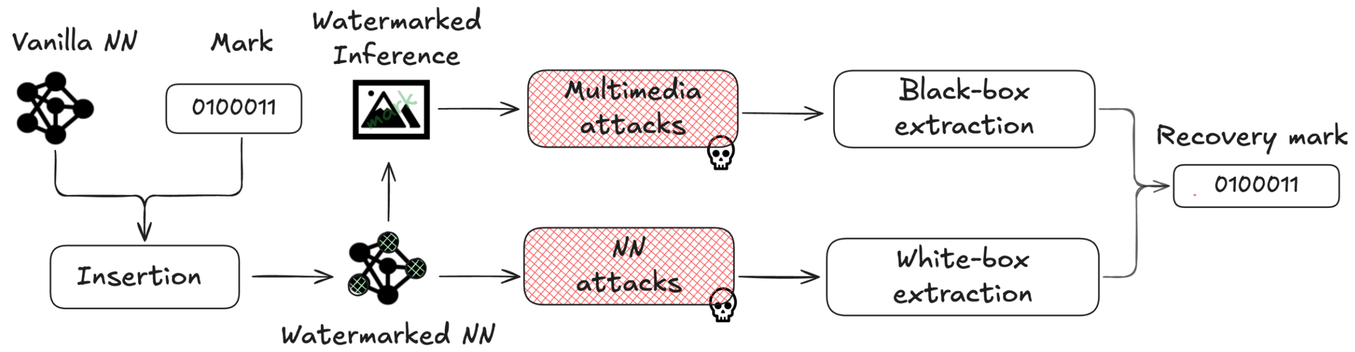
Watermarked image generative models. Case study: GAN models

M. Zoughebi, M. Mitrea, C. De Sousa Trias, H. Chaouchi, SAMOVAR, Telecom SudParis

Context *Rapid growth of AI-generated content*

- **New challenges** in sovereignty, disinformation, and ethics.
- **Mandatory tracking and authentication** of user requests and resulting inferences **(AI Act)**
- Watermarking research totals **100+** studies on **classifiers**, **10+** across **CV tasks**, and **5+** on **GANs**.

Background concepts *NN Watermarking*

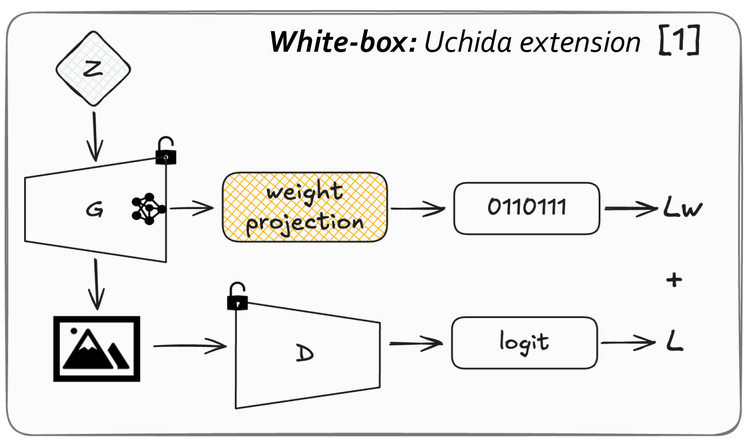


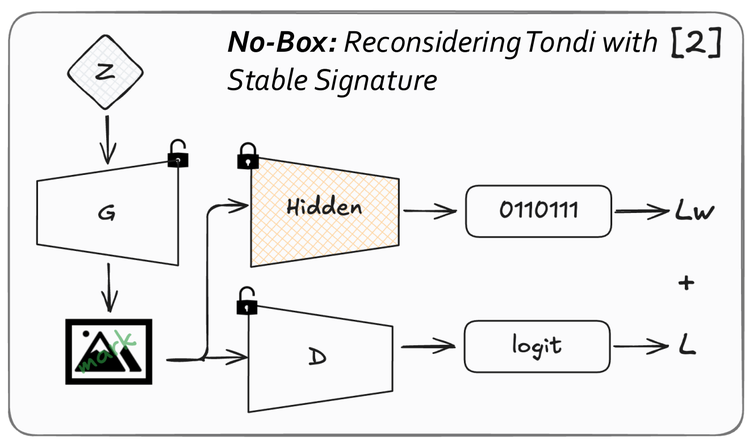
Objectif *End-to-end AI watermarking*

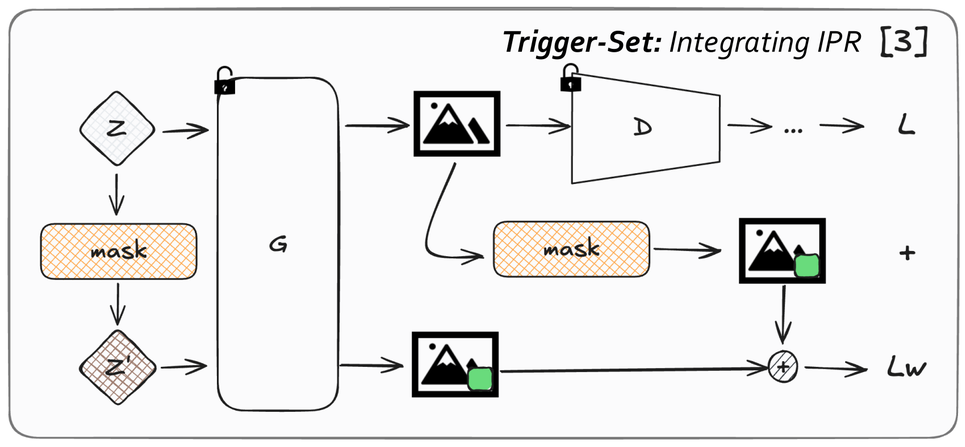
- **For generative models**
- **For images thus generated**

Solutions *Stemmed from the SotA*

- Contributions: **reconsidering**, **extending** and **integrating** existing solutions [1], [2], [3] for **Style GAN2-Ada** to jointly reach **imperceptibility** and **robustness**
  - **New initialization rule for [1]**
  - **New decoder for [2]**
  - **New architecture for [3]**



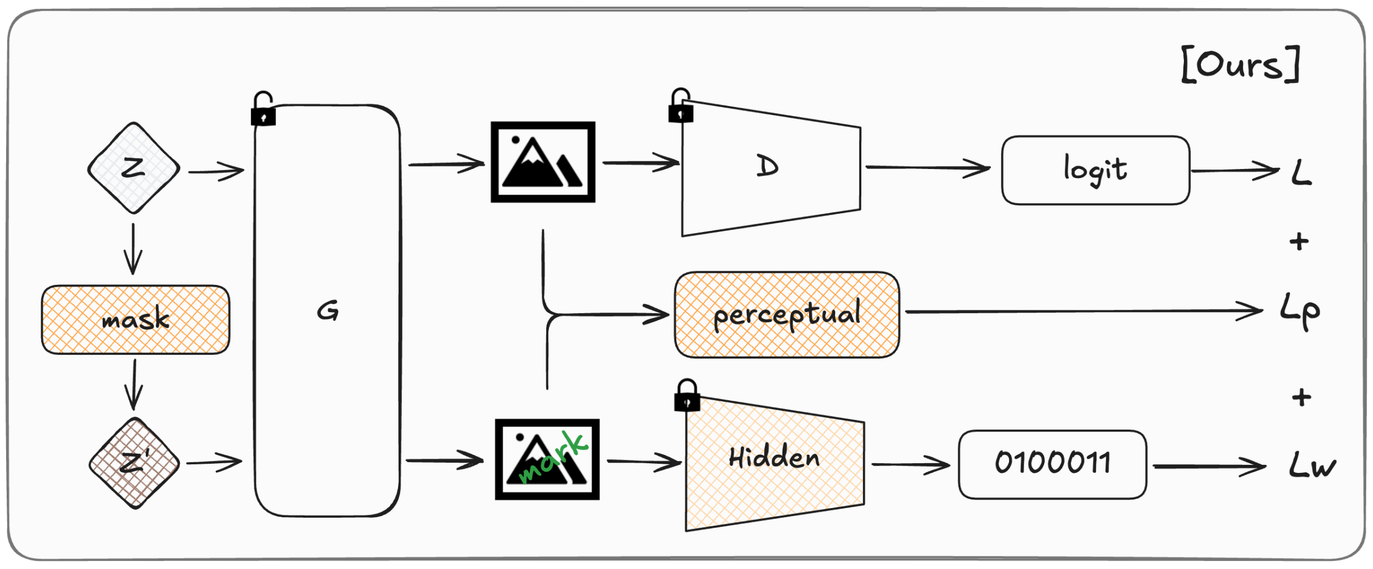




- **Improving insertion for [1]: -0.1 BER**
- **Extending and unifying the experimental results for [2]**
- **Integrating [3] into StyleGAN2 framework**

Advanced solution: *No-box, trigger watermarking*

- ***1st No-box & Trigger synergies***
- ***On/off Watermarking***



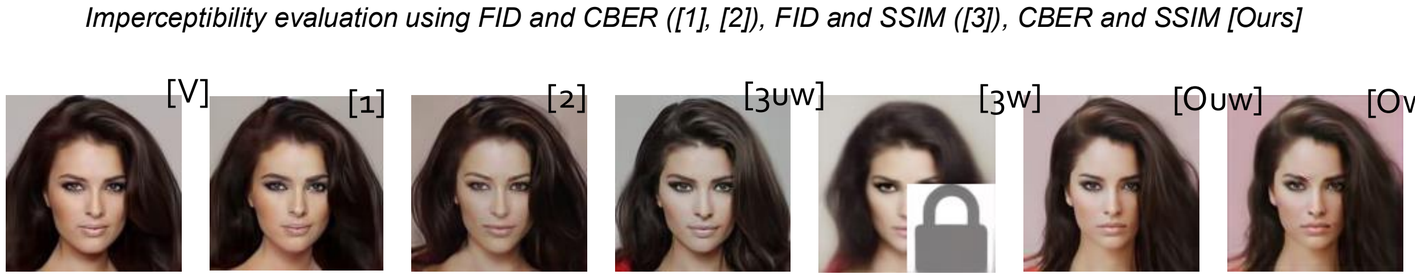
Objectives

- $quality(img_{UW}) = quality(img_V)$
- $perceptual(imgUW, imgW)$
- $insertion(img_W)=1$
- $insertion(img_{UW})=0$
- $quality(img_W) > quality(img_{[2]})$

Imperceptibility evaluation

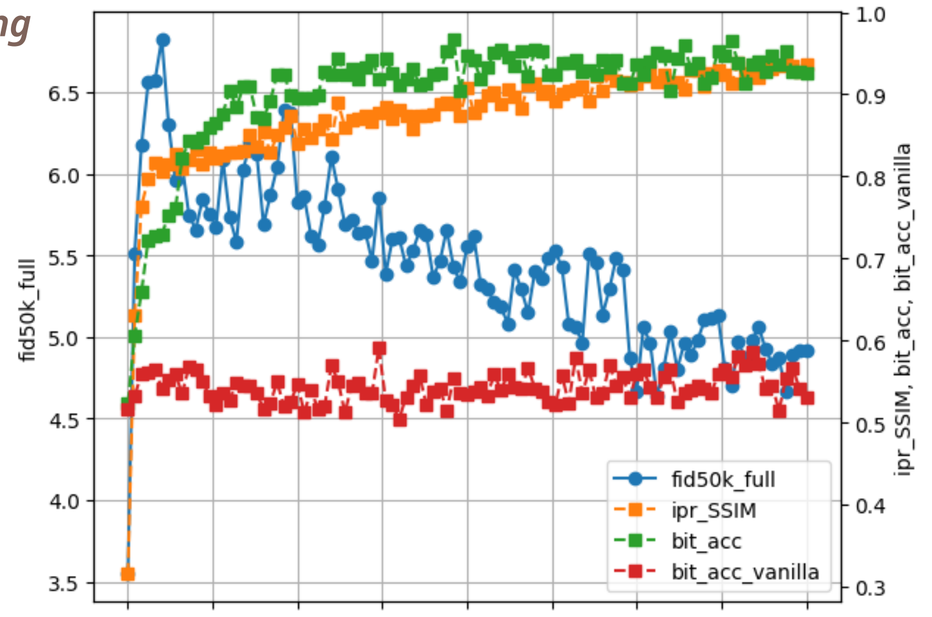
	Methods				
	Vanilla *V*	[1]	[2]	[3]	[Ours] (*UW -- W*)
FID	3.34	3.98	6.83	3.57	4.6 5.4
Insertion	/	1.00	0.97	0.72	0.54/- 0.94/0.92

Imperceptibility evaluation using FID and CBER ([1], [2]), FID and SSIM ([3]), CBER and SSIM [Ours]



Watermarked (W) and unwatermarked (UW) inferences

Training



Training with dual perceptual loss MSE and SSIM on 1600 k-imgs

Robustness evaluation

	Methods			
	[1]	[2]	[3]	[Ours]
Pruning	1.00	> 0.69	> 0.71	>0.79
Noise	1.00	> 0.79	> 0.71	>0.71
Quantization	1.00	> 0.76	> 0.68	>0.62

Robustness evaluation through minimum Insertion-metric range value for FID degradation after NN attacks: ~15.

	Multimedia Attacks					
	None	Crop	JPEG	Resize	Brightness	Contrast
[2]	0.96	0.72	0.65	0.61	0.85	0.96
[Ours]	0.94	0.77	0.71	0.67	0.84	0.93

Robustness of [2],[Ours] against multimedia attacks in terms of CBER

Short-time perspectives

- *Improvement of [Ours] in term of imperceptibility (FID)*
- *Publication of the global study*
- *Integration of [Ours] into the simpleRAN project*
- *Multiple-Key scenario*

This work is supported by BPI France SIMPLERAN Project https://www.simpleran.org/.

mateo.zoughebi@telecom-sudparis.eu
mihai.mitrea@telecom-sudparis.eu
hakima.chaouchi@imt.fr

[1] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, "Embedding Watermarks into Deep Neural Networks," in Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, 2017.

[2] Fei, Jianwei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. 'Supervised GAN Watermarking for Intellectual Property Protection'. 2022 IEEE International Workshop on Information Forensics and Security (WIFS), 2022.

[3] D. S. Ong, C. S. Chan, K. W. Ng, L. Fan, and Q. Yang, "Protecting Intellectual Property of Generative Adversarial Networks from Ambiguity Attack," 2021.