JIVP webinar series
2 November 2023

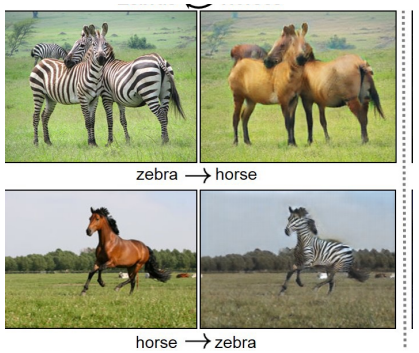
GAN fingerprint for active detection/attribution of fake media

Mauro Barni
University of Siena

Outline

- A little bit of history
- The AI media-revolution
- Limits of passive media forensics
- Active DNN fingerprinting
- DNN watermarking in a nutshell
- GAN watermarking: early and new solutions
- Looking ahead

A history of fakes



90's

2014

1826

CHAT GPT

The analogue ERA

The Photoshop ERA

The AI ERA

When alarm started raising

In the end of the 90's, the diffusion of easy-to-use image editing tools raised increasing alarm about the credibility of digital images and the possible use of manipulated images for malevolent purposes

- **Use of digital images in a court of law**
- **Gossip / defamation**
- **Bias the political debate**

The birth of a new discipline

First technical solutions were proposed in the early 2000's

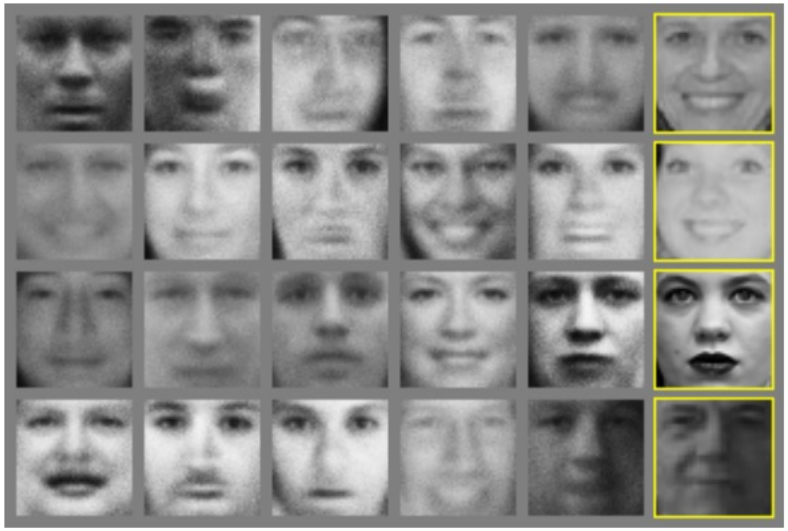
Interestingly They were proposed by researchers working in steganalysis

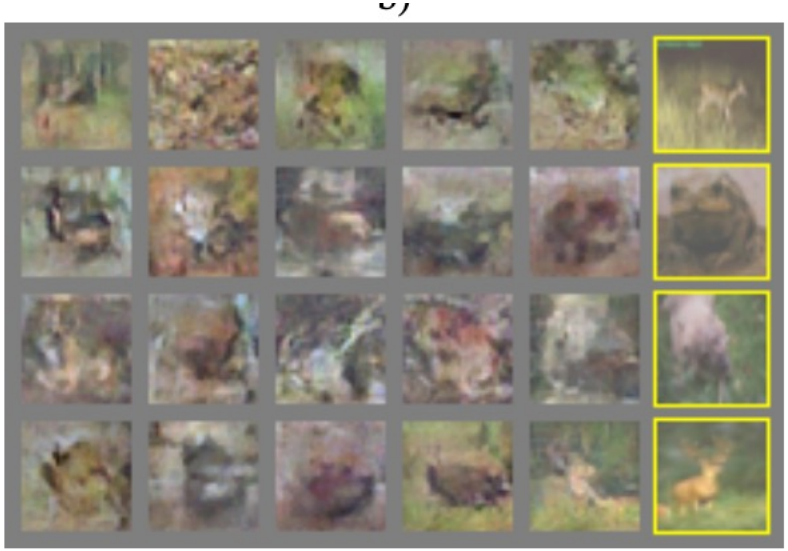
- Popescu, Alin C., and Hany Farid. "Statistical tools for digital forensics." *International workshop on information hiding*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.
- Fridrich, J., Soukal, D., & Lukas, J. (2003, August). Detection of copy-move forgery in digital images. In *Proceedings of digital forensic research workshop* (Vol. 3, No. 2, pp. 652-63).
- **Image (and multimedia) forensics research**
- **REWIND Project, 7PF, EU**

The AI revolution (yet another)

The AI revolution of photo-editing started in 2014*

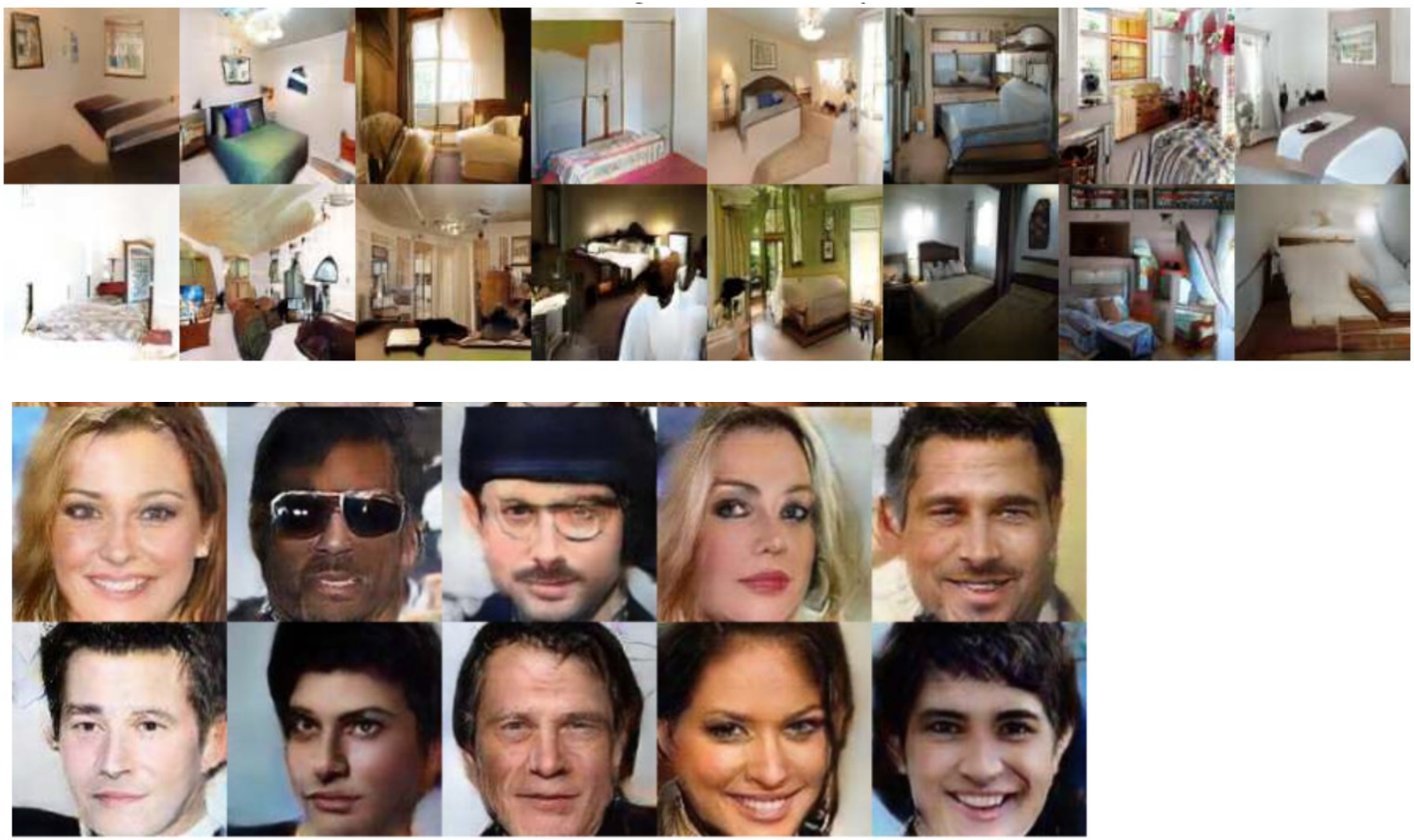
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.





*This paper is receiving more than 1000 citations **per month** (still increasing)

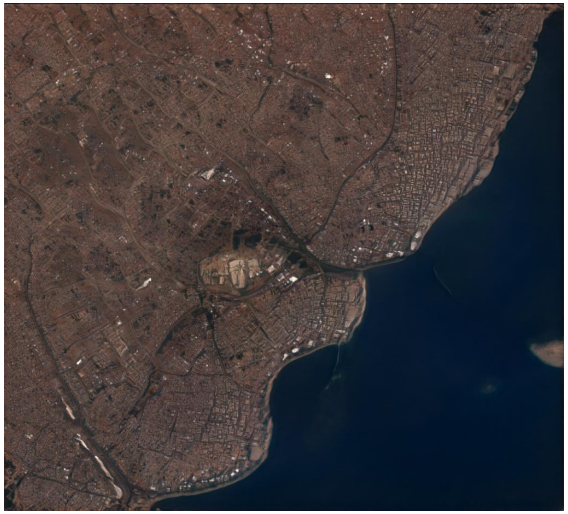
Since then the ball started rolling …



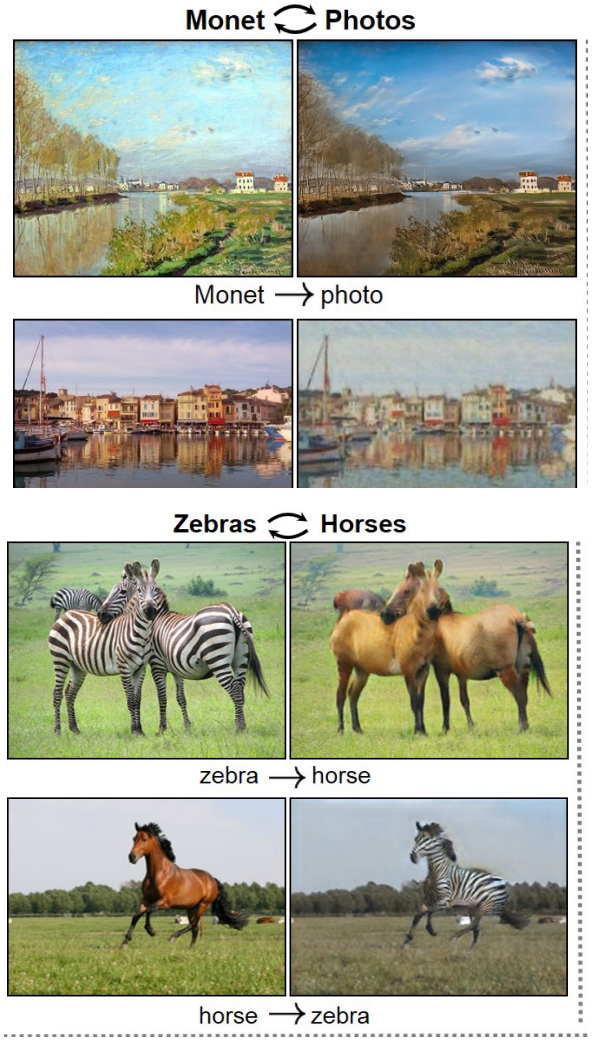
and rolling and rolling …

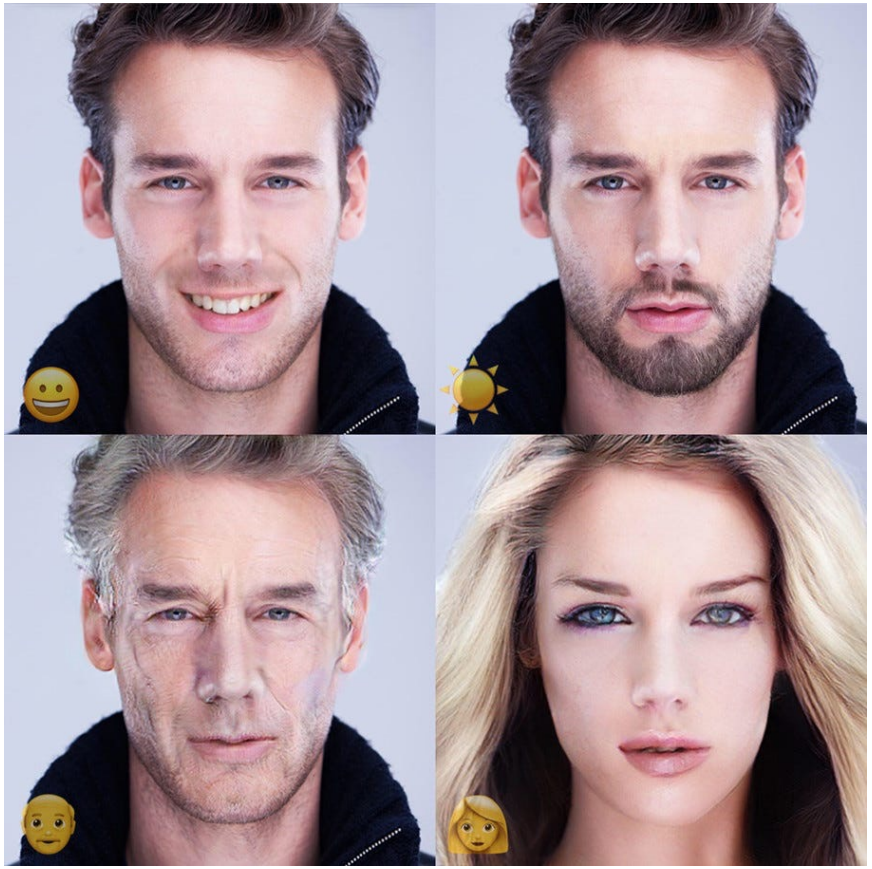
This X does not exist





Im2Im translations





And of course videos

Then chat-GPT, DALL-E …

Two Italian girls eating pizza un a sunny day

A motorbike blocked by a cow on a mountain road

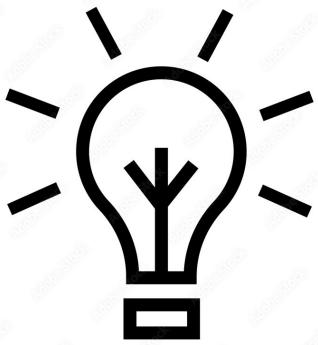
An ubiquitous presence

- Image taken by cell-phone are “doctored” at the origin
- All cell phones have AI-based apps to retouch/manipulate images easily and in a credible way
- AI image editing tools are freely available online
- All photoeditors are equipped with AI tools easier and easier to use
- Video enacting or puppeteering will soon be available on any video conferencing systems
- Companies are alteady working on a video version of DALL-E
- …and who knows what …

Does conventional MMF still make sense?

- The amount of possible modifications, their extent and the progress of image editing AI clearly outgun the capabilities of MMF research
- MMF forensics *IN THE WILD* is more difficult than ever
- Understanding that an image has been “manipulated” by an AI tool may (no longer) be a meaningful info

A paradigm shift



An old idea

Why don't we embed an invisible watermark in all AI images so to ease

- *Origin verification*
- *Trace manipulation history*
- *Detection of abuses*

Unfeasible solution: how can we enforce the watermarking of all images generated or edited by means of AI ?

A paradigm shift

Rather than engaging a hopeless race of arms with generative AI companies, **team up with them** to ease the ethical use of digital images and identify abuses

A new idea

Watermark AI generative models so that all the images they produce contain a watermark

Possibly feasible: only a handful of companies are capable to train *from scratch* a new AI generative model

DNN watermarking in nutshell

DNN watermarking has been proposed as a way to protect the IPR of DNN models [1,2]

It indissolubly embeds within a DNN model a piece of information to be used later for a given purpose

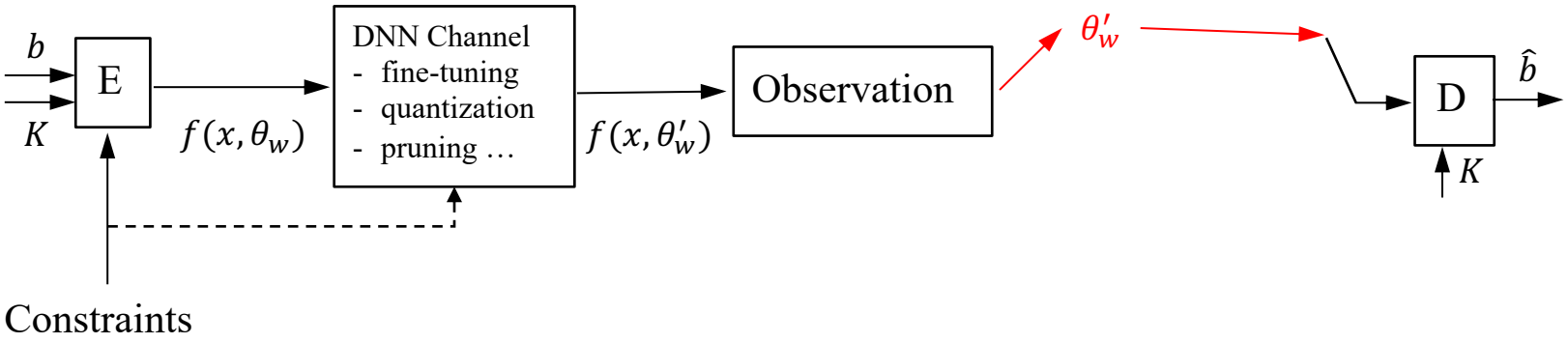
The watermark should resist moderate to strong model modifications, like pruning, compression, fine tuning and even transfer learning

[1] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, "Embedding watermarks into deep neural networks," in Proc. ICMR'17, 2017

[2] Li, Y., Wang, H., & Barni, M. (2021). A survey of deep neural network watermarking techniques. *Neurocomputing*, *461*, 171-193.

DNN watermarking = function watermarking

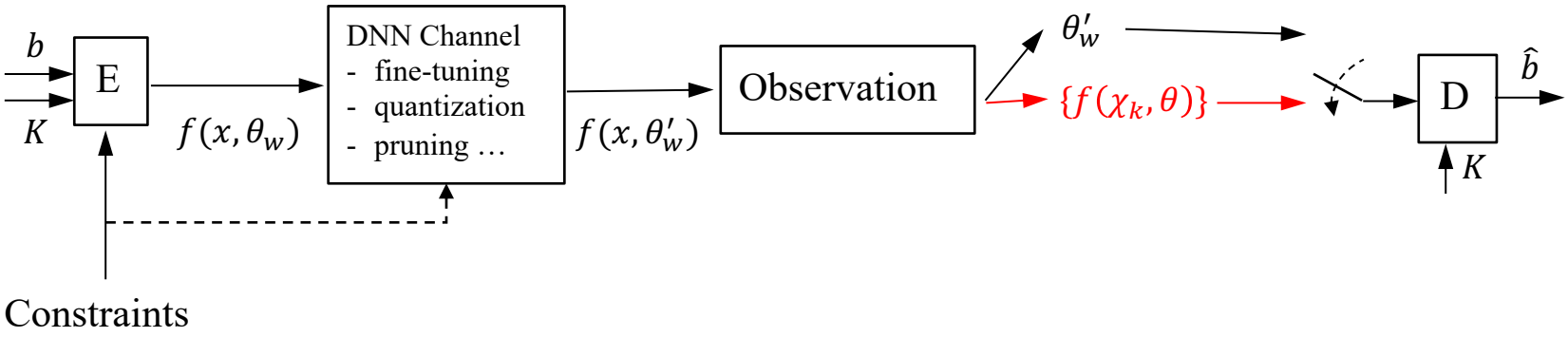
A digital communication problem (multi-bit watermarking)



White-box watermarking

DNN watermarking = function watermarking

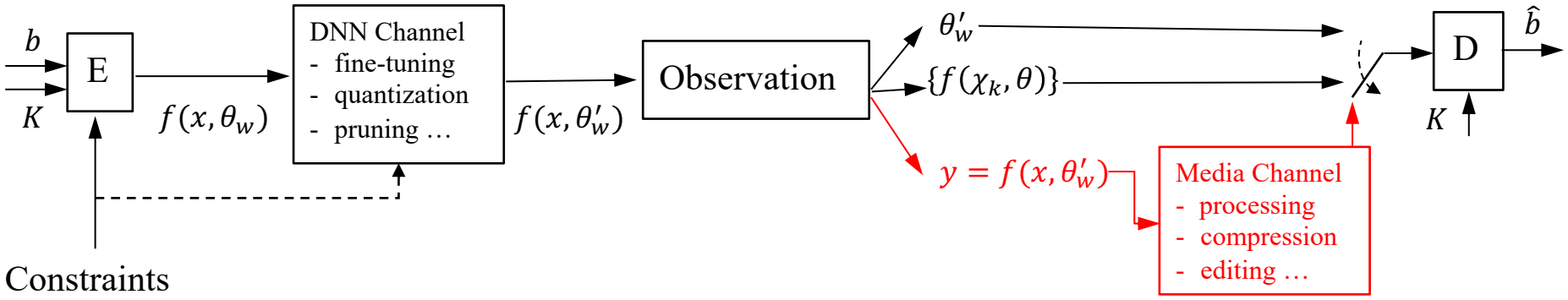
A digital communication problem (multi-bit watermarking)



Black-box watermarking

DNN watermarking = function watermarking

A digital communication problem (multi-bit watermarking)

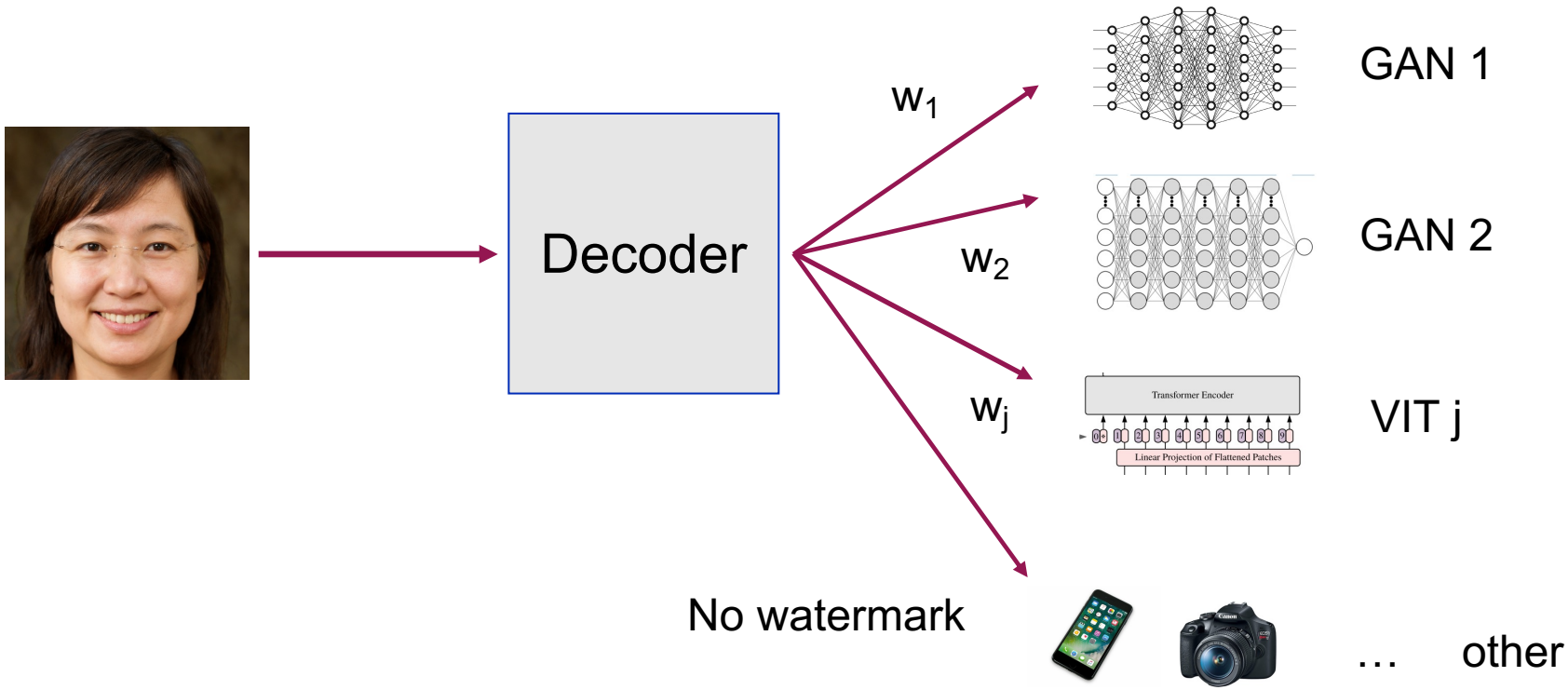


Box-free watermarking

Joint DNN and media watermarking

Kill two birds with one stone

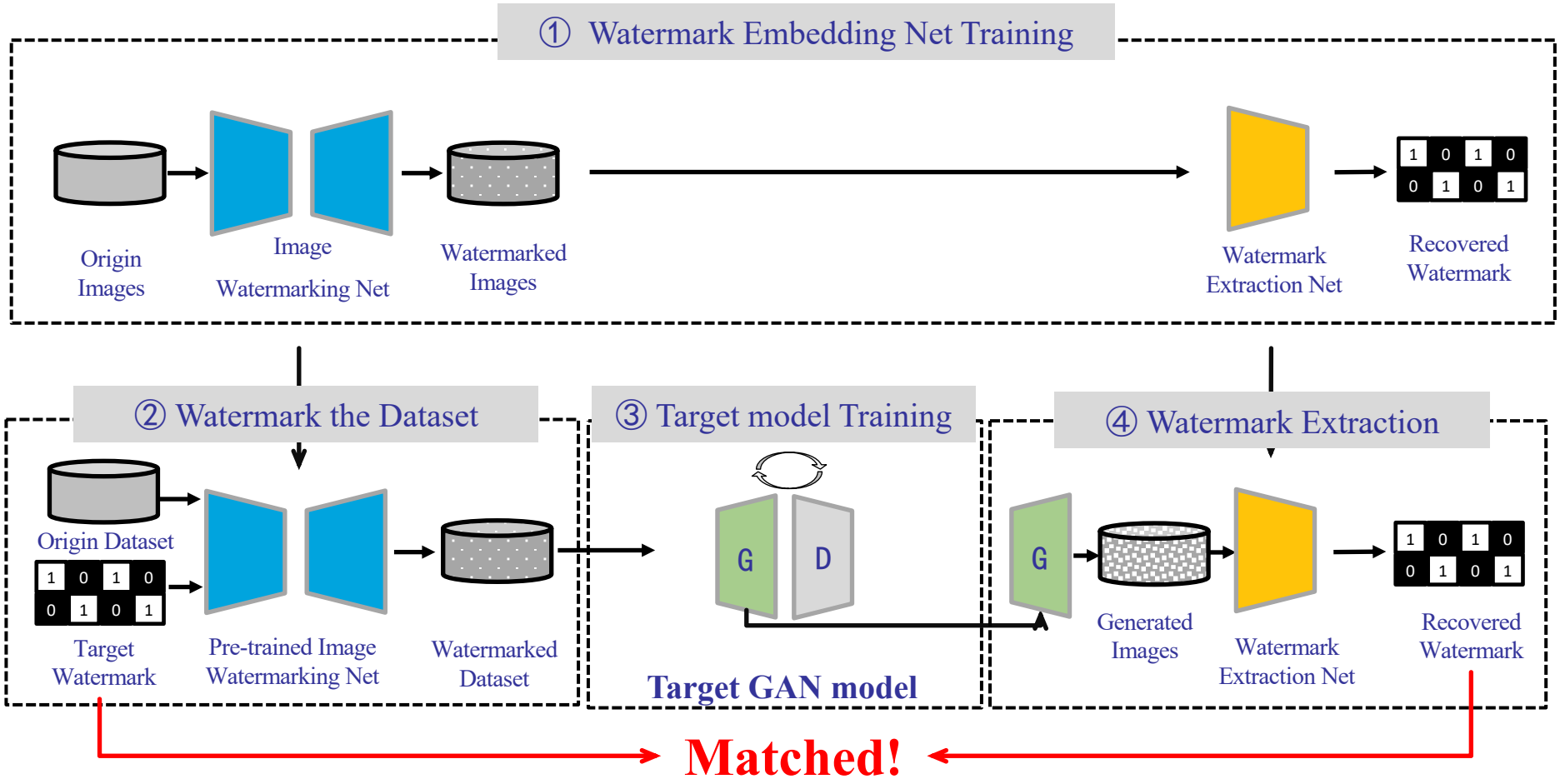
Box-free watermarking can be used to detect and trace synthetic contents to the model which generated them



Challenges

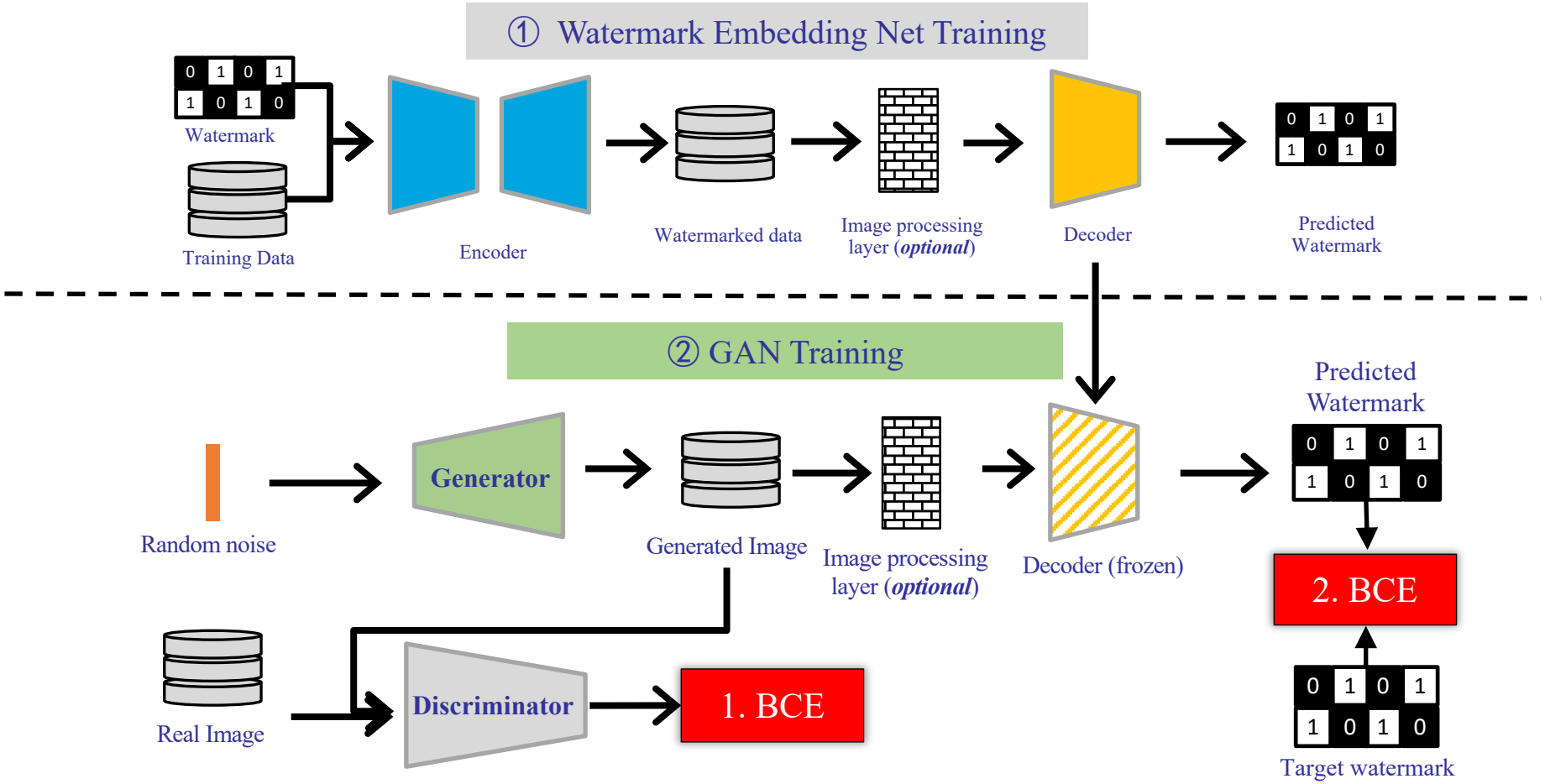
- Development of a brand new theory of function watermarking
- Model-level robustness
- Image-level robustness
- Security model
  - Embedded information
  - Who does what?
  - Keyword management
- Security against intentional attacks

Early attempts: the imitation game [1]



[1] N. Yu, et. al, Artificial fingerprinting for generative models: Rooting deepfake attribution in training data, ICCV, 2021

Supervised watermarking [2]



[2] Fei, J., Xia, Z., Tondi, B., & Barni, M. (2022, December). Supervised gan watermarking for intellectual property protection. In *2022 IEEE International Workshop on Information Forensics and Security* (WIFS)

Selected results [2]

StyleGAN2 | FFHQ | 256

Clean
FID 5.28

Marked
FID: 6.16
Bits 100
Acc 99.75

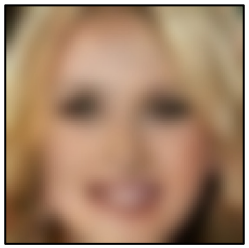
[2] Fei, J., Xia, Z., Tondi, B., & Barni, M. (2022, December). Supervised gan watermarking for intellectual property protection. In *2022 IEEE International Workshop on Information Forensics and Security* (WIFS)

Selected results: image processing

Original
Bit acc
100%

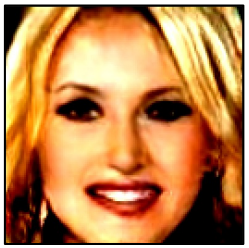
JPEG QF=30
Bit Acc 65%
PSNR 32.14

Noise Std=0.1
Bit Acc 77%
PSNR 20.05

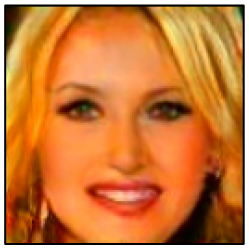


Blurring k=3
Bit Acc 92%
PSNR 25.09

Brightness ×2.5
Bit acc 78%
PSNR 14.97



Contrast ×2.5
Bit acc 84%
PSNR 15.12



Saturation ×2.5
Bit acc 98%
PSNR 14.96

Selected results: model pruning

5%
Acc = **99.95**%

10%
Acc = **99.92**%

15%
Acc = **98.93**%

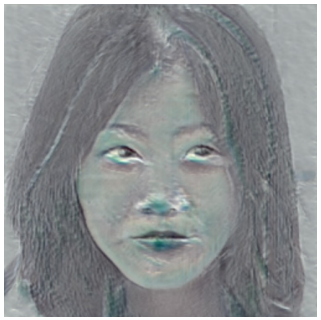
20%
Acc = **95.87**%

25%
Acc = **83.86** %

30%
Acc = **67.90** %

35%
Acc = **65.83** %

40%
Acc = **61.15**%



50%
Acc = **57.09**%

60%
Acc = **52.35**%

Selected results: model quantization

Precision 10^{-5}
Acc = **99.90**%

Precision 10^{-4}
Acc = **99.21**%

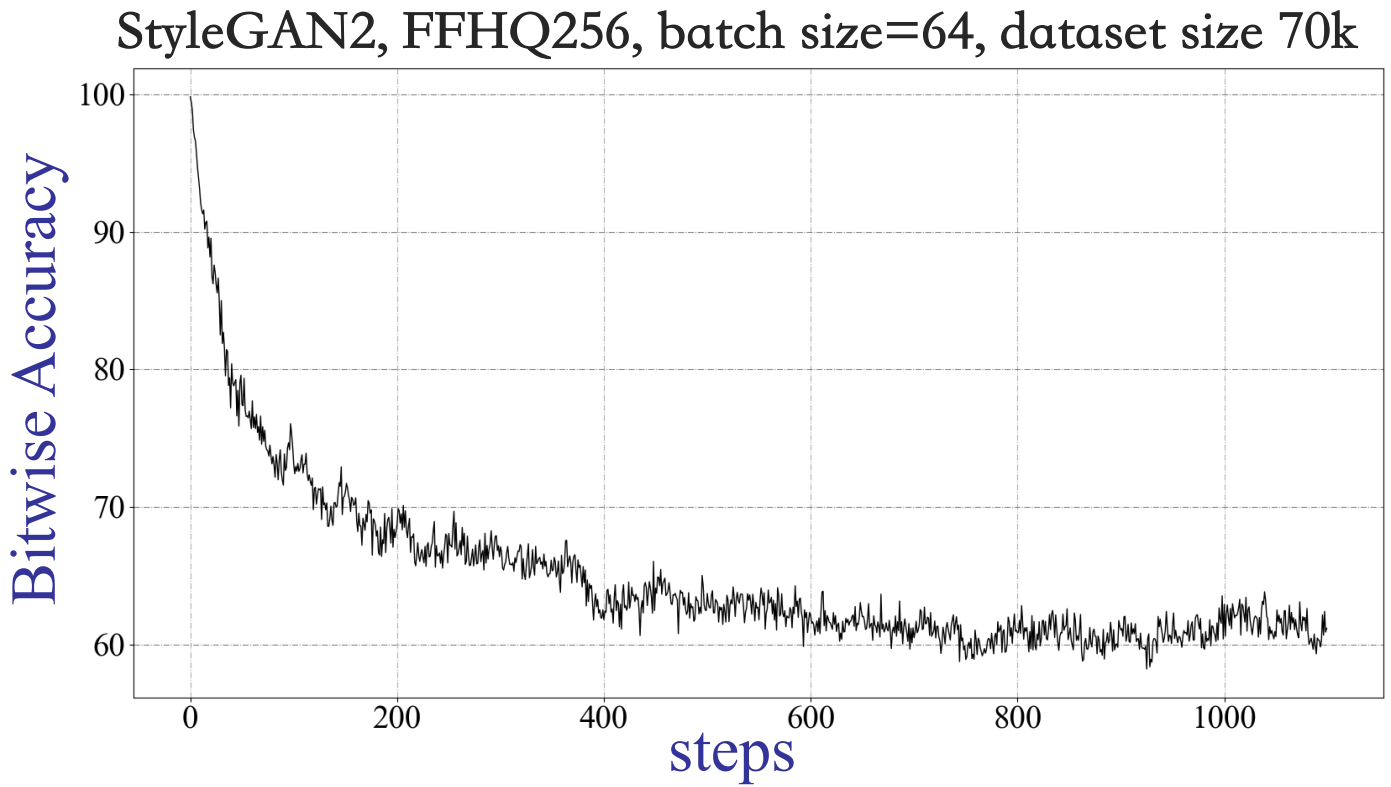
Precision 10^{-3}
Acc = **98.43**%

Precision 10^{-2}
Acc = **85.61**%

Precision 10^{-1}
Acc = **49.7**%

Precision 10^{-0}
Acc = **53.52**%

Selected results: fine tuning



Finetuning attack without WM loss (1 epoch ~ 1093 steps)

Other tasks: super-resolution

Clean

input Ground truth generated

input Ground truth generated

Marked

input Ground truth generated

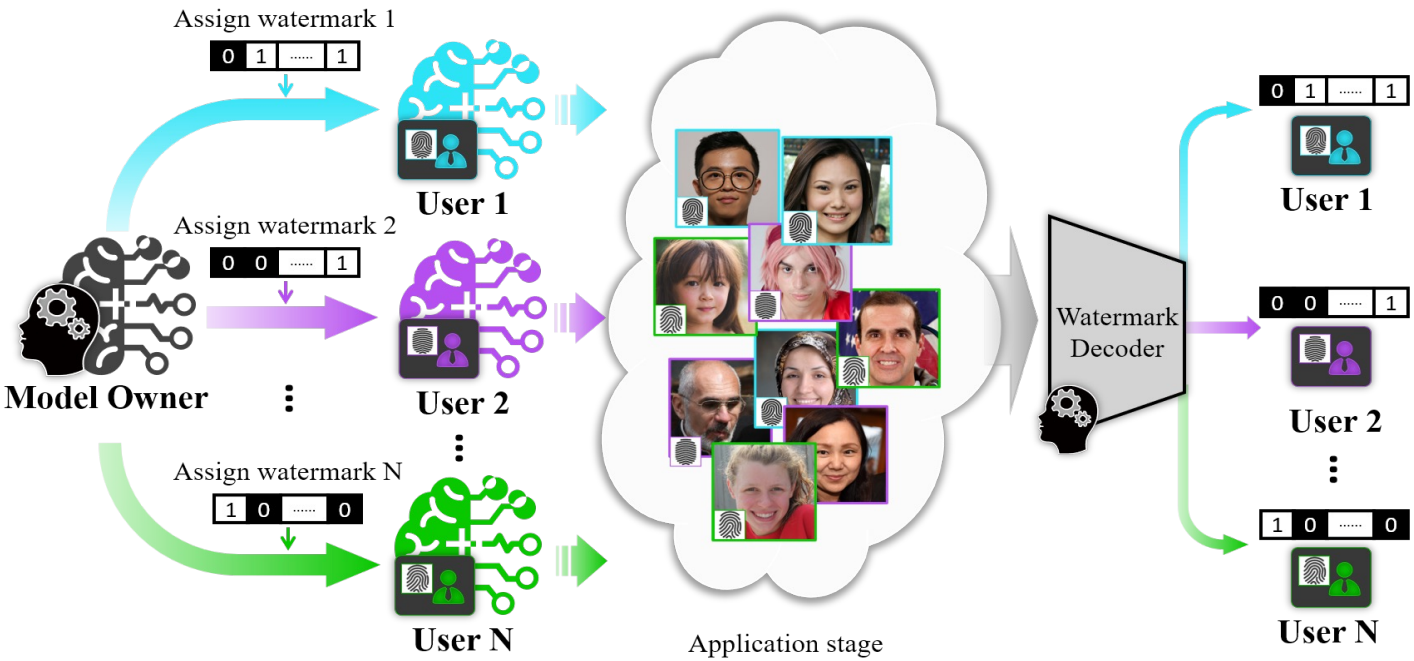
input Ground truth generated

PSNR 23.87 → 21.94 **SSIM** 0.7100 → 0.6206

Retraining free fingerprinting

For some applications it may be useful to embed different watermarks in different version of the *same* DNN model

With most methods this requires a heavy retraining



Solution [3,4]

Introduce a personalized layer in the generator whose parameters are responsible for watermark embedding

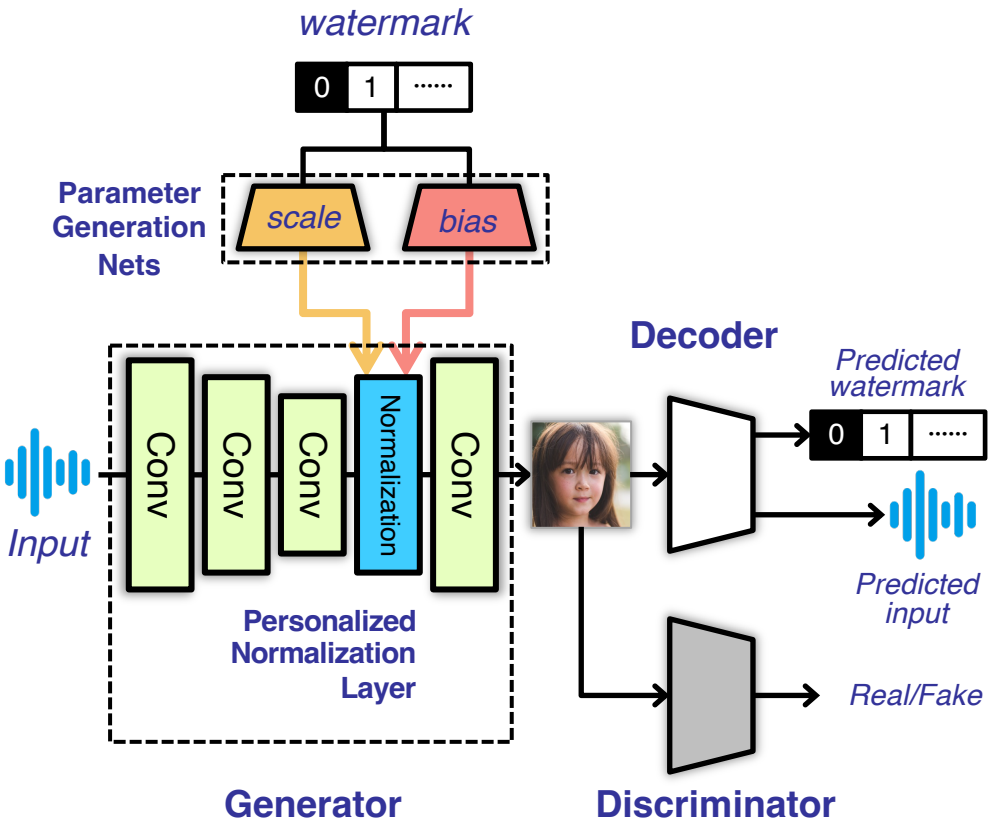
The parameters are generated (feedforward) by a separate parameter-generation network for each different watermark

The param gen network derives the parameters resulting in the embedding of the desired watermark without retraining

[3] J. Zhang, D. Chen, J. Liao, W. Zhang, G. Hua, and N. Yu, "Passport- aware normalization for deep model protection," Advances in Neural Information Processing Systems, vol. 33, pp. 22 619–22 628, 2020.

[4] J. Fei, Z. Xia, B. Tondi, M. Barni, Robust retraining-free GAN fingerprinting via Personalized Normalization, IEEE WIFS 2023

For instance [4]



Losses

$$\mathcal{L}_{wat} = BCE(\text{watermark prediction})$$
$$\mathcal{L}_{adv} = BCE(\textit{real/fake}\ \text{prediction})$$
$$\mathcal{L}_{z} = MSE(z, \text{predicted input})$$
$$\mathcal{L}_{const} = MSE(G_{w_1}(z), G_{w_2}(z))$$

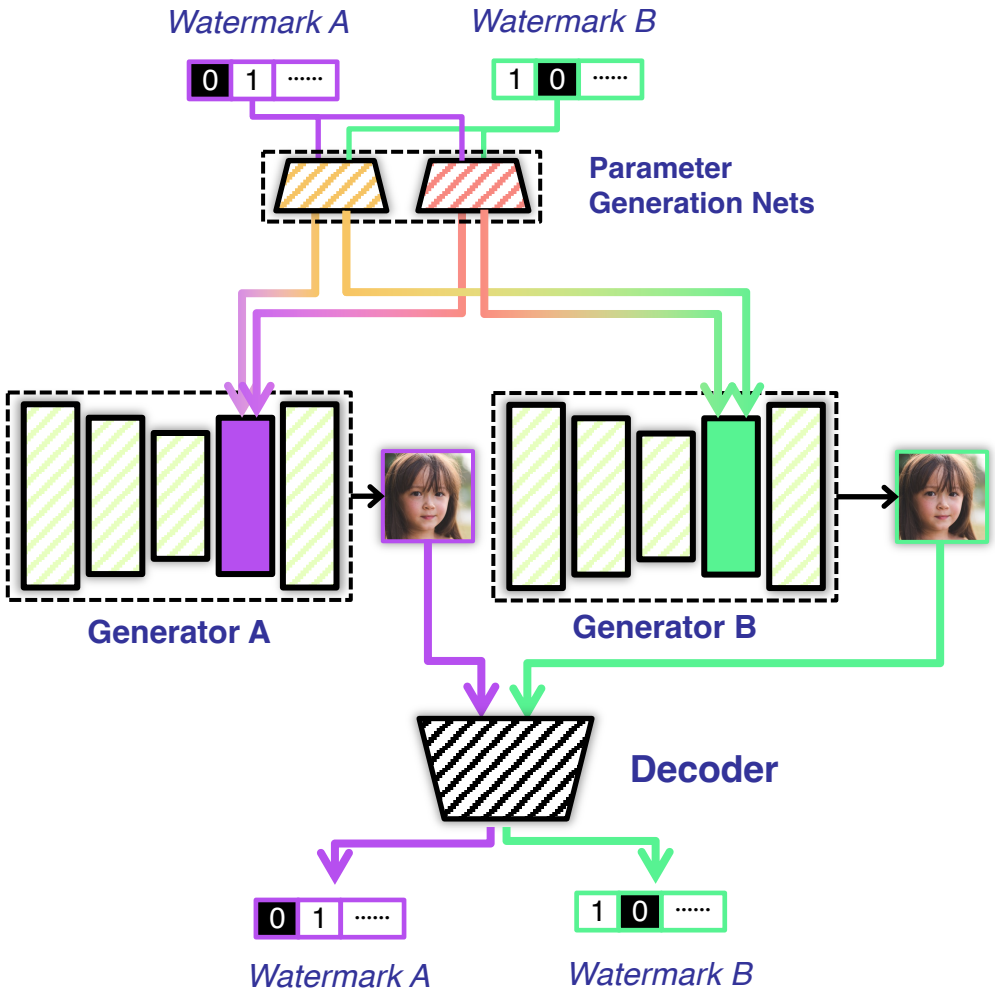
Discriminator: $\mathcal{L}_{adv}$

Decoder: $\mathcal{L}_{wat}$ $\mathcal{L}_{z}$

Generator, $\mathcal{L}_{adv}$ $\mathcal{L}_{const}$
ParamGen: $\mathcal{L}_{z}$ $\mathcal{L}_{wat}$

[4] J. Fei, Z. Xia, B. Tondi, M. Barni, Robust retraining-free GAN fingerprinting via Personalized Normalization, IEEE WIFS 2023

For instance [4]: distribution



To create a model with a given watermark it is only necessary to run the ParamGen networks and generate the weights of the personalized normalization layer

A sense of achievable results

- Boundary Equilibrium GAN (BEGAN)
- Spectral Normalization GAN (SNGAN)
- Progressive Growing GAN (PGGAN)
- Face generation, trained on CelebA dataset

- Penultimate layer for PN
- ParamGen networks: fully connected, ReLu, 128 wm bits

[4] J. Fei, Z. Xia, B. Tondi, M. Barni, Robust retraining-free GAN fingerprinting via Personalized Normalization, IEEE WIFS 2023

Impact of noise loss term

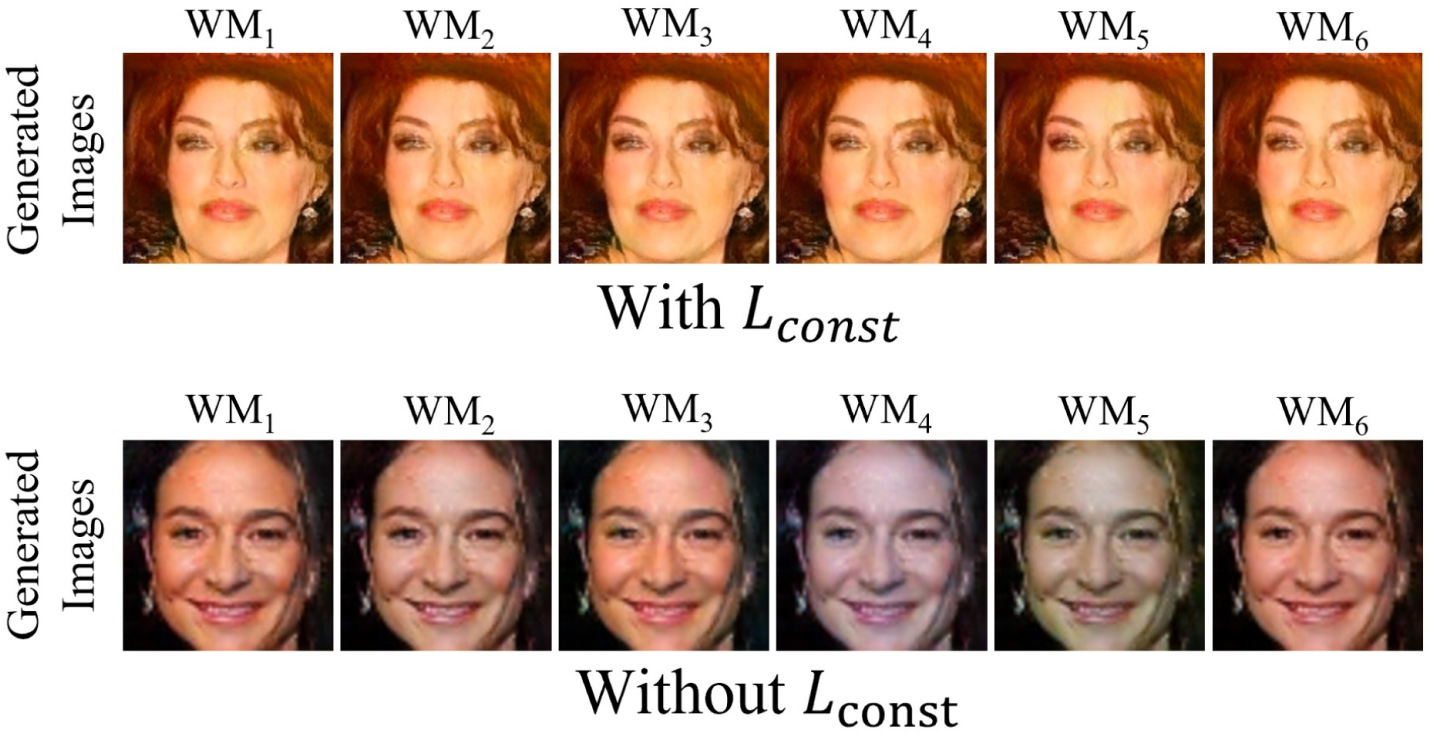
With L_z

Without L_z

The L_z term is essential to force the dependency of the generated images on the input noise

[4] J. Fei, Z. Xia, B. Tondi, M. Barni, Robust retraining-free GAN fingerprinting via Personalized Normalization, IEEE WIFS 2023

Impact of constant-loss term



The L_{const} is crucial to enforce a uniform behaviour of models watermarked with different bits

[4] J. Fei, Z. Xia, B. Tondi, M. Barni, Robust retraining-free GAN fingerprinting via Personalized Normalization, IEEE WIFS 2023

A sense of achievable results [4]

Model	Metric	No WM	Performance
BEGAN	Accuracy	-	100%
	FID	20.89	20.72
	Comput. overhead	-	100ms
SNGAN	Accuracy	-	99.9%
	FID	24.25	24.70
	Comput. overhead	-	100ms
PGGAN	Accuracy	-	99.8%
	FID	27.50	28.02
	Comput. overhead	-	100ms

[4] J. Fei, Z. Xia, B. Tondi, M. Barni, Robust retraining-free GAN fingerprinting via Personalized Normalization, IEEE WIFS 2023

A sense of achievable results [4]

Robustness to model-manipulation

Model	Fine tuning	Pruning (10%)	Pruning (20%)	Quantization 10^{-1}
BEGAN	85%	99%	68%	98%
SNGAN	88%	98%	80%	99%
PGGAN	75%	99%	85%	100%

Robustness to image procesing

Model	JPEG (QF 50)	Blurring (5x5)	Noise addition (std = 0.1)
BEGAN	92%	80%	89%
SNGAN	92%	80%	90%
PGGAN	94%	84%	88%

[4] J. Fei, Z. Xia, B. Tondi, M. Barni, Robust retraining-free GAN fingerprinting via Personalized Normalization, IEEE WIFS 2023

Looking ahead

- Limits of classical Multimedia forensics in the AI era
  - Still valid in specific, narrow, scenarios
  - Difficult (impossible) to cope with in a wide settings
    - Disinformation campaigns
- DNN-based active fingerprinting may provide a solution
  - Challenges to be solved (robustness [A] and security)
  - Do not think it will be a general solution, but can be a valid complement to passive MMF

[A] J Fei, Z Xia, B Tondi, M Barni "Wide Flat Minimum Watermarking for Robust Ownership Verification of GANs" - arXiv preprint arXiv:2310.16919, 2023

Thank you
for your attention