MCMI WORKSHOP · ICPR 2026 · AUG 21, 2026

Sunglasses-Region Adversarial Perturbations

for Feature-Space Face Recognition Attacks

Sheng-Pin Chang†, Chun-Yao Chen†, I-Syuan Lee, Chia-Po Wei *

Department of Electrical Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan

† Equal contribution. ** Corresponding author — cpwei@mail.nsysu.edu.tw*

BACKGROUND · 1

Why Face Recognition Security Matters

1. **Authentication**
   Phone unlock, payment, account access
2. **Access control**
   Border gates, campus/building entry
3. **Identity verification**
   KYC, enrolment, watchlists
4. **Media applications**
   Tagging, retrieval, organisation

LFW verification accuracy

99.67%

clean, no overlay

▼

22.17%

under adversarial sunglasses (dodging)

Key point: clean accuracy does not guarantee robustness.

The Modern Face Recognition Pipeline

Face image
aligned crop
112 × 112

→

Extractor
$F(\cdot)$, frozen
deep backbone

→

Embedding
$Y \in \mathbb{R}^d$
ℓ_2-normalised

→

Similarity
$\cos\_sim(Y_1, Y_2)$
or ℓ_2 distance

→

Decision
accept / reject
vs. τ

Representative embedding backbones

- **FaceNet —** triplet embedding learning
- **ArcFace —** angular margin loss
- **AdaFace —** adaptive margin — the attacked model

Decision lives in feature space

Identity is decided by feature-vector comparison, not a classifier head. The attacker only needs to move an embedding.

Verification Is Not Classification

Classification

closed-set — a fixed list of identities

INPUT
one image

OUTPUT
a label over N known identities

DECISION
argmax over classifier logits

ATTACK TARGET
the classifier head

Verification

open-set — the deployed setting

INPUT
a pair of images (X_1, X_2)

OUTPUT
same identity, or not

DECISION
$d(Y_1, Y_2) \leq \tau$ on embeddings

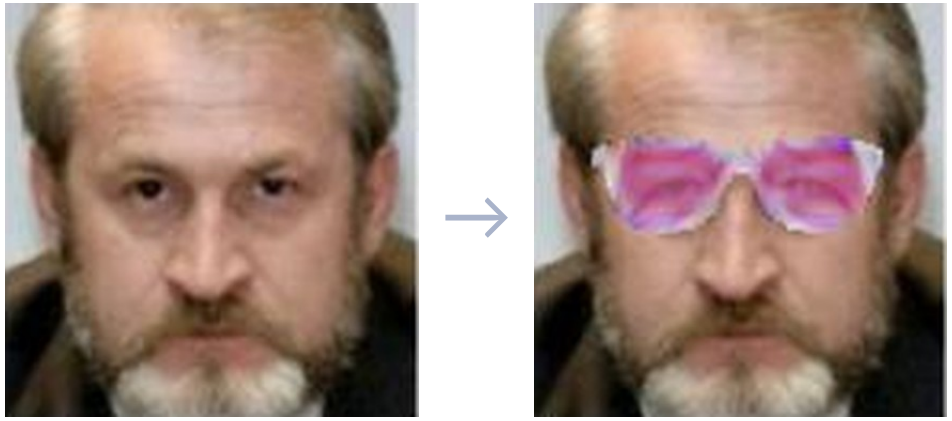
ATTACK TARGET
the embedding itself

Therefore: deployed attacks must manipulate feature similarity, not classifier output.

WEAK POINT · 1

A Local Edit, A Global Feature Shift

Image space — a small, localised edit

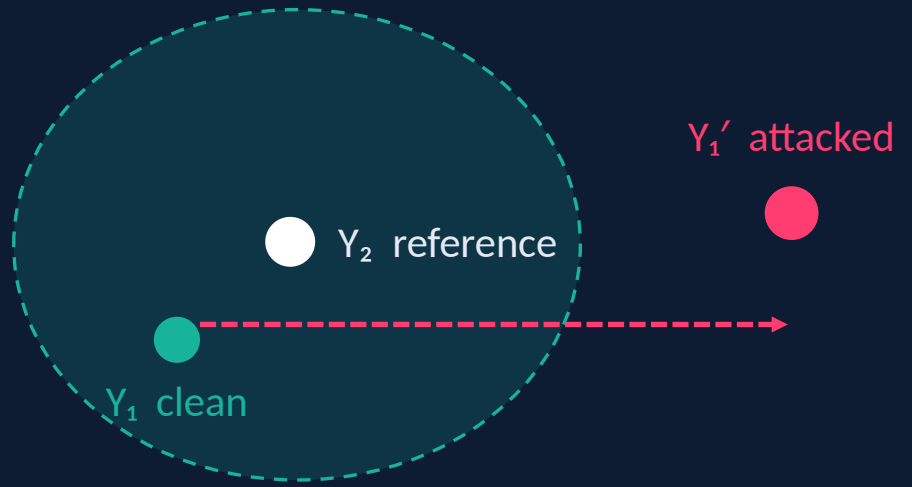


original probe

accessory ROI only

Outside the ROI the image is untouched: $X_1' = X_1$.

Feature space — a large shift



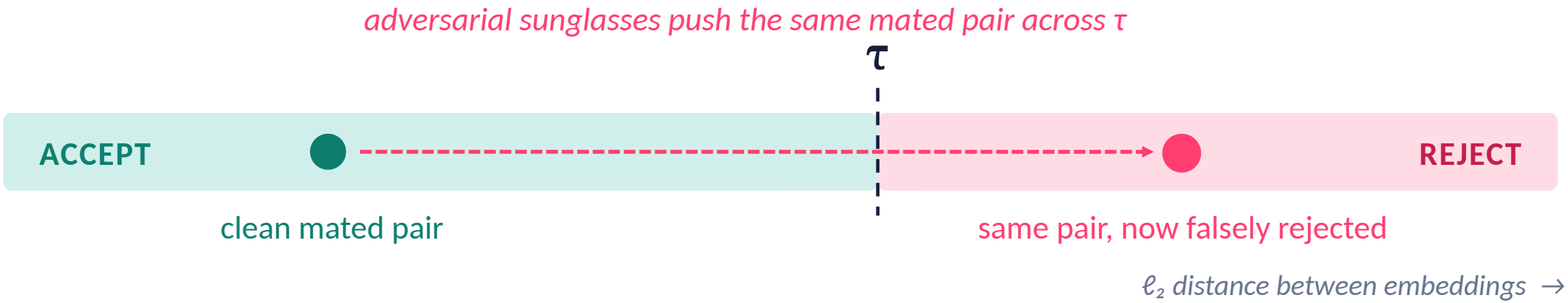
cos_sim 0.79 → 0.29

Same backbone, same face, one accessory-shaped patch.

WEAK POINT · 2

One Scalar Decides Everything

Verification rule: accept the pair if $d(X_1, X_2) = \|F(X_1) - F(X_2)\|_2 \leq \tau$



Fixed, not adaptive

Calibrated once on clean pairs, then frozen for every attack.

LFW $\tau = 1.5070$

AdaFace IR-50, View 2 protocol, 6,000 pairs in 10 folds.

CALFW $\tau = 1.6200$

Cross-age pairs start closer to the boundary.

THREAT MODEL 1

Two Attack Goals

Dodging

untargeted: evade own identity

reference X_2 — probe X_1 — attacked X_1'

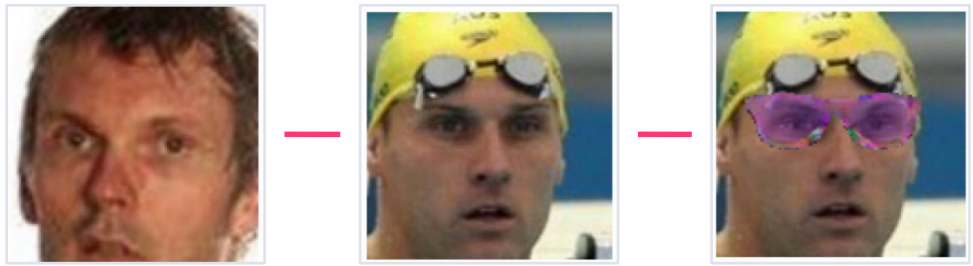
mated pair: same identity

increase distance → false reject

Impersonation

targeted: be accepted as target

target X_2 — attacker X_1 — attacked X_1'



non-mated pair: attacker vs. target

decrease distance → false accept

THREAT MODEL 2

Two Attack Goals, One Objective

Dodging

push apart

untargeted — evade your own identity

PAIR

mated (X_1, X_2 same person)

OPTIMISE

maximise $d(Y_1, Y_2)$

FAILURE INDUCED

false reject ($s = -1$)

Impersonation

pull together

targeted — be taken for someone else

PAIR

non-mated (attacker, target)

OPTIMISE

minimize $d(Y_1, Y_2)$

FAILURE INDUCED

false accept ($s = +1$)

One loss: $\mathscr{L} = s \cdot d(Y_1, Y_2)$ — the sign selects the attack; everything else is shared.

WEAK POINT · 3

A Local Patch Can Dominate the Decision

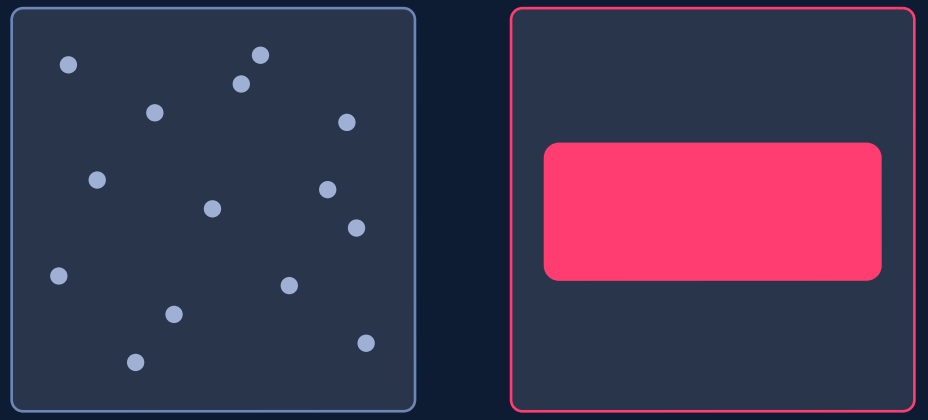
Why patches suit face recognition

1. **Localised, but not small**
   A patch replaces pixels outright rather than perturbing everywhere.

2. **Robust in practice**
   Transform optimisation keeps patches effective under pose and lighting changes.

3. **Plausible as an object**
   An accessory-shaped patch looks like ordinary occlusion.

Two ways to perturb an image



bounded, everywhere — confined region

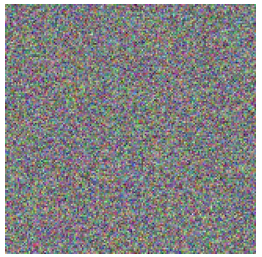
Accessory attacks are the second kind, so region choice is a design variable.

What an Adversarial Example Is

Original image

Face recognition model

✓ recognised correctly



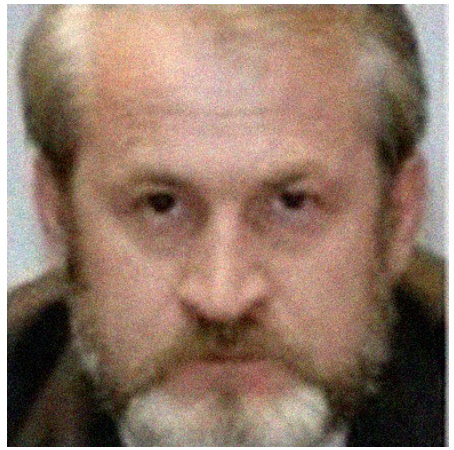
\+ optimised perturbation

gradient-computed, then scaled until invisible

Szegedy et al., 2014 · Goodfellow et al., 2015

Illustration — the perturbation is exaggerated here so it is visible on screen.

Adversarial image



Face recognition model

✗ recognised wrongly

- **The point:** invisible to people, decisive for the model; patches drop invisibility.

From Pixel Perturbations to Physical Attacks

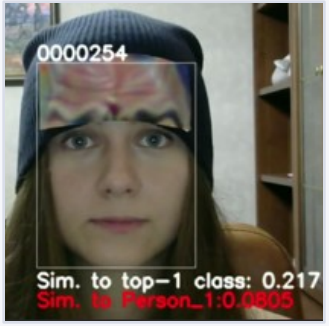
Pixel-level attacks

FGSM, PGD, and DeepFool add small image-space perturbations. They are easy to evaluate digitally, but hard to realise as stable physical objects.

Physical patch / accessory attacks

Adversarial Patch, AdvHat, NatMask, and transformation-robust patches restrict changes to a local object. They must survive lighting, viewpoint, pose, and re-capture.

Adversarial Patch



AdvHat

NatMask

Takeaway: the field moved from invisible pixel noise to visible, wearable attack surfaces.

The Gap: Nobody Varies the Region

Three consequences of a frame-only ROI

- **A thin annulus of pixels**
  The rim is narrow, leaving little editable area.

- **It frames the eyes, not covers them**
  The perturbation sits beside the identity-bearing region, not on it.

- **The mask is never a variable**
  Prior work fixes the shape, then tunes losses, priors and optimisers.

Editable area

- The lens area sits in front of the eyes, but is left untouched.

MAIN IDEA

From Frames to the Sunglasses Region

The reference point — Sharif et al., CCS 2016

- **An accessory as the attack surface**
  Perturbation as a wearable object.
- **Both goals demonstrated**
  Dodging and impersonation.
- **Physically realisable**
  Survives printing, wearing and recapture.

Limitation: rim-only area, classifier-level objective.

Our change: swap the mask

Same generator, objective, blending and thresholds; only the ROI differs.

eyeglass frame only

cos 0.42

frame + lenses

cos 0.29

Same probe, same AdaFace model, same losses.

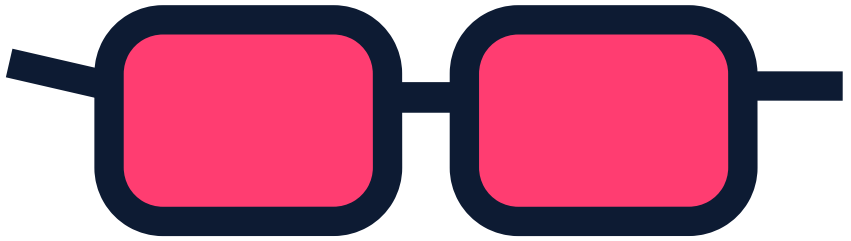
METHOD · 1

Two Masks, One Geometric Template

Eyeglasses ROI — the rim

editable pixels: a narrow band

Sunglasses ROI — rim + lenses



editable pixels: band + both lens areas

Shared template

One canonical geometry M, so alignment and blending are identical.

Different support

The lens area is added — the same optimiser simply gets more room.

Different smoothing

TV is enforced on the rim alone, or on rim and lens separately.

CONTRIBUTIONS

What This Paper Adds

01

Larger perturbation region

ROI covers frame + lenses; we quantify the region effect.

02

Embedding-space attacks

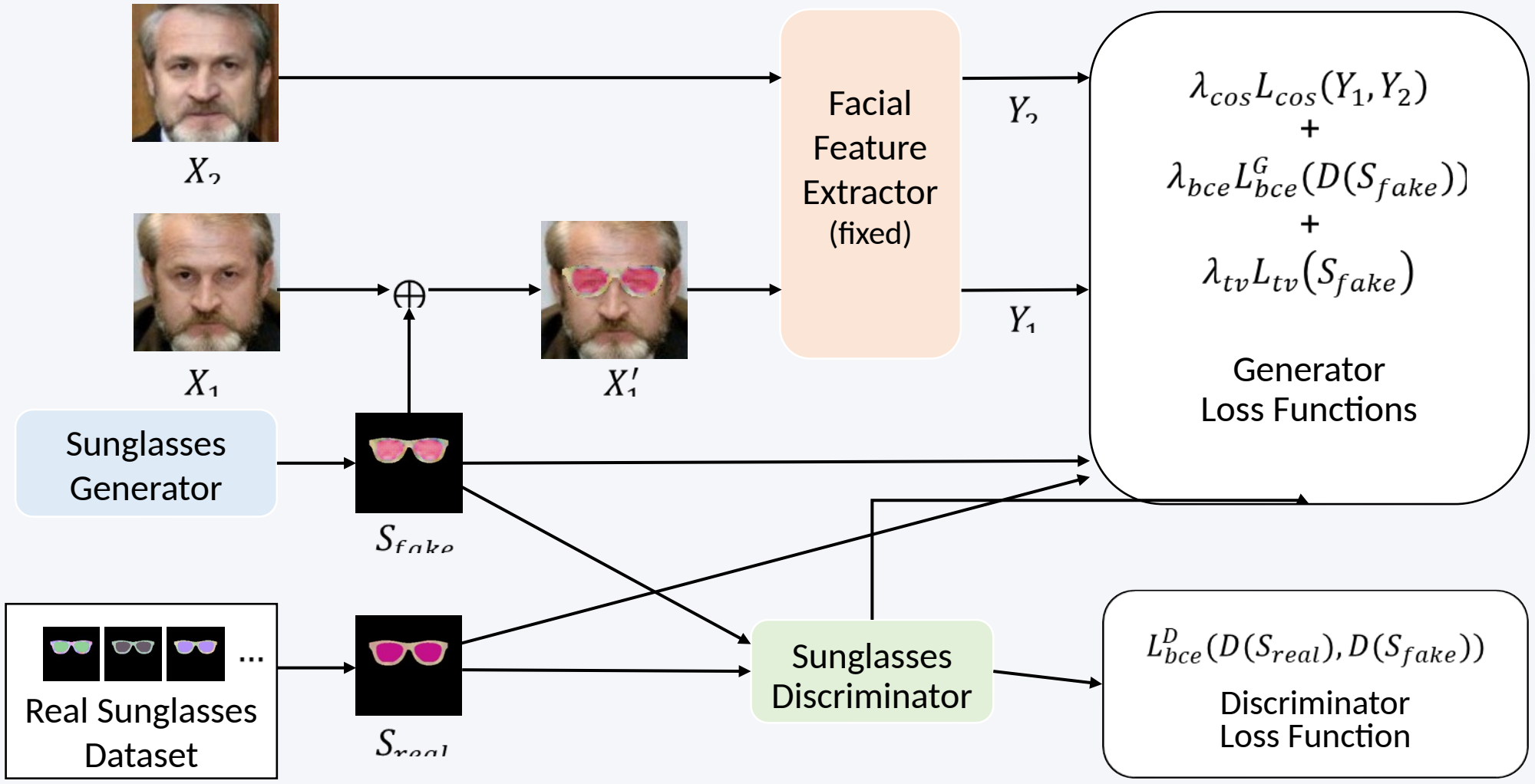
Dodging and impersonation use cosine distance; no classifier head.

03

Benchmark evidence

LFW/CALFW, fixed AdaFace thresholds, ArcFace transfer and lens patterns.

The Generation Pipeline



METHOD · 3

The RGBA Accessory Generator

$z \sim N(0, I)$
latent code, $z \in \mathbb{R}^{100}$

→

G
transposed convolutions, batch norm, ReLU, Tanh

→

$G(z) \in \mathbb{R}^{4x160x160}$
RGB + alpha, masked by M

Four channels

R G B α

First three channels carry colour, the fourth is an alpha matte. The frame is fully opaque; lens transparency $\alpha_\ell \in [0.2, 0.8]$.

Why a small DCGAN is enough

The attack is driven by the feature-space objective and region design, not generator capacity. A DCGAN-scale pair is enough to keep outputs eyewear-like.

METHOD · 4

Alignment and Alpha Blending

$$I_{out} = \alpha \odot I_{glass} + (1 - \alpha) \odot I_{face} \qquad X_1' = X_1 \text{ everywhere outside the ROI}$$

face X_1 aligned RGBA accessory attacked X_1'

Placement

T(· ; L)
Facial landmarks scale and translate the accessory.

One rule everywhere
Same blend for dataset building and attacks.

Differentiable
Gradients flow through the blend into G.

METHOD · 5

The Feature-Space Objective

$$\mathscr{L}_{cos} = s \cdot d(Y_1, Y_2), \qquad d(Y_1, Y_2) = 1 - \cos_{sim}(Y_1, Y_2), \qquad s = -1 \text{ or } +1$$

The reference is fixed

$Y_2 = F(X_2)$ is frozen. Only X_1 is modified.

Cosine matches the decision rule

AdaFace embeddings are ℓ_2-normalised, so distance and cosine similarity are monotone.

No classifier is needed

The loss uses only embeddings, so open-set identities are handled directly.

METHOD · 6

Keeping the Accessory Plausible

$$\mathcal{L}_G = \lambda_{cos}\mathcal{L}_{cos} + \lambda_{bce}\mathcal{L}^G_{bce} + \lambda_{rim}\mathcal{L}_{tv,rim} + \lambda_{lens}\mathcal{L}_{tv,lens}$$

Adversarial realism

D is trained on real accessories; G fools D to stay eyewear-like.

Masked total variation

TV is separate for rim and lens and normalised by region size.

Hinge thresholds

Only excess above τrim and τlens is charged. We use TVrim ≤ 0.4 and TVlens ≤ 0.1.

Datasets, Recogniser and Protocol

LFW

τ = 1.5070

13,233 images · 5,749 identities

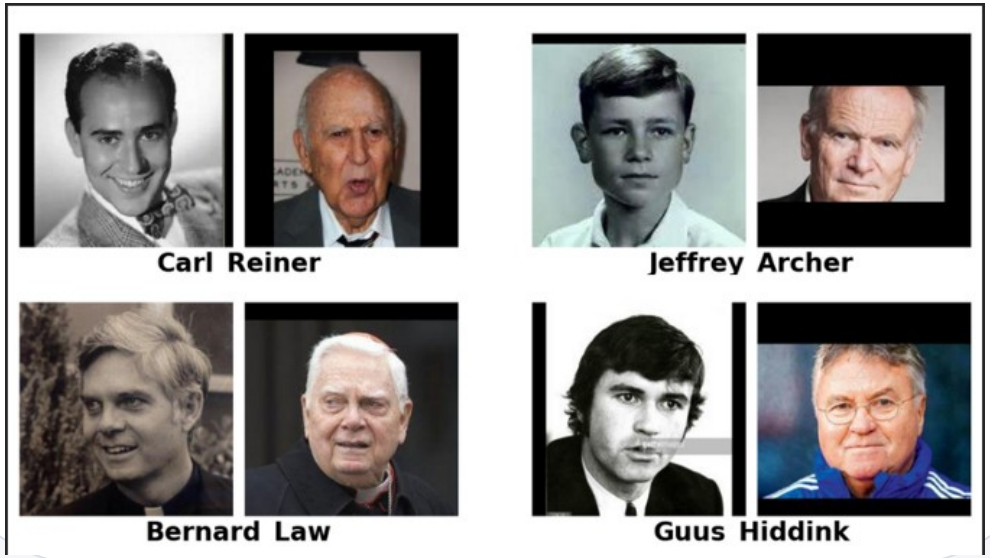
View 2 protocol · 6,000 pairs · 3,000 mated / 3,000 non-mated

CALFW

τ = 1.6200

12,174 images · 4,025 identities

Cross-age LFW · 6,000 pairs · age-gap positives, matched negatives



AdaFace IR-50 · 112×112 input · X_2 kept clean · thresholds fixed after calibration · metrics: accuracy + CSM

EXPERIMENTS · 2

Seven Settings, One Recogniser

Clean (AdaFace) — original aligned faces, no overlay.

Eyeglasses ROI

$LR\_G = 3 \times 10^{-4}$

Gallery-sampled overlay
real eyeglass overlay, non-adversarial control.

Ours, TV-constrained
adversarial rim, TVrim ≤ 0.4.

Ours, unconstrained
same optimisation, no smoothness bound.

Sunglasses ROI

$LR\_G = 9 \times 10^{-4}$

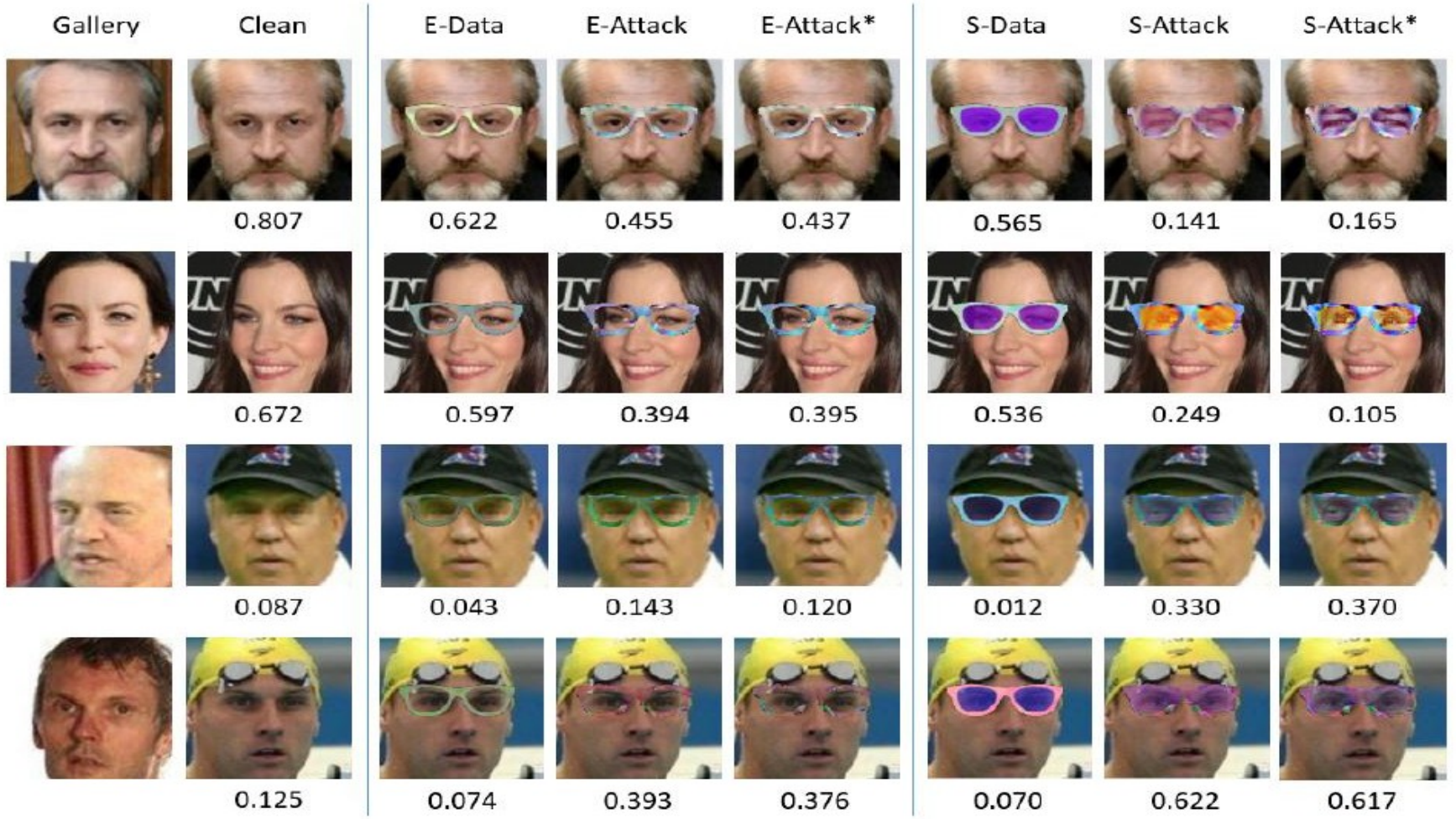
Gallery-sampled overlay
real sunglasses overlay, non-adversarial control.

Ours, TV-constrained
rim + lens, TVrim ≤ 0.4 and TVlens ≤ 0.1.

Ours, unconstrained
same optimisation, no smoothness bounds.

RESULTS · 0

What the Attacks Look Like



- **Read**

Gallery → Clean → E → S

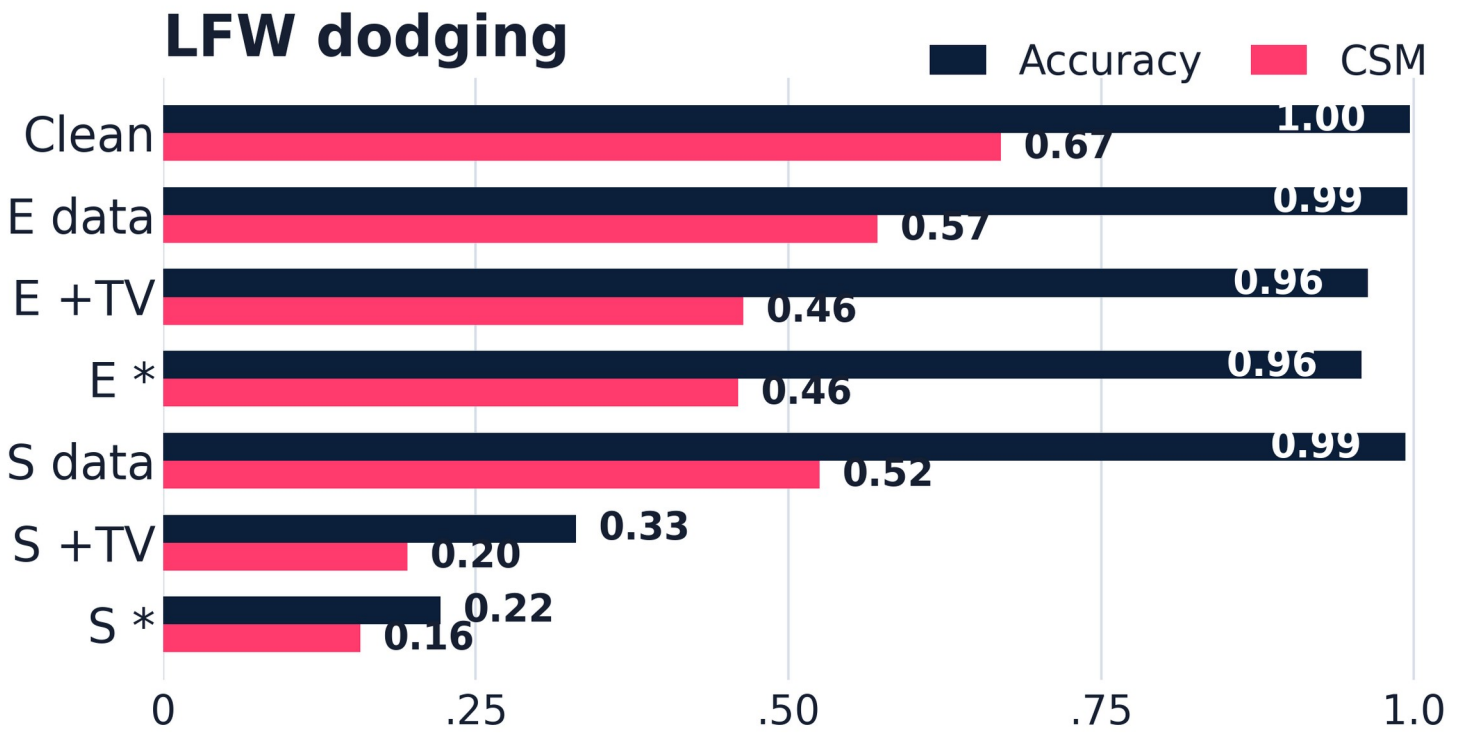
Number = CSM

★ unconstrained

S attacks shift most

RESULTS · 1

Dodging on LFW



E = eyeglasses, S = sunglasses, * = unconstrained

Sunglasses + TV

33.02%

clean baseline: 99.67%

Eyeglasses + TV

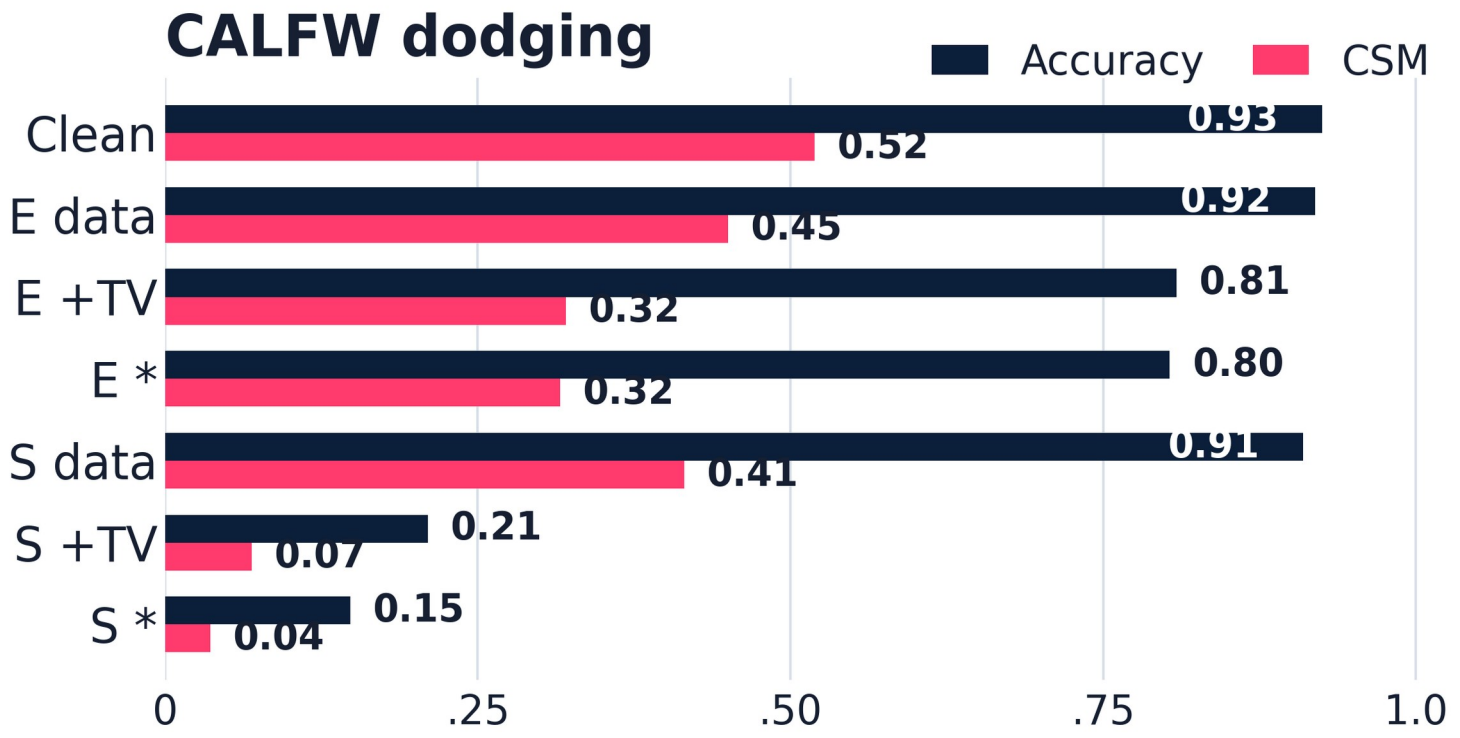
96.33%

same optimiser, rim only

Same pipeline: 3-point loss with frame, 67-point loss with frame + lenses.

RESULTS · 2

Dodging on CALFW



E = eyeglasses, S = sunglasses, * = unconstrained

Sunglasses + TV

21.03%

clean baseline: 92.50%

Eyeglasses + TV

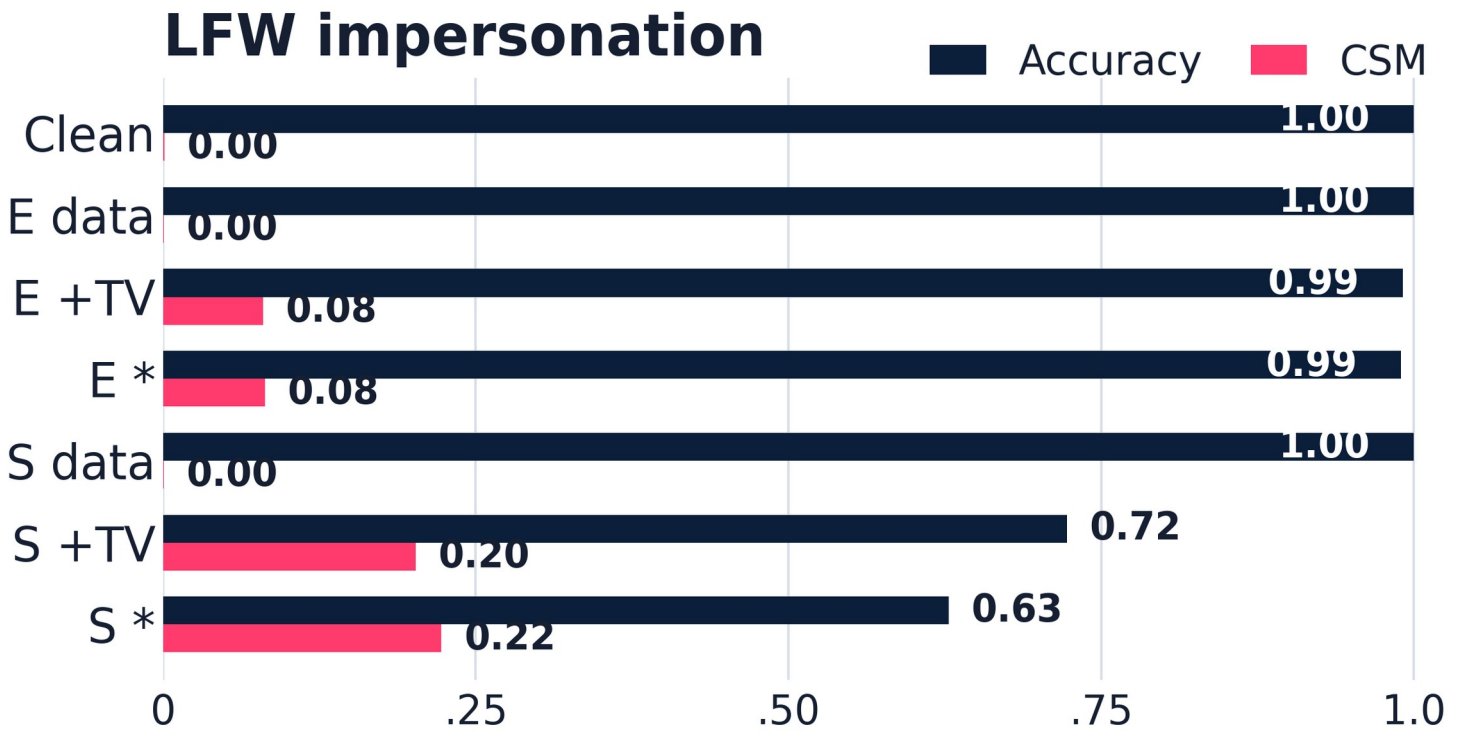
80.87%

same optimiser, rim only

Cross-age pairs sit closer to the boundary, so more cross it.

RESULTS · 3

Impersonation on LFW



Higher CSM means stronger impersonation

Sunglasses + TV

72.27%

clean baseline: 100.00%

Eyeglasses + TV

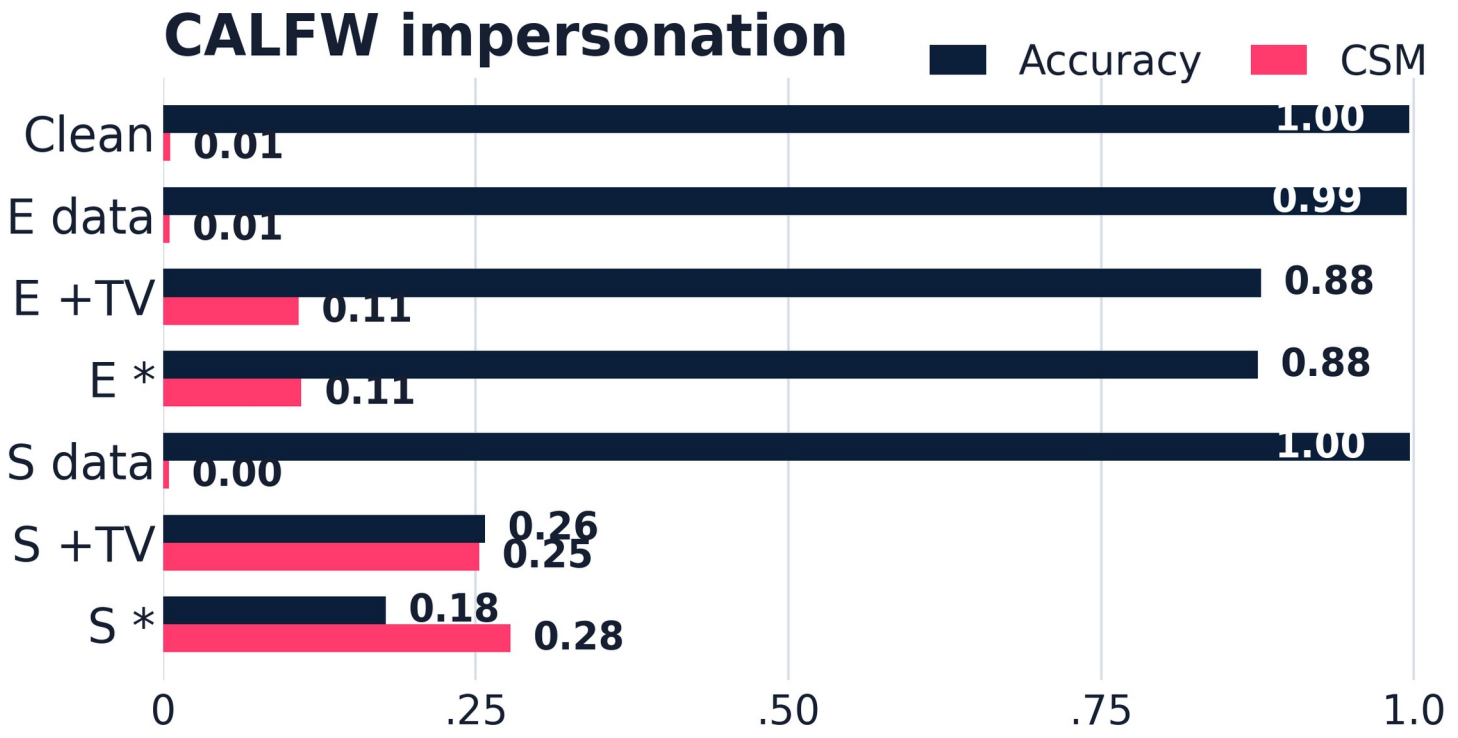
99.10%

barely moves the decision

Targeted attacks are harder: frame gains little; lens gains 28 points.

RESULTS · 4

Impersonation on CALFW



Higher CSM means stronger impersonation

Sunglasses + TV

25.73%

clean baseline: 99.63%

Eyeglasses + TV

87.77%

same optimiser, rim only

Largest region effect: three quarters of impostor pairs are accepted.

RESULTS · 5

Structure Appears Inside the Lens

Sunglasses attack — lens region

Structured, eye-like texture fills the lens.

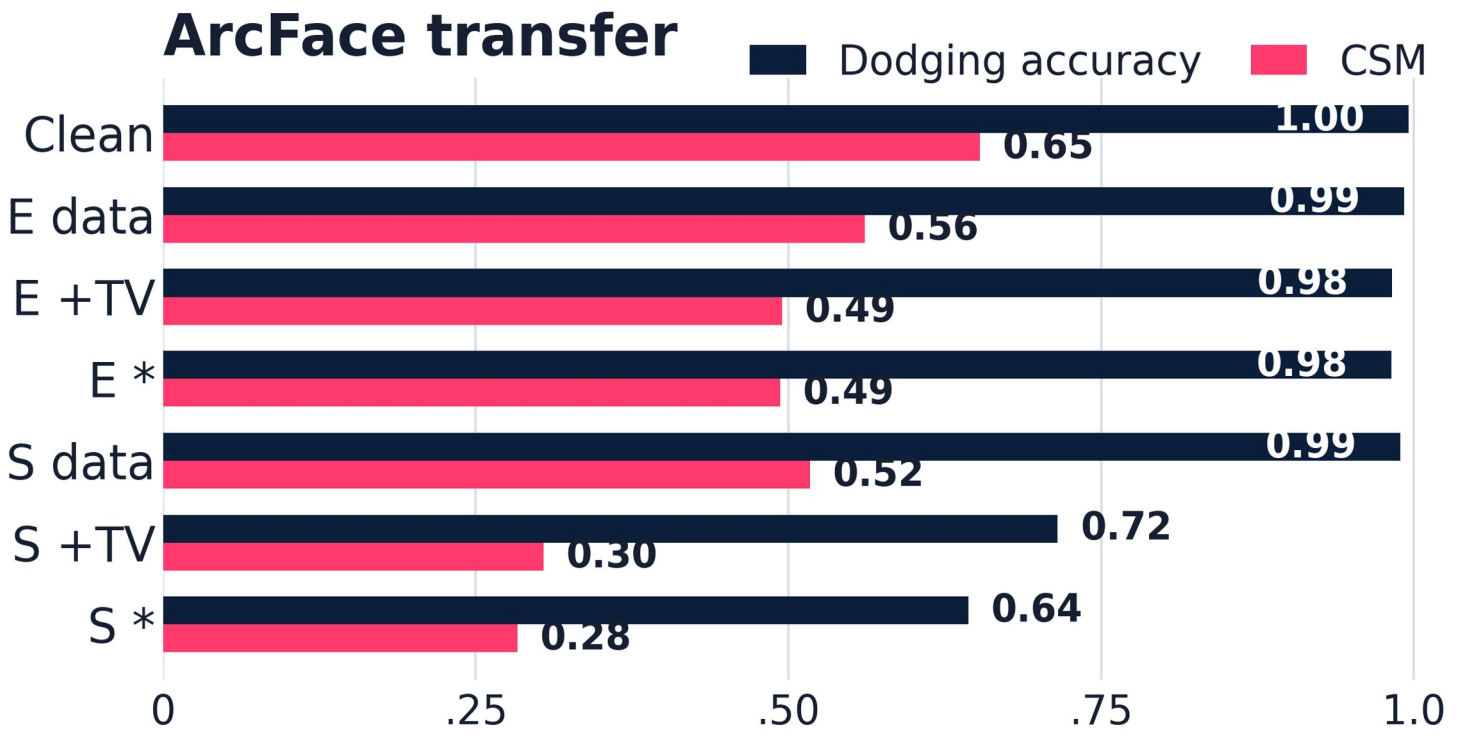
Eyeglasses attack — rim region

High-frequency colour noise around the frame.

- **Labelled honestly:** a qualitative observation, not a measured effect. It suggests the lens interacts with identity cues — quantifying that is future work.

RESULTS · 6

Transfer: AdaFace → ArcFace



Generated on AdaFace, evaluated on ArcFace

Sunglasses, transferred

71.53%

ArcFace dodging, τ = 1.5140 (clean 99.60%)

Eyeglasses, transferred

98.23%

frame-only perturbations barely transfer

Impersonation transfers weakly: CSM 0.037 → 0.102.

CONCLUSION

Where the Perturbation Acts

- **Region beats objective**
  Frame + lenses gains tens of accuracy points under the same loss, generator and protocol.
- **Both goals, and transfer**
  Stronger on LFW/CALFW; the only variant that transfers to ArcFace.
- **Not occlusion**
  Real accessories cost at most 1.5 points. Optimisation does the work.