



sAlfer Lab

Joint lab on Safety and Security of AI



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

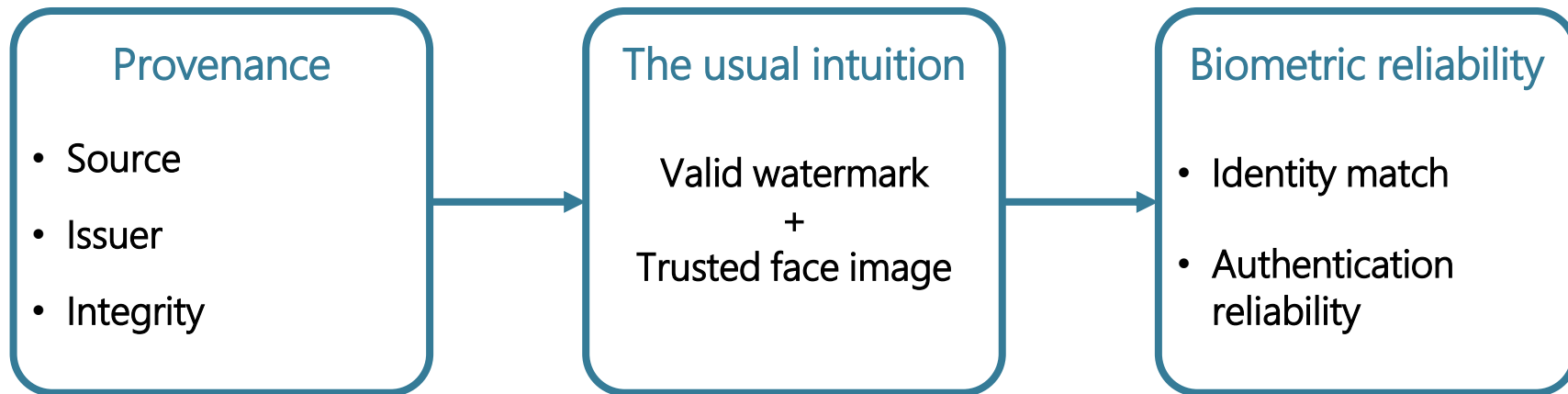
A Valid Watermark Is Not a Valid Identity: Spatial-Spectral Adversarial Face Evasion with Watermark Preservation

Jefferson David Rodriguez Chivata, Davide Ghiani, Simone Maurizio La Cava,
Marco Micheletto, Giulia Orrù, Gian Luca Marcialis

University of Cagliari (Italy)

When Biometrics Break (WBB) - IJCB 2026

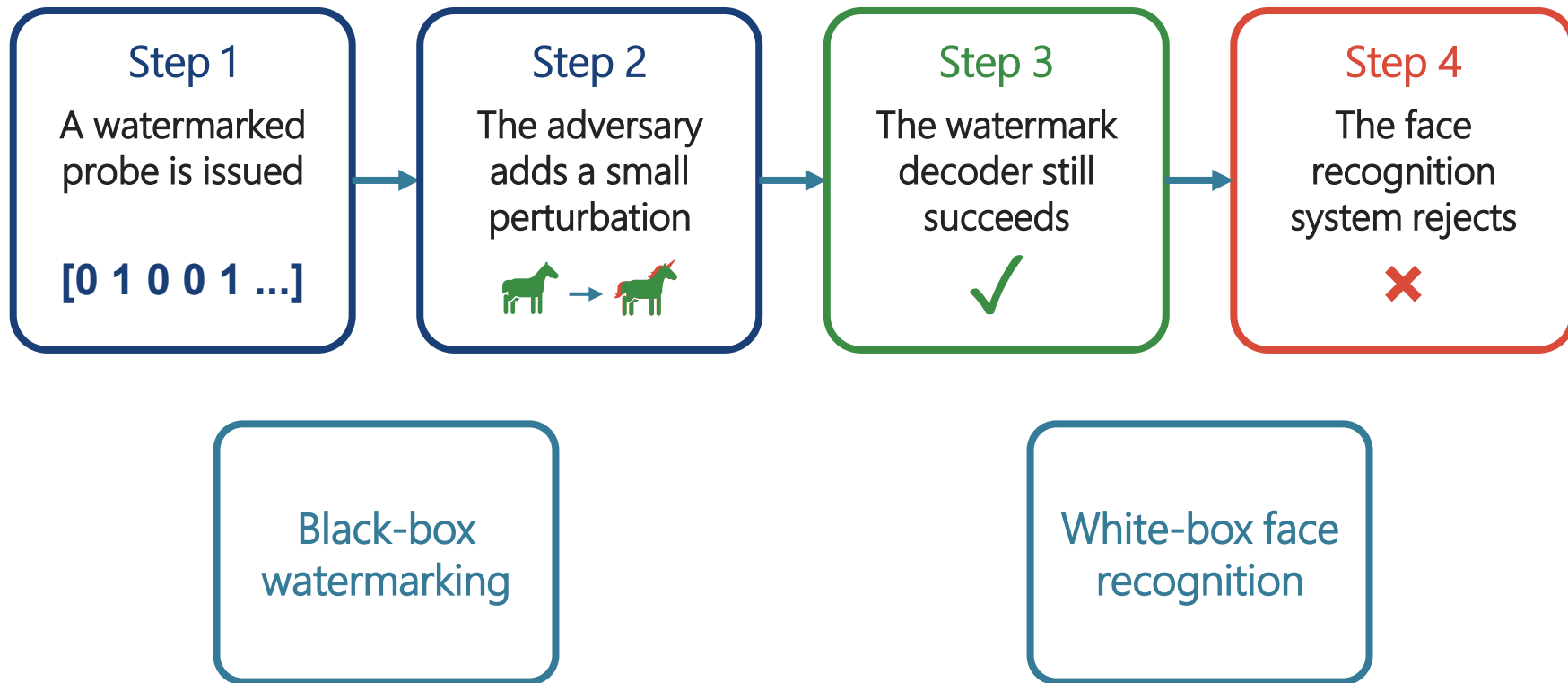
Motivation: Biometric documents require two independent guarantees



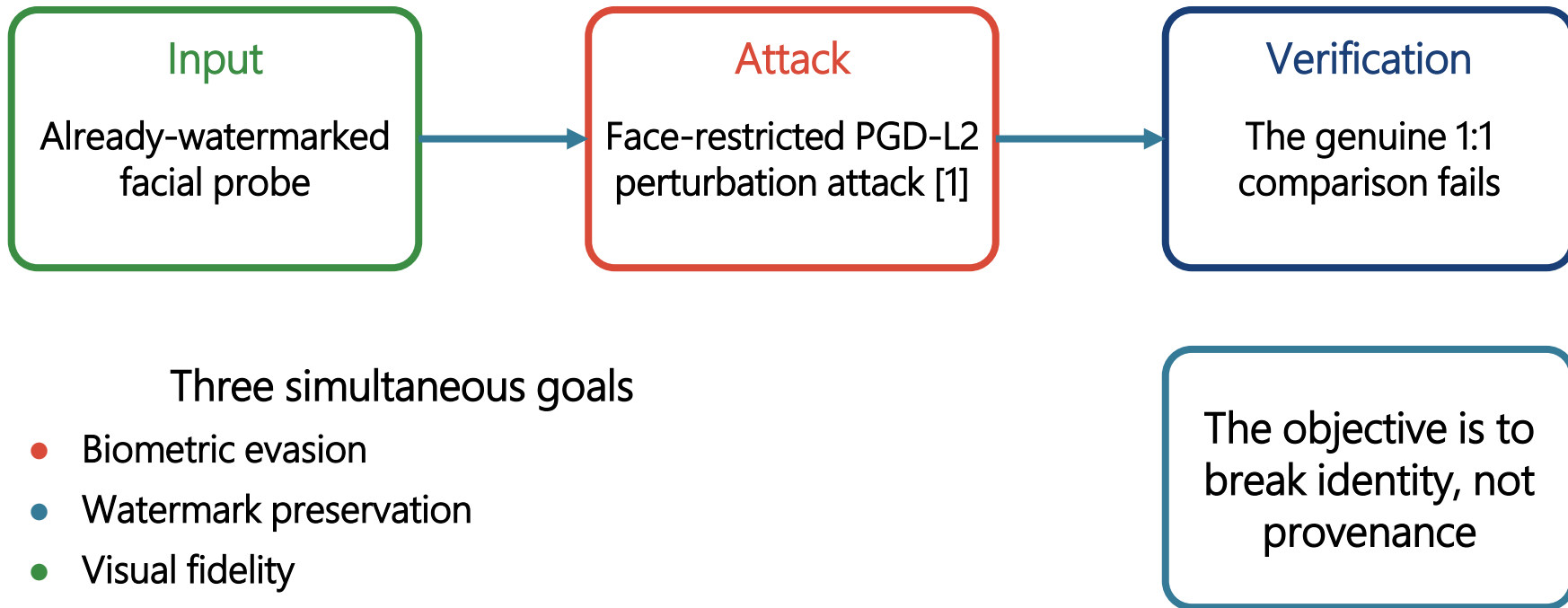
Our question

Can this implication fail under adversarial perturbations?

Failure mode and threat model

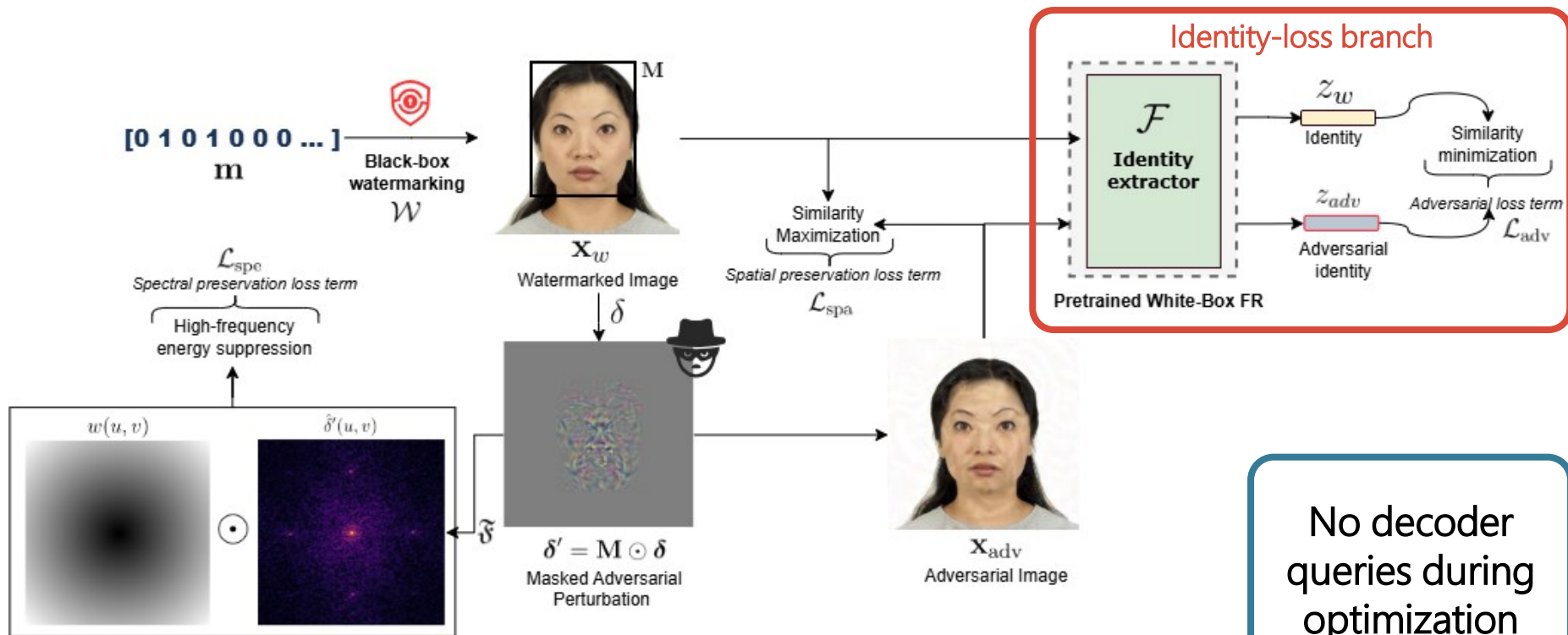


Attack objective: break identity, preserve provenance

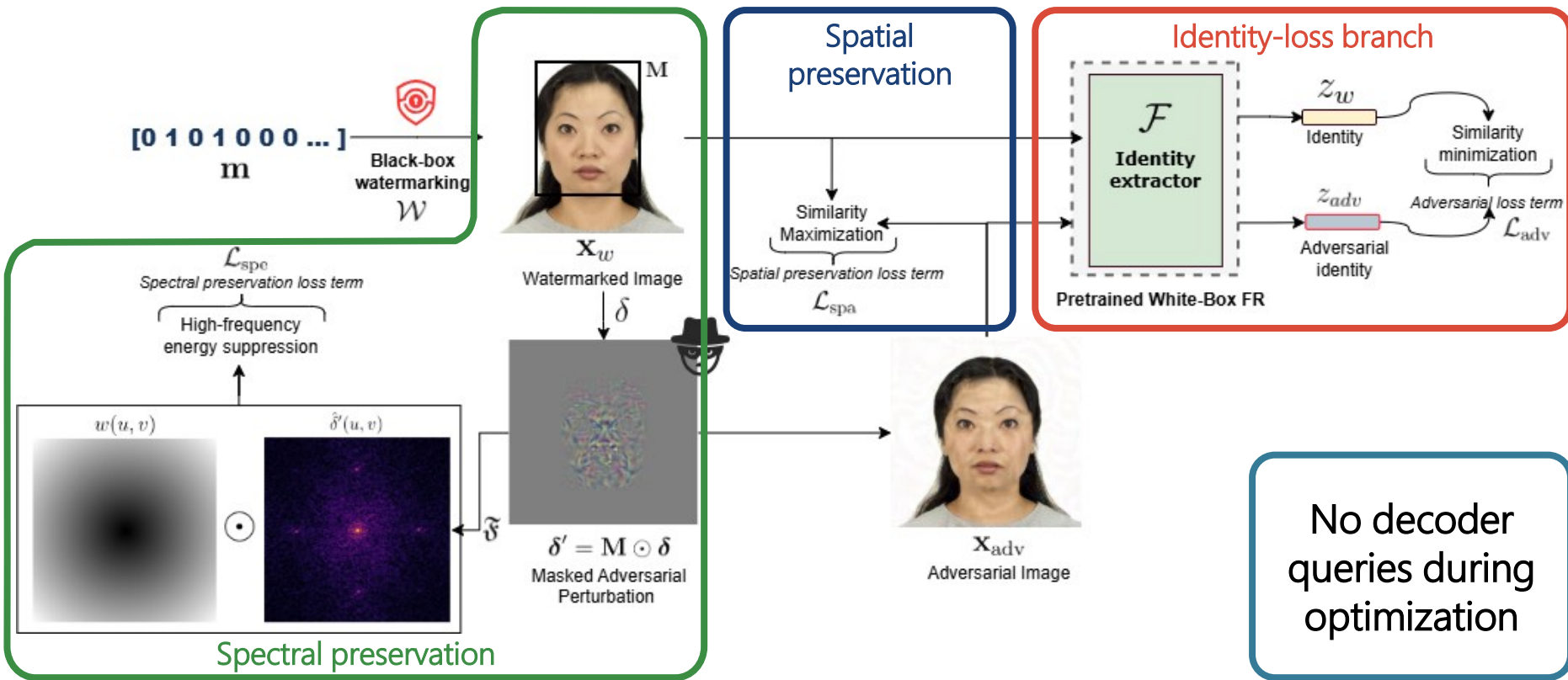


[1] Madry et al. "Towards deep learning models resistant to adversarial attacks," International Conference on Learning Representations, ICLR 2018.

Proposed attack: preserve the watermark without querying it



Proposed attack: preserve the watermark without querying it



Datasets

- CFD: 831 probes, 158 identities [2]
- FaceLab London: 102 probes, 102 identities [3]

Face recognizers

- ArcFace [4]
- FaceNet [5]

[2] Ma et al. "*The chicago face database: A free stimulus set of faces and norming data*," Behavior research methods, 47(4), 2015.

[3] DeBruine and Jones. "*Face research lab london set*," figshare, 2017.

[4] Deng et al. "*Arcface: Additive angular margin loss for deep face recognition*," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.

[5] Schroff et al. "*Facenet: A unified embedding for face recognition and clustering*," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

Watermarks

- StegaFormer (1 bpp, high-capacity) [6]
- FaceSigns (128-bit, low-capacity) [7]

Attack budgets

- $\epsilon_2 \in \{0.5, 1.0, 2.0, 3.0\}$
- Baseline: 80 PGD iterations
- Proposed: 180 PGD iterations

Metrics

- $ASR_{FAR=1\%} / ASR_{FAR=0.1\%}$
- BAR
- PSNR

Only the probe is watermarked and attacked, while the template remains unchanged

[6] Yu et al. "Effective message hiding with order-preserving mechanisms," 35th British Machine Vision Conference, BMVC 2024.

[7] Neekhara et al. "Facesigns: Semi-fragile watermarks for media authentication," ACM Transactions on Multimedia Computing, Communications and Applications, 20(11), 2024.

Reference condition: the watermarking systems are valid

Pre-attack performance

StegaFormer

- BAR: 99.6-100%
- PSNR: 57-59.5 dB

FaceSigns

- BAR: 99.9-100%
- PSNR: 31.6-32.6 dB

Why this matters

- Stable watermark recovery before attack
- The attack starts from valid watermarked probes
- Any post-attack degradation comes from the perturbation

Baseline PGD already exposes the gap at $\epsilon_2=2$

Performance at $\epsilon_2=2.0$

ASR

- $\geq 96\%$ at FAR=0.1%
- $\geq 95\%$ at FAR=1%

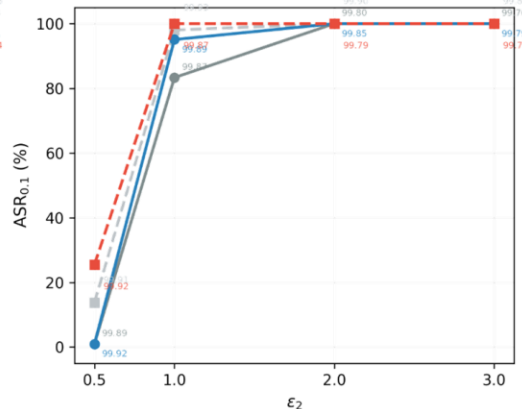
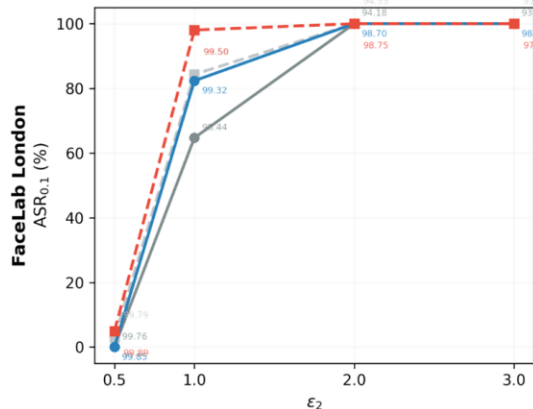
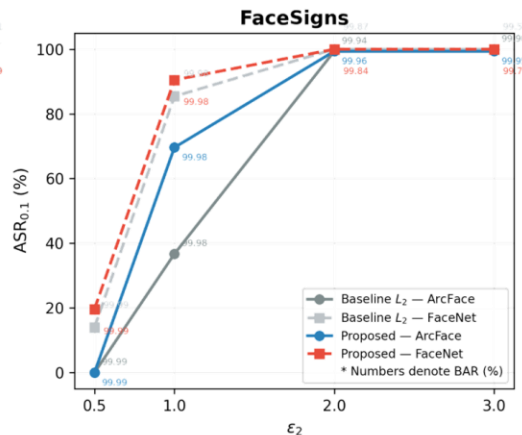
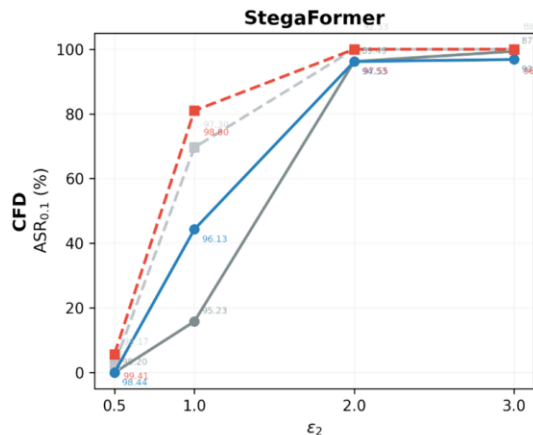
BAR

- StegaFormer: 89.5–94.5%
- FaceSigns: 99.8–99.9%

Key interpretation

- The baseline PGD already exposes the watermark–identity gap
- StegaFormer BAR drops by up to 10.1 pp, motivating the proposed preservation-aware regularization

Spatial-spectral preservation results at $\epsilon_2=2$



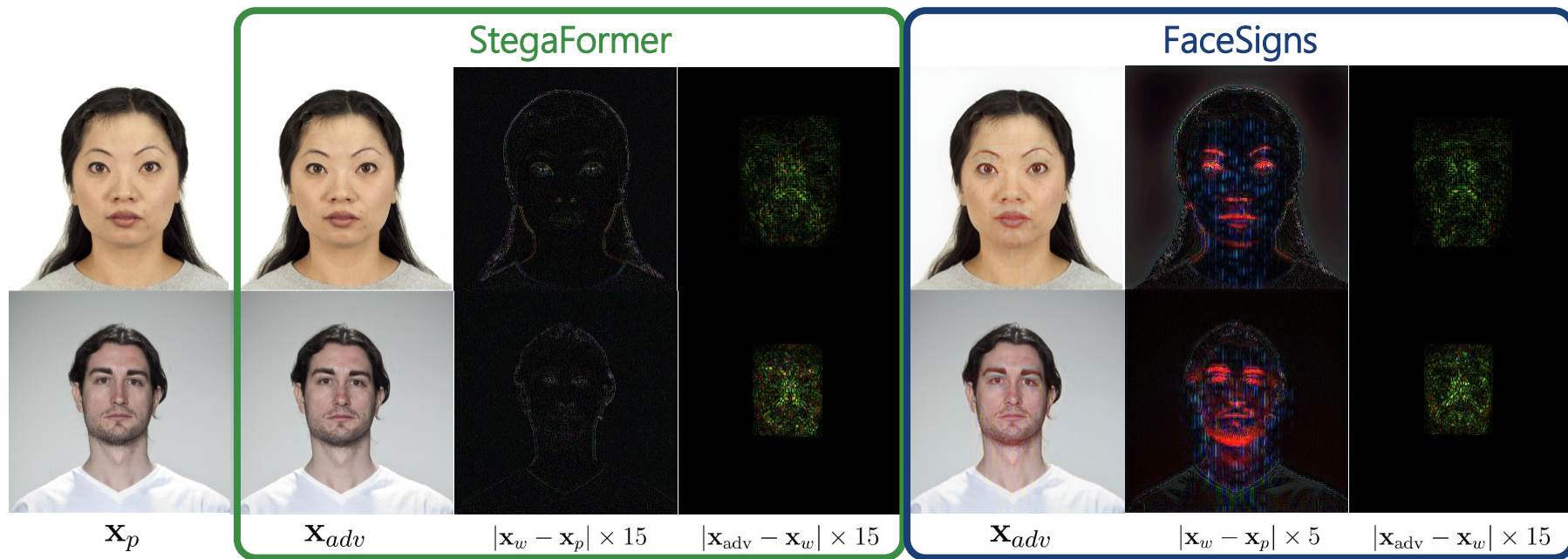
BAR improves by 4.7 pp on average for StegaFormer and remains essentially unchanged for FaceSigns

ASR does not decrease

Takeaway

Preservation improves watermark recovery where needed, without degrading attack success

Visual evidence: faithful probes, preserved recovery, biometric failure

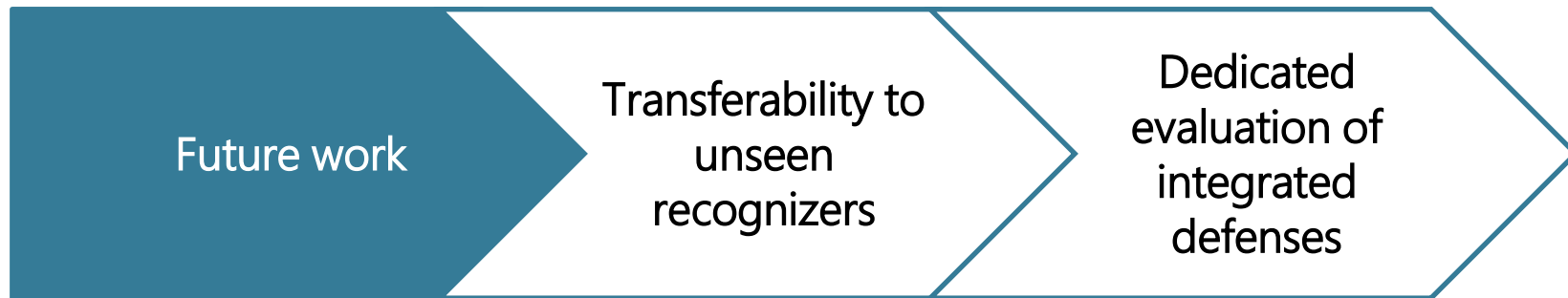


$\epsilon_2 = 2.0$: PSNR > 47 dB

The face-localized perturbation remains visually subtle and preserves the apparent identity

Conclusions: A valid watermark is not a valid identity

- Watermark-valid $\neq$ identity-valid
- Preservation improves recovery while maintaining high evasion
- Provenance and identity require separate guarantees





sAifer Lab

Joint lab on Safety and Security of AI



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Thank you for your attention

Simone Maurizio La Cava

simonem.lac@unica.it

www.saiferlab.ai

Davide Ghiani's PhD grant is partly funded by Dedem SpA under the PNRR program. The authors thank Federico Lama for his valuable advice.

When Biometrics Break (WBB) - IJCB 2026