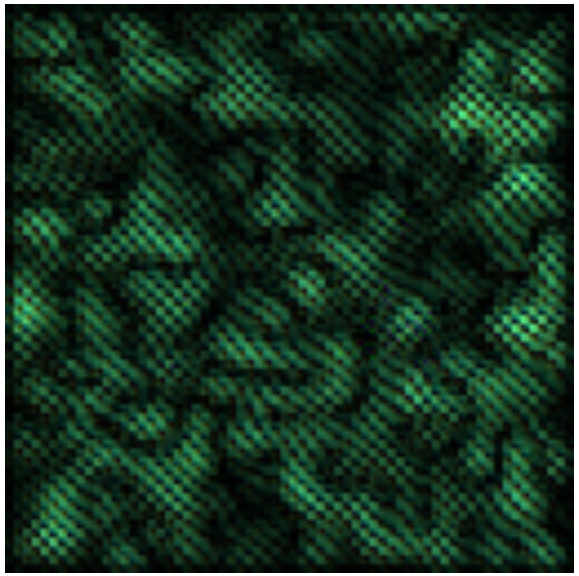
Watermarked

w=01001011…01010
(100 bit)



Attacked

w^t =10110100… (= $\bar{w}$, all bit flipped !!) bASR = 100%
PSNR=37.7 SSIM =0.978



Difference (scaled)

[*] Tondi, B., Guo, W., Pancino, N., & Barni, M. (2025). JMA: a General Algorithm to Craft Nearly Optimal Targeted Adversarial Examples. IEEE Transactions on Information Forensics and Security, Vol.20, Sept 25.

Attacking a watermark detector is a very easy job

- Case of MBRS decoder - Algorithm for multi-label targeted attack [+]



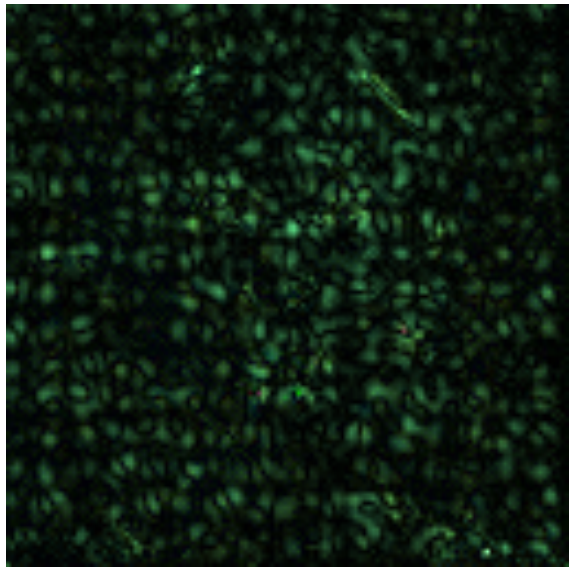
Watermarked

w=10101001….001101
(100 bit)



Attacked

w^t =01010110…..(= $\bar{w}$, all bit flipped !!) bASR = 100%
PSNR=58.9 SSIM=0.999



Difference (scaled)

[*] Tondi, B., Guo, W., Pancino, N., & Barni, M. (2025). JMA: a General Algorithm to Craft Nearly Optimal Targeted Adversarial Examples. IEEE Transactions on Information Forensics and Security, Vol.20, Sept 25.

Secure watermarking

Possible strategies offered by AI (nonlinear mappings, content-adaptive embedding, adversarial training …) may help but DO NOT SOLVE *Adversarial examples are unavoidable* !

- The necessity of a secret key: a lesson from history
- No public (or freely accessible) decoders
- Develop AI watermarking methods that work with a key (?)
- The definition of suitable threat models (often neglected) is important

Wrapping up

- Most AI-based watermarking methods are quite naïf and empirical
- No attempt to link to established media watermarking theory
- No attempt to derive general results on the achievable tradeoff between payload, robustness, quality
- No rigorous security analysis

Takeaway: learning from the past

- Security is not robustness
- Do not confuse 0-bit and multi-bit watermarking
- Forget embedding logos: bits are bits
- Don’t use ad-hoc performance metrics, rather Bit Error Rate, False Positive and False Negative error probabilities (and AUC)
- Do not confuse steganography and watermarking

  The StegaStamp example: *nomina sunt consequentia rebus*
- Payload is not capacity

Thank you
for your attention

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND