

(Face) biometrics under attack

On provenance, trust and threat robustness

Guilherme Schardong, Tiago Novello, Luiz Schirmer, Nuno Gonçalves

UFSM IMPA ISR-UC / INCM

20/10/2026



About me



**Guilherme G.
Schardong**

Assistant Professor
UFSM – Cachoeira do Sul

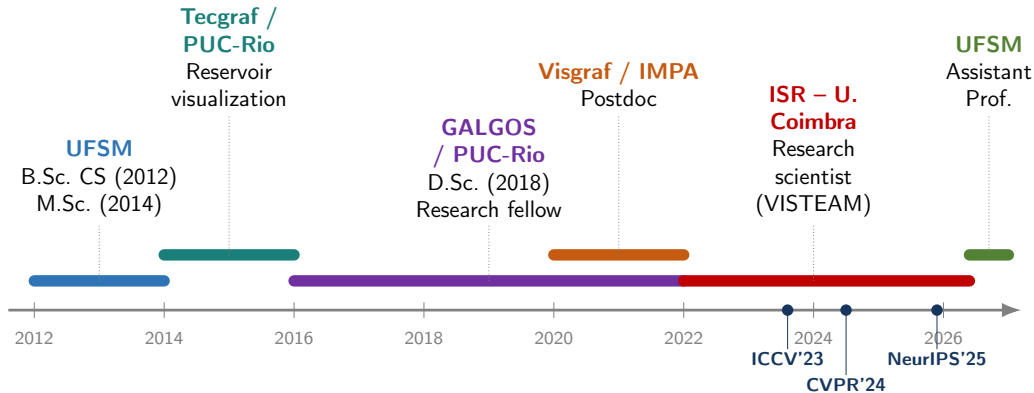
Research interests

- ▶ Implicit neural representations for graphics and vision
- ▶ Morphing, deformation, and registration (2D images, 3D)
- ▶ Biometrics, media security, and digital watermarking

Selected work

- ▶ **ifmorph** – Neural Implicit Morphing of Face Images (CVPR 2024)
- ▶ **FLOWING** – Implicit Neural Flows for Structure-Preserving Morphing (NeurIPS 2025)
- ▶ Neural Implicit Surface Evolution (ICCV 2023)
- ▶ BlockDFake (FCT) – deepfake detection and media certification

Trajectory



Graphics → neural implicit representations → morphing → biometric security.

Roadmap (50 min)

1. Introduction: four dimensions of trust – 8 min
2. Watermarking as a defense layer – 13 min
3. Face morphing as identity manipulation – 15 min
4. Evaluating defenses and biometrics jointly – 7 min
5. Conclusions and research agenda – 5 min

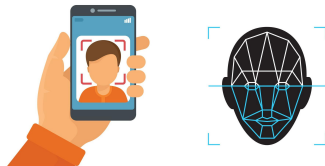
Thesis: a successful match is not, by itself, evidence that the media should be trusted.

Facial biometrics are becoming pervasive

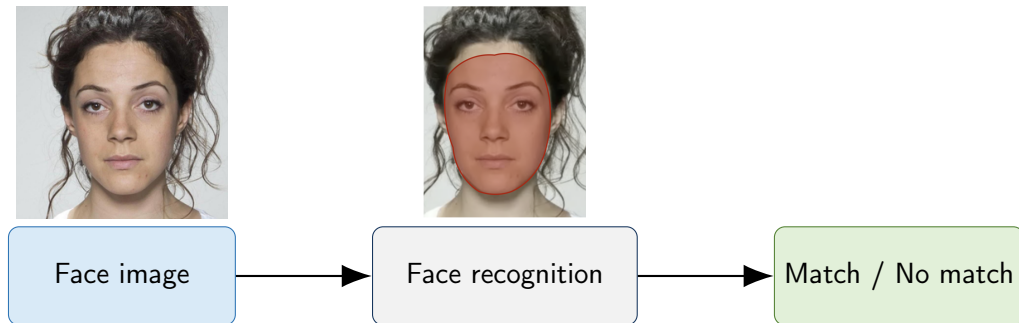
- ▶ Identity documents and e-gates
- ▶ Remote onboarding and digital identity
- ▶ Mobile authentication and access control
- ▶ Financial and government services



A face image is increasingly used not only as data, but as evidence supporting an identity claim.

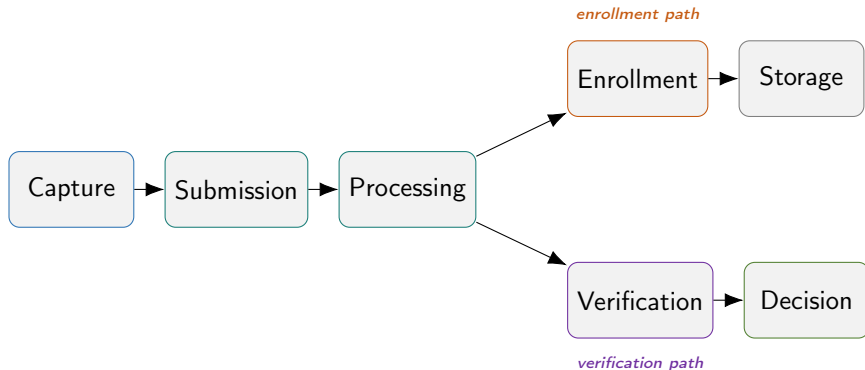


The conventional view



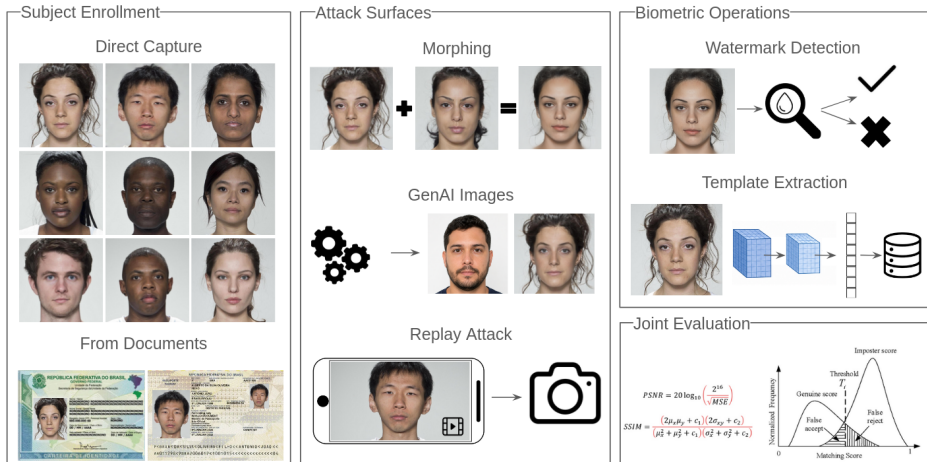
Is a successful match sufficient evidence that the media should be trusted?

The actual media lifecycle



- ▶ Manipulation may happen at any stage
- ▶ Thus, recognition score is only one signal inside a larger trust workflow.

The actual picture: inputs, attacks, defenses, evaluation



Manipulation may happen at any stage; the recognition score is only one signal inside a larger trust workflow.

Four dimensions of trust

Identity	Who does this media represent?
Integrity	Has the media been altered relative to a trusted state?
Provenance	Where did it come from, and what processing history can be verified?
Recognition performance	Does the biometric matcher still operate reliably?

These objectives are related, but they are not equivalent – and the literature often conflates them.

Identity, integrity, provenance: examples

Identity

A morph may match two contributors. The file can look perfectly valid while the identity semantics are compromised.

Integrity

A retouched, recompressed or cropped photo may still match the right person – but it is no longer the original acquisition.

Provenance

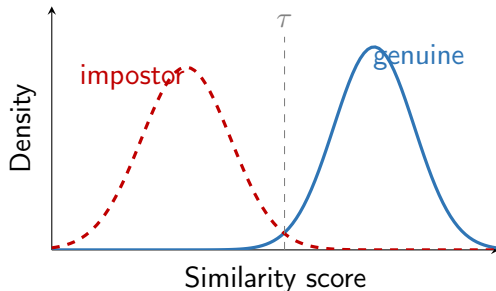
An unmodified file from an unknown source may have good integrity but weak provenance.

“The right person is still in the image” $\neq$ “the recognition system still works well”.

Embedding-based verification in one slide

$$s(l_1, l_2) = \frac{f(l_1)^\top f(l_2)}{\|f(l_1)\| \|f(l_2)\|}$$

- ▶ f : ArcFace-like embedding.
- ▶ Verification (1:1): accept if $s \geq \tau$.
- ▶ FAR, FRR, EER, TAR@FAR summarize the threshold trade-off.



Any media transformation that reshapes these two distributions changes operational behavior.

Watermarking: basic idea



Host image I



Watermarked I_w

Embedder E

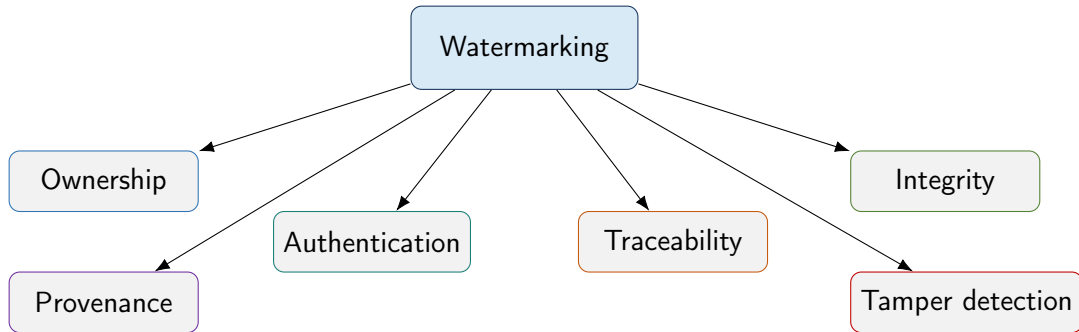
Detector D

Decision / $\hat{m}$

Message m

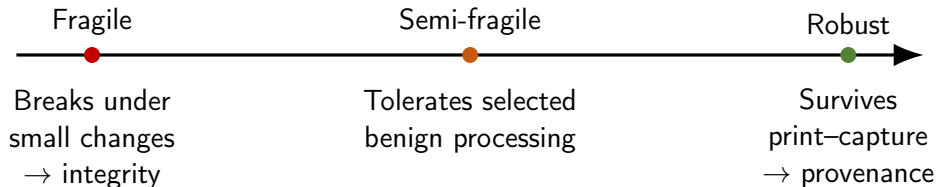
$$I_w = E(I, m, k), \quad \hat{m} = D(I_w, k) \approx m$$

What is a watermark for?



A detected watermark means only what the scheme and its trust assumptions allow it to mean.

Fragile, semi-fragile, and robust



Practical catch

Benign preprocessing (resize, recompression, cropping, contrast normalization) can break a fragile mark with no attack $\Rightarrow$ false tamper alarms. Options: semi-fragile schemes, or *reversible* watermarks removed before forensic processing.

Choose the class from the threat model *and* the expected processing pipeline.

How the mark is embedded

Spatial domain

$$I_w = \text{clip}(I + \alpha W)$$

Simple, controllable; usually fragile to compression and geometry.

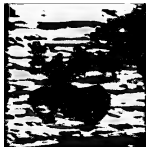
Transform domain

DCT / DWT coefficients (mid-frequencies).

Better robustness / visibility balance.

Learned

Encoder–decoder trained jointly with simulated attacks (HiDDeN, StegaStamp, StampOne, RSTEGA).



All families trade imperceptibility, robustness, payload, and detectability.

Where does the mark enter the workflow?

Trusted capture

Authorized sensor → Capture



Watermark / signature → Enrollment

Unknown media

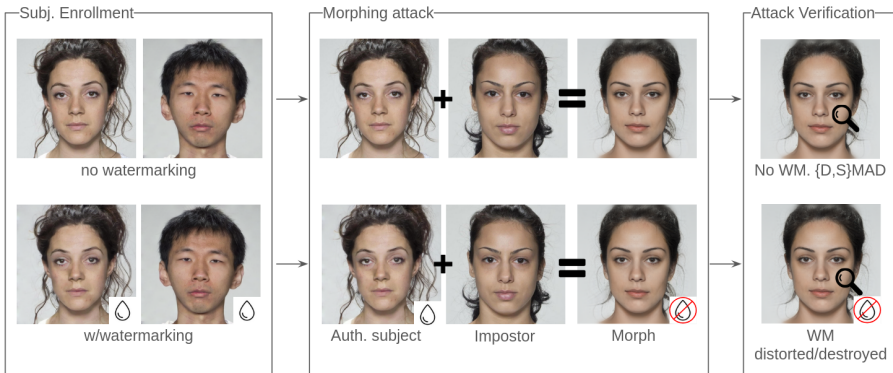
Unknown image



Watermark added later → Enrollment

A mark added after an unknown acquisition cannot retroactively prove that acquisition was trustworthy.

Watermarking at capture as a protection layer



- ▶ **Top:** non-watermarked enrolled image – a morph may pass unless dedicated detectors are used.
- ▶ **Bottom:** watermarked image – morphing alters/destroys the mark $\Rightarrow$ flag for manual check.

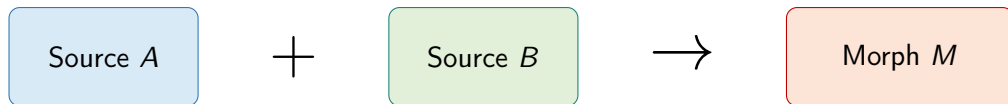
The transition to biometrics

$$I \longrightarrow I_w \quad \Longrightarrow \quad f(I) \longrightarrow f(I_w)$$

$$\text{PSNR}(I, I_w) \uparrow \not\Rightarrow f(I_w) \approx f(I)$$

A watermark can be visually imperceptible and still be biometrically relevant: a defense for integrity may cost recognition performance.

Face morphing as identity manipulation

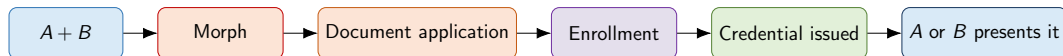


$$s(M, A) > \tau \quad \text{and} \quad s(M, B) > \tau$$

The attacker does not need to “break” recognition. The attack succeeds because the matcher behaves exactly as designed.

Venkatesh et al., IEEE TTS, 2021; Scherhag et al., 2019.

Document-forgery scenario



Enrollment is the critical trust boundary: once the morph is accepted, future matching looks perfectly normal – the one-person/one-identity assumption is broken.

From graphics to biometric attack generation



- ▶ **Classical:** landmarks $\rightarrow$ warp to intermediate geometry $\rightarrow$ blend $\rightarrow$ post-process.
- ▶ **Generative:** MIPGAN, MLSD-GAN; MorDIFF, LADIMO (template inversion), DCMorph.
- ▶ **Implicit neural:** ifmorph (CVPR'24), FLOWING (NeurIPS'25) – smooth, continuous, structure-preserving deformations.

The objective shifts from “good interpolation” to “realistic, operational biometric attack”.

Measuring attack strength: MMPMR and FMMPMR

Morph m with contributors C_m ; $s(m, k, i)$ = score against the i -th probe of contributor k .

$$\text{MMPMR}(\tau) = \frac{1}{M} \sum_m [\bigwedge_{k \in C_m} (\max_i s(m, k, i) \geq \tau)]$$
$$\text{FMMPMR}(\tau) = \frac{1}{M} \sum_m [\bigwedge_{k \in C_m} \bigwedge_i (s(m, k, i) \geq \tau)]$$

Toy example, $\tau = 0.86$:

	Probe 1	Probe 2
Alice	0.95	0.40
Bob	0.90	0.88

MMPMR = **1.0** (best probe each)

FMMPMR = **0.0** (Alice's 0.40 fails)

FMMPMR $\leq$ MMPMR; equal with one probe per contributor.

τ comes from the system's operating point. RMMR = MMPMR + FRR exposes "safe" numbers obtained by over-strict thresholds.

Scherhag et al., BIOSIG 2017; Venkatesh et al., 2020; Ferrara et al., 2022; MorFacing, 2024.

Morphing and the four trust properties

Identity	Compromised: the sample is deliberately compatible with multiple contributors.
Integrity	Depends on the reference point – the file may be internally consistent.
Provenance	Often weak or misleading when the source chain is unknown.
Recognition performance	May remain excellent – that is part of the attack.

Proactive vs. reactive defense

Watermarking (proactive)

- ▶ Requires trusted acquisition.
- ▶ Cannot help with media manipulated before marking.
- ▶ May degrade recognition performance.

Morph detection (reactive)

- ▶ No cooperation from the capture device.
- ▶ Generalization depends on attack family and image history.
- ▶ Frequency-based detectors inspect the same spectrum frequency-domain marks modify.

Detection $\neq$ secure enrollment. The two must be evaluated jointly – an open problem.

Two communities, two metric sets

Watermarking

PSNR, SSIM
BER, payload
Robustness to expected processing

Biometrics

FAR, FRR, EER
ROC / DET, TAR@FAR
MMPMR / FMMPMR

Embedding drift: $d(I, I_w) = 1 - \cos(f(I), f(I_w))$ – useful, but per-image only.

The question is not whether each component works in isolation, but whether the combined system remains trustworthy.

What a joint protocol should report

Perceptual	PSNR, SSIM
Watermark	Detection, BER, payload
Robustness	Survival after JPEG, resize, crop, recapture
Representation	Embedding drift
Verification	Genuine/impostor dists., TAR@FAR
Morph threat	MMPMR / FMMPMR
Trust	Integrity / provenance checks

Comparison scenarios

Clean $\leftrightarrow$ Clean
Clean $\leftrightarrow$ Watermarked
Watermarked $\leftrightarrow$ Clean
Watermarked $\leftrightarrow$ Watermarked

Asymmetric cases matter: clean legacy gallery vs. newly marked probes.

Expect a trade-off surface (utility $\times$ robustness $\times$ quality), not a single optimum.

From joint metrics to a trust architecture



**Trust emerges from coordinated controls and explicit decision logic,
not from a single algorithm.**

What a watermark can – and cannot – establish

Potentially supports

- ▶ provenance
- ▶ integrity
- ▶ source authentication
- ▶ traceability

Does not automatically prove

- ▶ identity validity
- ▶ absence of morphing
- ▶ recognition reliability
- ▶ unchanged content (for robust marks)

Research agenda

1. **Identity-aware watermarking:** embedding preservation in the objective; avoid identity-sensitive regions/frequencies.
2. **Joint benchmarks:** watermark, morph, provenance, and recognition metrics in one protocol.
3. **Adaptive attackers:** jointly optimize morph success, detector evasion, and watermark removal/preservation.
4. **Privacy-preserving provenance:** authenticity without unnecessary linkage or metadata exposure.
5. **Literacy:** developers, operators, decision makers, and users must know what each signal proves.

Take-home message

Face recognition asks:

“Do these samples match?”

Trustworthy biometric media asks:

“Why should this identity claim
be trusted?”

Selected references

-  S. Venkatesh, R. Ramachandra, K. Raja, C. Busch. "Face Morphing Attack Generation and Detection: A Comprehensive Survey." IEEE TTS, 2021.
-  U. Scherhag et al. "Biometric Systems under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting." BIOSIG, 2017.
-  I. Medvedev, N. Gonçalves. "MorFacing: A Benchmark for Estimation Face Recognition Robustness to Face Morphing Attacks." 2024.
-  G. Schardong et al. "Neural Implicit Morphing of Face Images." CVPR, 2024.
-  A. Bizzi et al. "FLOWING: Implicit Neural Flows for Structure-Preserving Morphing." NeurIPS, 2025.
-  W. Wan et al. "A Comprehensive Survey on Robust Image Watermarking." Neurocomputing, 2022.
-  M. Tancik, B. Mildenhall, R. Ng. "StegaStamp: Invisible Hyperlinks in Physical Photographs." CVPR, 2020.
-  Young et al. Media integrity and provenance, 2026.

Full bibliography in the accompanying paper.

Thank you!

Guilherme G. Schardong

UFSM – Cachoeira do Sul

E-mail guilherme.schardong@ufsm.br

Web schardong.github.io

ORCID [0000-0003-3927-0852](https://orcid.org/0000-0003-3927-0852)

Code github.com/schardong/facetrust

Co-authors

Tiago Novello (IMPA)

Luiz Schirmer (UFSM)

Nuno Gonçalves (ISR-UC, INCM)

Questions?

Funding: EU Horizon Europe ACHILLES (GA 101189689); FCT UID/00048 (ISR-UC).