(Face) biometrics under attack

On provenance, trust and threat robustness

Guilherme Schardong, Tiago Novello, Luiz Schirmer, Nuno Gonçalves

SIBGRAPI 2026 Tutorial Session

29/09/2026







Tutorial goals

By the end, participants should be able to

- Distinguish identity, integrity, provenance, and recognition performance.
- Explain key watermarking paradigms and their trade-offs.
- Describe face morphing as an enrollment/document threat.

And also

- Interpret biometric verification metrics.
- Design a joint watermarking–biometrics evaluation protocol.
- Identify open problems for trustworthy biometric media.

Tentative roadmap and timing

1. Introduction – 20 min
2. Watermarking concepts and developments – 35 min
3. Face morphing for identity manipulation – 35 min
4. Coffee break – 30 min
5. Joint evaluation – 60 min
6. Conclusions and future directions – 20 min

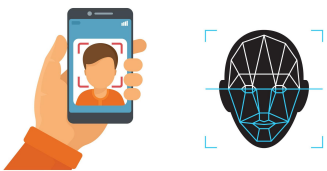
The first half separates the concepts. The second half focuses on how to evaluate them together.

Facial biometrics are becoming pervasive

- Identity documents and e-gates
- Remote onboarding and digital identity
- Mobile authentication
- 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

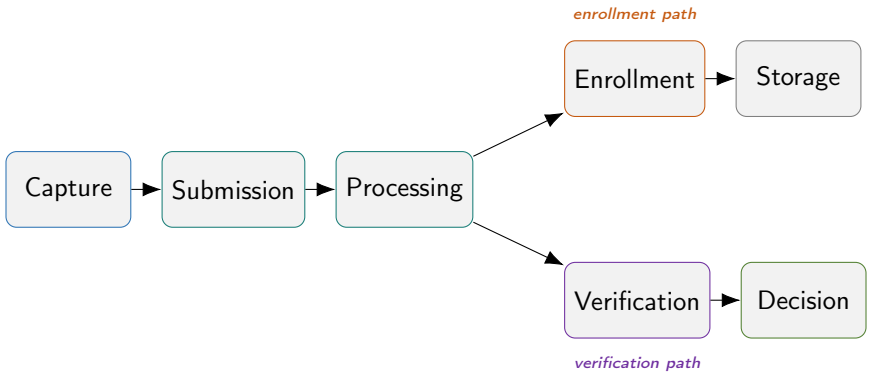




Face image → Face recognition → Match / No match

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.

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.

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 correct subject may still be represented by an image altered after acquisition.

Provenance

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

Recognition performance is yet another question: how does the matcher behave on each case?

Verification versus identification

Verification: 1:1

Claimed identity + probe

$\Downarrow$

same person or not?

Identification: 1:N

Probe + gallery

$\Downarrow$

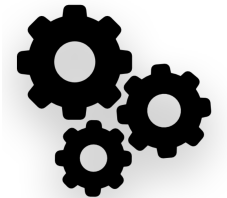
which enrolled identity?

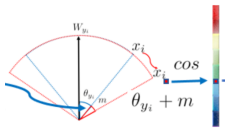
This tutorial focuses primarily on verification because it provides a clean framework for studying score distributions and perturbations.

Embedding-based face recognition







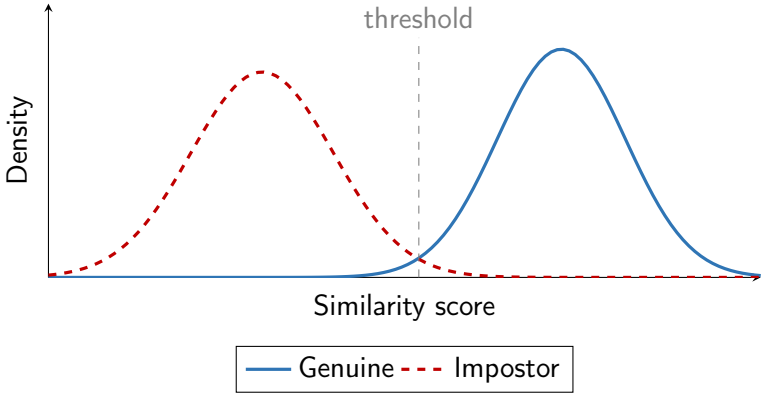


ArcFace

Face image → Alignment → ArcFace-like model → Embedding $f(I)$

$$s(I_1, I_2) \approx \frac{f(I_1)^\top f(I_2)}{\|f(I_1)\| \, \|f(I_2)\|}$$

Genuine and impostor score distributions



Watermarking: basic idea





Host image I

Message m

Embedder E

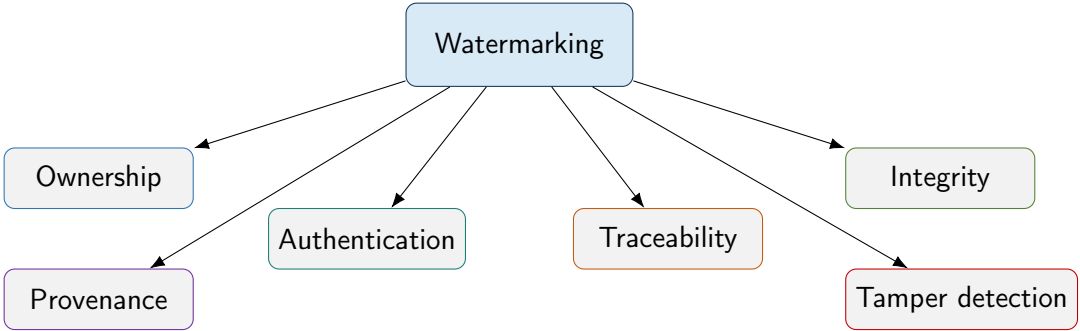
Watermarked I_w

Detector D

Decision / $\hat{m}$

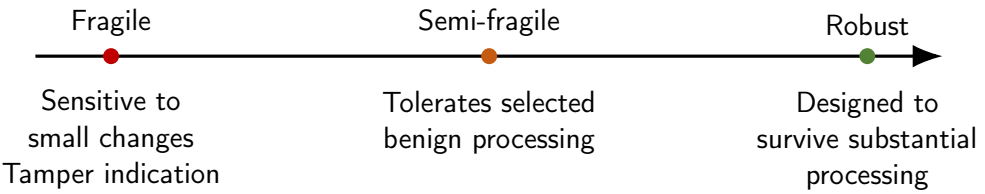
$$I_w = E(I, m, k), \qquad \hat{m} = D(I_w, k) \approx m$$

What is the purpose of a watermark?



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

Fragile, semi-fragile, and robust watermarking

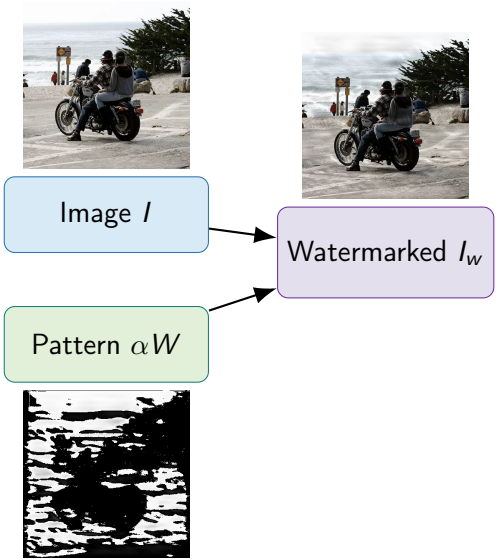


Robustness is not always the desired property: integrity and provenance can require different behaviors.

Spatial-domain watermarking

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

- Directly modifies pixels.
- Simple and controllable baseline.
- Embedding strength α controls distortion.
- Typically fragile to compression, resizing, and geometric changes.

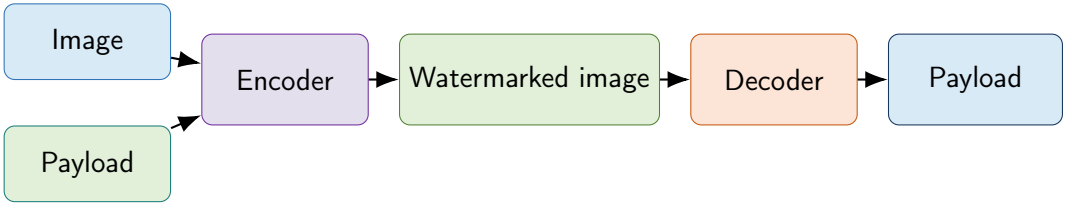


Transform-domain watermarking



- Mid-frequency coefficients often provide a compromise between visibility and robustness.
- The host signal is perturbed in a structured spectral manner.

Learned watermarking



- Embedding and recovery are optimized jointly.
- Differentiable attacks can be injected during training.
- The loss can include visual, decoding, and task-specific objectives.

0-bit versus multi-bit watermarking

0-bit

"Is the expected mark present?"

No arbitrary user-selected payload is

decoded.

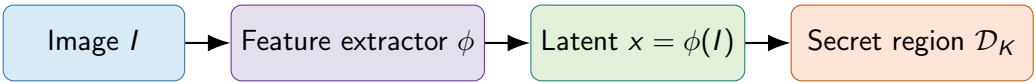
Multi-bit

"What message is encoded?"

A bit string or message is recovered.

0-bit means zero payload bits, not zero information or zero perturbation.

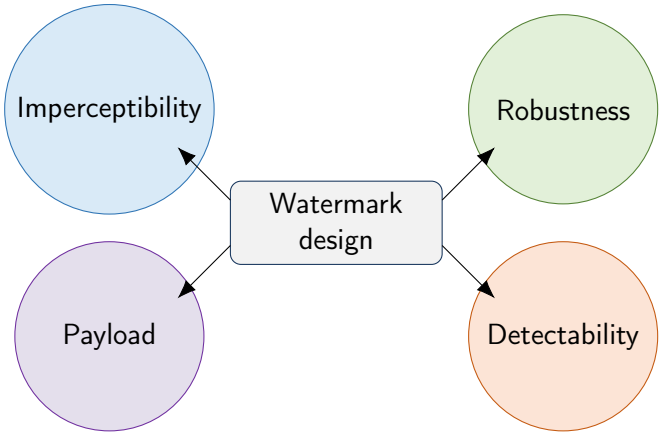
0-bit watermarking as a latent-space test



$$H_0 : x \notin \mathcal{D}_K \qquad \text{vs.} \qquad H_1 : x \in \mathcal{D}_K$$

Example: SSL watermarking / self-supervised latent-space watermarking.

Classical watermarking trade-offs



Robustness tests

Signal processing

- JPEG compression
- Blur
- Noise
- Contrast changes

Geometric

- Resize
- Crop
- Rotation
- Translation

Physical / pipeline

- Screen recapture
- Print–scan
- Document re-encoding

Robustness must be defined relative to an expected processing or attack model.

Watermark quality metrics

Media quality

- PSNR
- SSIM
- Perceptual metrics

Watermark performance

- Detection rate
- Bit Error Rate (BER)
- Payload
- Robustness after transformations

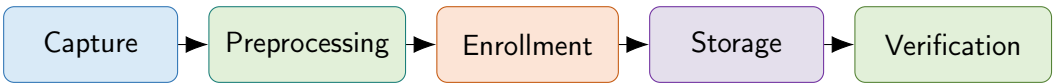
These metrics say little, by themselves, about downstream biometric behavior.

Where should a biometric watermark be inserted?

trusted-capture mark?

enrollment mark?

storage mark?



The same watermark algorithm can carry very different trust meaning depending on where the mark enters the workflow.

Trusted capture versus watermarking unknown media

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.

The transition to biometrics

$$I \longrightarrow I_w$$

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

A watermark can be visually imperceptible and still be biometrically relevant.

Face morphing as identity manipulation



Goal: create one image that retains biometric similarity to multiple contributors.

Venkatesh et al., IEEE TTS, 2021.

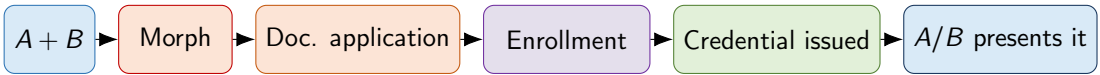
Biometric objective of a morph attack

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

- M: morphed enrollment image
- A, B: contributing identities
- τ: acceptance threshold

The attacker does not need to "break" recognition. The attack can succeed because the matcher behaves exactly as designed.

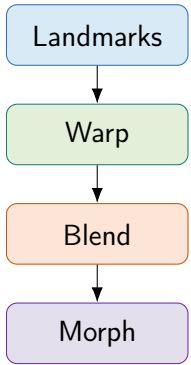
Document-forgery scenario



Enrollment is the critical trust boundary: once the manipulated image is accepted, future matching may look perfectly normal.

Classical morph generation

1. Detect or annotate corresponding landmarks.
2. Warp source images to an intermediate geometry.
3. Blend appearance information.
4. Post-process for visual plausibility.



Classical graphics lineage: feature-based morphing and halfway-domain formulations.

From graphics to biometric attack generation



The objective increasingly shifts from "good interpolation" to "strong, realistic, operational biometric attack".

GAN-based morphs

- Latent interpolation enables photorealistic synthesis beyond explicit landmark warping.
- Identity-aware methods optimize for contributor similarity, not only appearance.
- Attack evaluation becomes a first-class design criterion.

Representative directions

MIPGAN; analyses of GAN-based morph threats; MLSD-GAN and latent semantic disentanglement.

Zhang et al.; Sarkar et al.; Aravinda Reddy et al.

Diffusion and biometric template inversion

- Diffusion-based synthesis offers strong reconstruction quality and flexible conditioning.
- Template inversion brings morph generation closer to the biometric representation itself.
- Modern methods can jointly target realism and recognition vulnerability.

Representative directions

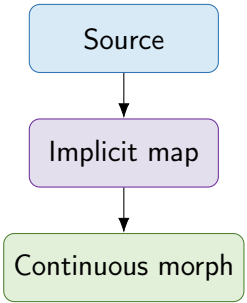
MorDIFF; LADIMO; DCMorph.

Damer et al.; Grimmer and Busch; Chettaoui et al.

Neural implicit face morphing: ifmorph

- Coordinate-based neural networks represent continuous deformation and blending.
- Smooth time-dependent mappings connect the source images.
- Variational regularization can be expressed through derivatives of the implicit map.

Schardong et al., CVPR 2024.



Structure-preserving neural flows: FLOWING

- Recasts morphing as an implicit neural flow.
- Emphasizes continuity, invertibility, and temporal coherence.
- Encodes useful structural properties in the model rather than relying only on external regularization.

Implicit neural representations broaden the morphing landscape beyond discrete warp fields and latent interpolation.

Bizzi et al., NeurIPS 2025.

Two traditions of morphing

Computer graphics	Biometric attack perspective
Smooth deformation	Attack success
Correspondence	Dual identity
Visual coherence	Matcher vulnerability
Structure preservation	Detector evasion / generalization
Continuous transformation	Operational security impact

Modern morphing methods increasingly cross this boundary.

Morphing attack potential

- Attack success should be measured against the target recognition system and operational threshold.
- A strong attack must preserve contributor similarity under realistic capture and document conditions.
- Aggregate metrics such as MMPMR / FMMPMR / MAP variants summarize different attack criteria.

Visual quality alone is not a meaningful measure of biometric attack strength.

Ferrara et al.; Medvedev and Gonçalves, MorFacing.

MMPMR: formal definition

Mated Morph Presentation Match Rate – Given a morph m with contributors C_m, and probe scores $s(m, k, i)$ (i-th probe of contributor k):

$$\mathrm{MMPMR}(\tau) = \frac{1}{M} \sum_{m=1}^{M} \Big[\bigwedge_{k \in C_m} \left(\max_i s(m, k, i) \geq \tau \right) \Big]$$

- Per contributor: take the **best** probe (attacker only needs one good photo of each accomplice).
- Across contributors: **all** of them must be beaten (logical AND).
- τ comes from the system's own operating point (e.g., threshold at FAR = 0.1%).

MMPMR is optimistic FOR THE ATTACKER: one good photo per accomplice is enough.

Scherhag et al., BIOSIG 2017.

FMMPMR: formal definition

Fully Mated Morph Presentation Match Rate – Same setup as MMPMR: morph m, contributors C_m, probe scores $s(m, k, i)$ (i-th probe of contributor k).

$$\mathrm{FMMPMR}(\tau) = \frac{1}{M} \sum_{m=1}^{M} \Big[\bigwedge_{k \in C_m} \bigwedge_{i} \big(s(m, k, i) \geq \tau \big) \Big]$$

- Per contributor: **every** probe must pass – no "best photo" exception.
- Across contributors: still all of them, same as MMPMR.
- Direct consequence: $\mathrm{FMMPMR}(\tau) \leq \mathrm{MMPMR}(\tau)$ always, for the same threshold and data – the two coincide only when each contributor has a single probe.

FMMPMR is pessimistic FOR THE ATTACKER: the morph must survive every captured photo of each accomplice, not just their best one.

Venkatesh et al., 2020.

FMMPMR: when does the distinction matter?

Probes per contributor	MMPMR vs. FMMPMR	Relevant here?
1 (single reference photo)	identical	no
$\geq$2 (multiple captures)	can diverge	**yes**

Our own dataset already has this: FRLL gives multiple expressions per identity – exactly the situation where FMMPMR earns its keep over MMPMR.

If your evaluation protocol only compares a morph against *one* photo per contributor, don't bother computing both – report MMPMR and say so. If you have several photos, the gap between the two numbers is itself informative: a wide MMPMR/FMMPMR gap means the attack is fragile (works only with a lucky reference photo); a narrow gap means it is robust across capture conditions.

MMPMR vs. FMMPMR: a toy example

Morph m_1 = Alice + Bob, two probes each. $\tau = 0.86$.

Contributor	Probe 1	Probe 2	max
Alice	0.95	0.40	**0.95**
Bob	0.90	0.88	**0.90**

At $\tau = 0.86$

MMPMR: Alice ✓ (0.95), Bob ✓ (0.90) $\rightarrow$ **1.0**

FMMPMR: Alice's 2nd probe (0.40) fails $\rightarrow$ **0.0**

Same morph, same threshold, opposite verdicts – because they ask different questions. They only agree when there is a single probe per contributor.

RMMR: is that MMPMR number honest?

$$\mathrm{RMMR}(\tau) = \mathrm{MMPMR}(\tau) + \mathrm{FRR}(\tau)$$

$\mathrm{FRR}(\tau)$: false-reject rate on GENUINE bona-fide comparisons, at the SAME threshold.

Threshold	MMPMR	FRR	RMMR
τ_1 (calibrated, FAR=0.1%)	5%	1%	**6%**
τ_2 (overly strict)	5%	40%	**45%**

Same MMPMR, wildly different RMMR: τ_2 only looks safe because it also locks out legitimate users. RMMR is not a value in [0,1] – it is a comparative index, not a probability.

Scherhag et al., BIOSIG 2017.

Morphing and the four trust properties

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

Detection is not prevention

Morph detection $\neq$ secure enrollment

- Detection is one control inside a larger process.
- Detector generalization depends on attack families and image history.
- Trusted acquisition and provenance can prevent some ambiguity before detection is needed.

The fundamental enrollment problem

Once manipulated media is accepted as trusted enrollment evidence, subsequent biometric recognition may operate exactly as designed while the identity claim itself remains invalid.

Coffee break

30 minutes

After the break: how do we evaluate watermarking and biometrics together?

Recap: before the break

Watermarking

- $I_w = E(I, m, k)$, $\hat{m} = D(I_w, k)$
- Fragile $\leftrightarrow$ semi-fragile $\leftrightarrow$ robust
- 0-bit (detection) vs. multi-bit (payload)
- Trade-off: imperceptibility, robustness, payload, detectability

Face morphing

- $s(M, A) > \tau$ **and** $s(M, B) > \tau$
- Landmark warping $\rightarrow$ GAN $\rightarrow$ diffusion $\rightarrow$ INR
- Enrollment is the critical trust boundary
- Attack succeeds *because* the matcher works as designed

Recap: before the break

Both sections kept returning to the same four pillars:

Identity | Integrity | Provenance | Recognition performance

Next: we stop treating watermarking and morphing separately and evaluate both against the same recognition system, simultaneously.

Two communities, two metric sets

Watermarking
PSNR
SSIM
BER
Payload
Robustness

Biometrics
FAR
FRR
EER
ROC / DET
TAR@FAR

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

Visual quality is not biometric utility

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

- Small perturbations may affect discriminative structures disproportionately.
- Spatial location and frequency content may matter more than global energy.
- Perceptual metrics were not designed to measure identity preservation.

Embedding drift

$$d(I, I_w) = 1 - \cos\big(f(I), f(I_w)\big)$$

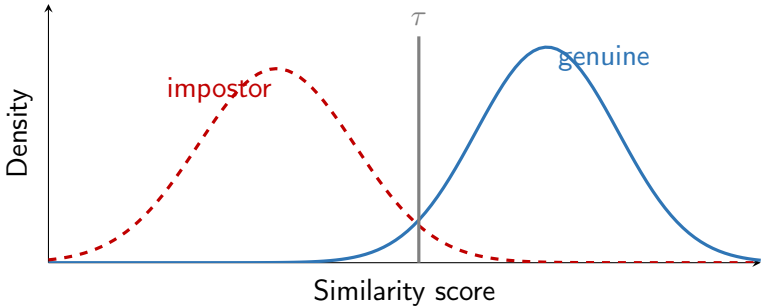
Useful because

It directly measures representation change induced by watermarking.

But insufficient because

Operational performance depends on genuine and impostor score distributions, not only per-image drift.

FAR, FRR, and threshold trade-offs



$$\mathrm{FAR}(\tau) = P(s \geq \tau \mid \text{impostor}), \qquad \mathrm{FRR}(\tau) = P(s < \tau \mid \text{genuine})$$

ROC, DET, EER, and operational points

ROC / DET

Threshold sweeps expose the trade-off between false accepts and genuine acceptance / false rejects.

EER and TAR@FAR

EER is a compact summary; TAR at a fixed low FAR is often more operationally meaningful.

Watermarking should be judged by its effect on the operating region that matters, not only by average cosine similarity.

Four core comparison scenarios

Clean $\leftrightarrow$ Clean

Watermarked $\leftrightarrow$ Clean

Clean $\leftrightarrow$ Watermarked

Watermarked $\leftrightarrow$ Watermarked

Asymmetric cases are especially important in real enrollment/probe pipelines.

Watermark strength sweep

$$\alpha \in \{0, 0.25, 0.5, 1, 2, 4, \ldots\}$$

For each α, record:

Visual	Watermark	Biometric
PSNR	Detection	Embedding drift
SSIM	BER	EER
	Robustness	TAR@FAR

Region-aware watermarking experiments

Whole image	Central face	Periocular	Background
same energy	identity-rich	highly discriminative	low biometric relevance

Is equal distortion energy equally harmful everywhere?

Robustness transformations: evaluate both systems

Transformation	WM detection / BER	Image quality	ArcFace behavior
JPEG	?	?	?
Resize	?	?	?
Blur	?	?	?
Crop	?	?	?
Rotation	?	?	?
Screen recapture	?	?	?

A transformation can be benign for biometrics but destructive to the watermark – or vice versa.

Joint evaluation matrix

Dimension	Example metrics
Perceptual quality	PSNR, SSIM, perceptual similarity
Watermark reliability	Detection rate, BER, payload
Robustness	Survival after expected processing/attacks
Representation	Embedding drift
Verification	Genuine/impostor distributions, EER, TAR@FAR
Morph threat	Attack success / morph attack potential
Trust	Provenance or integrity verification success

Dataset used in this tutorial: FRLL

Face Research Lab London Set (FRLL)

- 102 identities (53 M / 49 F), 1350×1350px, uniform illumination.
- Multiple views per identity: neutral/smiling, frontal and 3/4 profile.
- 189-point landmark templates included (Psychomorph / WebMorph-ready).
- Signed consent for research use, incl. publications and presentations.
- Released under CC BY 4.0.



FRLL – Where to get it?

Where to get it

- `https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666`
- Pre-built morphs on top of FRLL (AMSL protocol; OpenCV, FaceMorpher, WebMorph, StyleGAN2): FRLL-Morphs, `https://www.idiap.ch/en/scientific-research/data/frll-morphs`

We use FRLL specifically because consent and licensing are explicit – fitting, for a tutorial about media provenance and trust.

DeBruine & Jones, 2017 (figshare, DOI 10.6084/m9.figshare.5047666).

Practical activity: goal

Question

How does increasing watermark strength change visual quality, watermark reliability, and biometric similarity at the same time?

- Use a small set of aligned face images.
- Use ArcFace (or provided embeddings).
- Use a simple spatial watermark baseline.
- Sweep α and apply a few post-processing transformations.

Suggested duration: 20 minutes.

Practical activity: starter loop

```
alphas = [0.25, 0.5, 1.0, 2.0, 4.0]

emb_clean = arcface(image)

for alpha in alphas:
    wm, meta = embed_spatial_bits(
        image, bits, alpha=alpha, seed=1234)

    emb_wm = arcface(wm)

    log(alpha,
        psnr(image, wm),
        cosine_similarity(emb_clean, emb_wm),
        detect_watermark(wm, meta))
```

Practical activity: plots to produce

Plot A

Watermark strength α
vs.
PSNR / SSIM

Plot B

Watermark strength α
vs.
embedding drift

Then repeat after JPEG or resize and add watermark detection/BER.

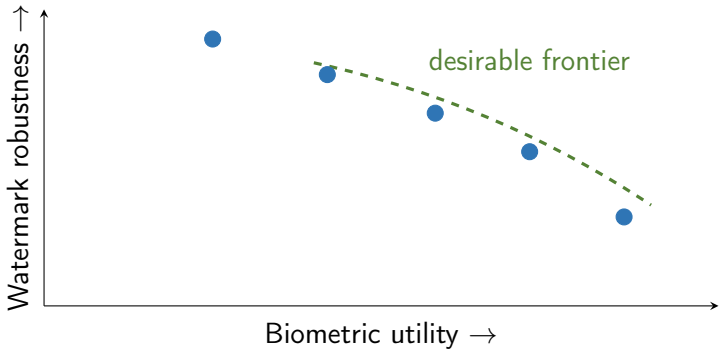
Does visual degradation track biometric degradation?

Practical activity: discussion prompts

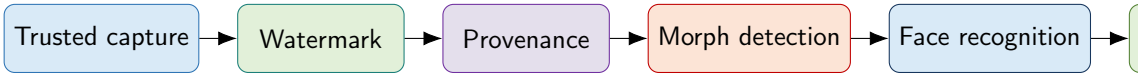
1. At what α does the watermark become visibly objectionable?
2. At what α does embedding drift become substantial?
3. Does watermark detection fail before biometric similarity does, or after?
4. Would you choose the same operating point for enrollment and verification media?
5. Is there a single "best" α?

The expected conclusion is a trade-off surface, not a single scalar optimum.

Pareto view of joint design



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)

Toward identity-aware watermarking

- Add embedding preservation to the watermark optimization objective.
- Control perturbations in identity-sensitive spatial/frequency regions.
- Optimize for biometric operating points, not only perceptual quality.
- Study recognition-aware robustness under realistic processing.

The goal is not merely an invisible watermark, but a watermark whose biometric impact is understood and controlled.

Research agenda

1. **Joint benchmarks**: watermark, morph, provenance, and recognition metrics in one protocol.
2. **Provenance-assisted morph defense**: trusted capture as an active security signal.
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. **Beyond still images**: video, multimodal biometrics, and audiovisual provenance.

Literacy is part of trustworthy deployment

Developers	Operators	Decision makers	Users
What does the model measure?	What does a signal prove?	What assumptions hold?	What are the limits?

As facial biometrics become everyday infrastructure, technical progress alone is insufficient.

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.
- 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.
- Self-supervised image watermarking / 0-bit latent-space watermarking work and codebase.
- NIST FATE MORPH operational guidance and evaluations.

Full bibliography and practical notebook can be distributed with the tutorial material.

(Face) biometrics under attack

On provenance, trust and threat robustness

Guilherme Schardong, Tiago Novello, Luiz Schirmer, Nuno Gonçalves

SIBGRAPI 2026 Tutorial Session

29/09/2026







https://schardong.github.io