Robustness and Generalization of Lightweight Face-Verification Models: A Reproducible Cross-Dataset Evaluation Framework for Resource-Constrained Deployment

MH Rahman, HafizLab (corresponding author)

Manuscript status: protocol with measured results for three lightweight models and one reference on three benchmarks. No target-domain data have been evaluated; items still to be completed are marked in square brackets.

Abstract

Lightweight face-verification models suit deployments with limited compute, including CPU-only servers, but performance on one benchmark may not transfer to other poses, ages, resolutions, compression levels or capture conditions. This paper specifies a reproducible framework for measuring robustness and generalization of lightweight face-verification models across datasets and capture conditions, with within-dataset, fixed-threshold cross-dataset, target-recalibrated, controlled-degradation and CPU-efficiency tracks, identity-aware bootstrap intervals, and an open toolkit implementing every metric and degradation. In an executed study, three lightweight models (MobileFaceNet-w600k, SFace and a Qualcomm MobileFaceNet) and an ArcFace-R50 reference were evaluated on LFW, CFP-FP and AgeDB-30 and on a 22-condition degradation grid on LFW. The lightweight models had overlapping clean LFW EER (0.43%, 0.57% and 0.67%; reference 0.30%) but differed from 5.51% to 14.60% on CFP-FP and from 3.70% to 5.70% on AgeDB-30; the paired difference between the Qualcomm model and MobileFaceNet-w600k on CFP-FP was -9.1 points (A − C). A threshold calibrated on LFW at FMR = 1% gave FMR above 1% on AgeDB-30 for all four models (1.83%–3.23%); recalibration restored FMR at the cost of higher FNMR and needed about 1,000 target pairs. Under strong blur, 20-pixel resolution and a lower-face band the models diverged by several-fold, and under the largest band a frozen clean threshold gave FMR of 32.53%, 4.20%, 93.13% and 18.73% (A, B, C, reference). Single-thread CPU latency was 13.7, 24.9, 11.3 and 287.6 ms per image; speed did not predict robustness. CFP-FP and AgeDB-30 came from an unofficial redistribution with approximated identities, so those results are provisional pending replication on the official releases. The framework helps practitioners identify where a lightweight model remains reliable, where thresholds become unstable, and when recapture or human review is preferable to automatic decisions.

Keywords: face verification; lightweight deep learning; robustness; generalization; cross-dataset evaluation; domain shift; threshold calibration; CPU inference

1. Introduction

Face verification compares two facial samples and estimates whether they belong to the same person. Deep embedding models make this practical at scale, and lightweight architectures reduce latency and resource consumption. Yet computational efficiency does not guarantee robustness: a model can perform well on a familiar benchmark and degrade when the capture device, image-processing pipeline, pose, illumination, resolution, or compression changes.

A deployment threshold also affects generalization. A threshold calibrated on one dataset may not achieve the same FMR or FNMR on another. When a system supports admission, account access, or another consequential workflow, a shift in error rates creates both security risk (false accepts) and usability cost (false rejects). Robustness evaluation must therefore examine not only average verification accuracy but also operating-point stability, condition-specific errors, failure-to-process rates, and computational cost.

This paper proposes a protocol for assessing lightweight face-verification models across datasets and capture conditions. It emphasizes controlled comparisons, strict separation of calibration and test data,

transparent reporting, and avoidance of unsupported claims. The protocol is released together with executable evaluation code so that any group can populate the result tables with identical procedures.

1.1 Research objectives

- Compare lightweight face-verification models across multiple public benchmarks using their official protocols.
- Measure performance degradation under controlled changes in image quality and capture conditions.
- Quantify how a fixed threshold calibrated on one dataset transfers to other datasets and conditions.
- Evaluate the relationship between robustness, failure-to-process rate, and CPU inference cost.
- Identify evaluation practices that reduce data leakage and improve reproducibility.

1.2 Research questions

- RQ1: How much do verification error rates vary across datasets for the same lightweight model?
- RQ2: Which capture degradations — blur, reduced resolution, illumination change, compression, occlusion — cause the largest measured performance changes?
- RQ3: How does a threshold calibrated on a source dataset perform on a different target dataset without recalibration?
- RQ4: Does target-domain threshold recalibration improve operating-point performance, and how much target data does it require?
- RQ5: What robustness–efficiency trade-offs appear when accuracy, failure rate, latency and memory are evaluated together?

1.3 Pre-specified hypotheses

To prevent post-hoc interpretation, the following directional hypotheses are fixed before any test-set evaluation. They are hypotheses, not findings, and will be reported as supported or not supported according to the confidence intervals defined in Section 4.5.

- H1: For every model, fixed-threshold FMR on at least one non-source dataset will differ from the calibration target FMR by more than the 95% bootstrap interval of the source estimate (threshold non-transfer).
- H2: Target-calibrated thresholds will reduce the absolute deviation of FMR from the declared target relative to the source-calibrated frozen threshold.
- H3: Low-resolution and strong-blur conditions will produce larger relative EER degradation than JPEG compression at the mild levels in the grid.
- H4: Model ranking by within-dataset EER will not be identical to ranking by worst-condition EER (rank instability).

1.4 Contributions

- A cross-dataset evaluation protocol for lightweight face-verification models.
- A controlled, parameterized degradation grid for testing sensitivity to capture conditions.
- A threshold-transfer analysis that separates fixed-threshold generalization from target-domain recalibration, with a defined transfer-gap metric.
- An identity-aware resampling and paired-comparison procedure for uncertainty reporting.

- A reporting template and open evaluation code for verification errors, uncertainty, failure rates and CPU resource use.

2. Related Work and Background

Modern face verification extracts a feature embedding from each image and compares embeddings with cosine similarity or another declared distance. FaceNet [1] introduced a learned embedding trained with a triplet objective, and ArcFace [2] introduced an additive angular-margin loss that has influenced many subsequent systems. MobileFaceNets [3] adapted compact convolutional backbones for real-time verification on mobile hardware. Lightweight variants aim to reduce inference cost, but comparisons are meaningful only when preprocessing, alignment, image resolution and threshold protocols are controlled.

LFW [4] is a widely used unconstrained verification benchmark. CFP-FP [5] emphasizes frontal-to-profile verification, and AgeDB-30 [6] includes age variation. These benchmarks represent different variation types but do not cover every camera, demographic or operational domain, and results across datasets must not be compared as if they used identical protocols. Several of these benchmarks are close to saturation for large models, which motivates examining operating-point behavior and degradation sensitivity rather than headline accuracy alone.

Robustness can be studied through naturally occurring dataset variation, controlled image transformations, or evaluation on a separate target domain. Synthetic degradations isolate sensitivity but are not equivalent to real capture changes and should complement, not replace, cross-dataset testing. Threshold transfer is particularly important because score distributions can shift between datasets even when the embedding model is unchanged. Biometric testing practice, including the evaluation methodology tradition of FERET [7], the ISO/IEC 19795 performance-testing series [8], and NIST face-recognition evaluations [9], stresses declared protocols, error-rate trade-off reporting, and counting failures to acquire or enroll; the present protocol adopts these principles for lightweight embedding models.

3. Study Design

3.1 Candidate models

Select at least three lightweight or deployment-relevant embedding models, subject to available checkpoints, licensing, and reproducible inference support. Examples include a MobileFaceNet-family model [3] and compact variants from established open face-recognition implementations. A larger reference model may be included as a non-lightweight comparator if compute and licensing permit. For every model record the exact architecture, checkpoint hash (SHA-256), input resolution, embedding dimension, normalization, alignment pipeline, inference runtime and version, and licence. In the executed study the models are: Model A, MobileFaceNet-w600k (InsightFace w600k_mbf.onnx; SHA-256 9cc6e4a7…19eb4f; 512-d; mean = std = 127.5); Model B, SFace (OpenCV Zoo face_recognition_sface_2021dec.onnx; SHA-256 0ba9fbfa…4e79; 128-d; mean 0, std 1, RGB input); Model C, MobileFaceNet-QC (Qualcomm AI Hub MobileFaceNet, ONNX float; graph SHA-256 d0e8723a…cc36; weights in a separate 3.76 MiB external data file, 4.13 MiB in total; 128-d; mean 0, std 255, RGB; two-input graph embedding both images of a pair per call); and the reference comparator ArcFace-R50 (InsightFace w600k_r50.onnx; SHA-256 4c06341c…9e43; mean = std = 127.5). All models use 112×112 input. Models A and R50 are named w600k, understood to indicate WebFace600K training; the model card for Model C states MS-Celeb-1M training with an ArcFace loss and an Apache-2.0 licence; SFace is distributed in OpenCV Zoo under Apache-2.0 as a MobileFaceNet trained with the SFace loss (Zhong et al., arXiv:2205.12010), but neither the OpenCV Zoo page nor the authors' code repository states which dataset trained the released file (the SFace paper reports experiments on CASIA-WebFace, VGGFace2 and MS1MV2), so its training data are treated as unknown; the InsightFace model zoo states that all its models are for non-commercial research use only. Licence terms were read from the cited pages on 27 October 2026 and must be re-checked before release. Models are

used as released; if any model is fine-tuned, the training data must be identity-disjoint from all evaluation sets, and this must be stated.

Table 1. Model slots and required metadata

Model slot	Role	Details to record
Model A	Primary lightweight candidate	Architecture, SHA-256, version, input size, licence
Model B	Alternative lightweight candidate	Architecture, SHA-256, version, input size, licence
Model C	Alternative lightweight candidate	Architecture, SHA-256, version, input size, licence
Reference model (optional)	Contextual performance baseline	Architecture, size, runtime, licence

3.2 Datasets and evaluation protocols

Table 2. Datasets

Dataset	Primary variation of interest	Evaluation notes
LFW	Unconstrained image variation	Official verification protocol and 10-fold pair list
CFP-FP	Frontal-to-profile pose variation	Official protocol and declared preprocessing
AgeDB-30	Age-related variation (30-year gap)	Official protocol and declared pair list
Target-domain dataset (optional)	Local camera, capture instructions, compression, operational conditions	Lawful basis, identity-disjoint partitions, restricted access, documented consent/governance

Each benchmark is used according to its official protocol and licence. Identities and pair lists are not merged across datasets in a way that leaks identity information between calibration and test partitions; where identities may overlap across public datasets, overlap is checked and overlapping identities are removed from the target side. Public benchmarks support reproducible component evaluation but do not establish performance for a particular institution or population. A target-domain dataset must be collected and governed separately, with data minimization and retention controls.

Data sources in the first executed run: LFW images were used with their official pair list and identity labels. CFP-FP and AgeDB-30 pair data were obtained from a third-party Hugging Face redistribution (cat-claws/face-verification) in Parquet format with image pairs and same/different labels but without identity labels; its provenance and licence statement are not given on the dataset card, so results on these two benchmarks should be confirmed against the official releases; the terms of use of the official releases are given under Data availability.

3.3 Evaluation tracks

Table 3. Evaluation tracks

Track	Procedure	Question addressed
T1 Within-dataset	Evaluate under the benchmark protocol; calibrate thresholds only on permitted development folds (cross-validated folds for LFW-style protocols)	Baseline performance on each benchmark
T2 Cross-dataset fixed-threshold	Calibrate on source development data to a declared FMR; freeze the threshold; test on a different dataset	Threshold transfer and domain shift (RQ1, RQ3)
T3 Cross-dataset recalibrated	Use a separate target calibration partition (varying n), then evaluate on disjoint target test identities	Benefit and data cost of recalibration (RQ4)
T4 Controlled degradation	Apply predeclared transformations to test images, preserving pair labels	Sensitivity to image-quality changes (RQ2)
T5 CPU efficiency	Run identical preprocessing and inference configuration on documented hardware	Latency, throughput, memory, trade-off (RQ5)

4. Methodology

4.1 Preprocessing and pair scoring

Model-specific normalization was selected where the released models do not document it unambiguously: for Models B and C a probe scored eight fixed variants (mean/std of 127.5/127.5, 127.5/128, 0/255, 0/1, each with RGB or BGR channel order) on 600 pairs from LFW fold 1, and the variant with the lowest EER was used for all datasets. For SFace the probe tied between RGB and BGR (EER 0.67% for both), RGB was used, and a full-data BGR check, which was worse on all three benchmarks, is reported in Section 7. Because the probe uses a subset of the test data of one benchmark, probe selection could slightly favour that model on LFW; models A and R50 used their documented normalization with no selection.

Apply the same face detection, alignment, crop, colour conversion, resizing and normalization procedure to all models where their documented input requirements allow. If model-specific preprocessing is necessary, record it and treat it as part of the model system. For each pair (a, b), extract embeddings e_a, e_b and compute the cosine similarity

$$s(a, b) = \langle e_a, e_b \rangle / (\|e_a\| \, \|e_b\|), \quad \text{accept if } s \geq \tau.$$

Larger scores indicate stronger matches. Face-detection failures, invalid images and embedding failures are recorded rather than silently dropped: a pair with any failed image receives a missing score, is counted in the failure-to-process rate, and is excluded from FMR/FNMR/EER computed over processable pairs. Accuracy among processable pairs is always reported alongside the failure rate, and a failure-inclusive sensitivity analysis (counting failed genuine pairs as false non-matches and failed impostor pairs as non-matches) is provided so that a model is not rewarded for failing on difficult images.

4.2 Controlled capture-condition transformations

A predeclared transformation grid evaluates sensitivity (Table 4). Transformations are applied to both images of each pair in the main analysis and to one image only in a secondary analysis (to mimic enrolment/probe quality mismatch). Pose robustness is studied with genuine frontal/profile data (CFP-FP) rather than geometric warping. Each transformation is documented with parameter values and software versions. Random components use a fixed seed, and the grid is frozen before final test evaluation.

Table 4. Predeclared degradation grid (for 112×112 aligned inputs)

Condition	Manipulation	Levels	Reporting requirement
Blur	Gaussian blur, σ in pixels	0.5, 1, 2, 3	σ, library version, affected image(s)
Resolution	Downsample to s×s then bilinear resize back	56, 40, 28, 20	Original and intermediate dimensions, interpolation
Illumination (brightness)	Additive shift on [0,255], clipped	−60, −30, +30, +60	Equation, clipping rule
Illumination (contrast)	$x' = (x - \mu)\cdot c + \mu$, clipped	c = 0.5, 0.75, 1.5	Equation, clipping rule
Compression	JPEG re-encoding	Quality 90, 50, 30, 10	Encoder, version, quality
Occlusion	Black band over lower face (mask-like)	Band height 20%, 40%, 60% of image height	Mask shape, area, location
Pose	Natural variation via CFP-FP	Frontal–profile	Dataset protocol, pose grouping

4.3 Threshold calibration and transfer

The operating objective is declared before testing as a target FMR α (primary: $\alpha = 10^{-2}$, with $\alpha = 10^{-3}$ reported where impostor-pair counts support it). The calibration threshold is the smallest score $\hat{\tau}(\alpha)$ whose empirical impostor acceptance rate on the calibration set does not exceed α:

$$\hat{\tau}(\alpha) = \min \{ \tau : FMR_{cal}(\tau) \leq \alpha \}.$$

In T2, $\hat{\tau}$ is estimated on source development data, frozen, and applied unchanged to the target test set. In T3, $\hat{\tau}$ is re-estimated on a separate target calibration partition of n pairs (n varied over a declared ladder, e.g. 500, 1000, 5000, 10000 pairs where available), and evaluated on disjoint target test identities. Identities and, where feasible, sessions are disjoint between calibration and test. The threshold-transfer gap is

$$\Delta FMR = FMR_{target}(\hat{\tau}_{src}) - FMR_{target}(\hat{\tau}_{tgt}), \quad \Delta FNMR = FNMR_{target}(\hat{\tau}_{src}) - FNMR_{target}(\hat{\tau}_{tgt}).$$

The deviation of observed FMR from the declared α is also reported. A threshold that transfers poorly is evidence of score-distribution shift under the tested conditions, not proof of a specific causal mechanism. If a target set is too small to estimate the operating point reliably, uncertainty is reported and strong claims are avoided. Because the empirical FMR resolution is limited by the number of impostor pairs, target FMRs below roughly $10/N_{imp}$ are reported only as upper bounds.

In the first executed run, the transfer analysis was performed within LFW as a condition shift: the threshold was calibrated on the first five folds of clean images, frozen, and evaluated on the last five folds under each degradation; the recalibrated threshold used the first five folds of the same degradation. Cross-dataset transfer uses the same procedure with a different source and target dataset.

4.4 Metrics

Table 5. Metrics

Metric	Definition / use
FMR(τ)	Proportion of impostor comparisons with $s \geq \tau$
FNMR(τ)	Proportion of genuine comparisons with $s < \tau$
EER	Operating point where estimated FMR = FNMR; reported as the mean of the two at the closest threshold on the empirical curve
TAR at FMR = α	1 − FNMR at the threshold calibrated to FMR = α, when sample size supports it
Failure-to-process rate	Proportion of pairs in which detection, alignment or embedding extraction fails for either image
Relative degradation	(M_cond − M_clean) / M_clean for error metric M; positive = worse; undefined if M_clean = 0 (then absolute change is reported)
Threshold-transfer gap	Δ_FMR and Δ_FNMR as defined in Section 4.3
Latency	Median, p95, p99 of embedding-only and end-to-end image-to-score time; timing boundary stated
Peak memory, throughput, model size	Peak resident set size, pairs or images per second, and parameter/file size under a documented configuration

4.5 Statistical analysis

The number of identities, images, genuine pairs and impostor pairs is reported for every split and condition. Pair-level observations are not independent because images and identities recur across comparisons. Confidence intervals therefore use an identity-level (cluster) bootstrap: identities are resampled with replacement (B = 1000 or more), a pair is retained, with multiplicity, when both of its identities are drawn, and the metric is recomputed; percentile 95% intervals are reported. Models are compared on exactly the same pairs and the same resamples, and the interval of the paired difference of a metric (model A − model B) is reported; differences whose interval excludes zero are described as distinguishable under this protocol. Where p-values are given, multiple comparisons across models and conditions are controlled with the Holm–Bonferroni procedure, and the family of tests is declared in advance. Condition-specific results are reported with counts and intervals rather than ranked from point estimates alone. Where identity labels are unavailable (CFP-FP and AgeDB-30 in the first executed run), identities are approximated by grouping images connected through genuine pairs; this produces more groups than true identities and therefore captures dependence only partially, and a fold-level bootstrap is used as a sensitivity check. Demographic subgroup claims are made only if labels were lawfully and appropriately obtained, subgroup sizes are adequate, and the analysis is justified in advance.

For the four-model comparison, paired differences in EER were computed on identical identity resamples for all six model pairs in 25 settings (clean LFW, 22 LFW degradations, CFP-FP and AgeDB-30). With B = 1000 the smallest attainable bootstrap p-value is about 0.002, which cannot survive Holm correction over 150 tests; two-sided p-values were therefore obtained from a normal approximation to the bootstrap distribution (standard error = width of the 95% percentile interval ÷ 3.92) and Holm-adjusted over all 150 tests. Percentile intervals reported alongside are unadjusted.

4.6 CPU efficiency protocol

Record CPU model, physical/logical cores, RAM, operating system, inference runtime and version, thread settings (primary: single thread, secondary: all cores), input size, batch size (primary: 1), and whether face detection is in the timed pipeline. Run at least 20 warm-up iterations and at least 200 timed iterations under controlled background load, repeated over three independent process launches. Report median and tail latency, throughput, peak resident memory and model size, for both embedding-only and end-to-end image-to-score pipelines where feasible.

4.7 Implementation and reproducibility package

All procedures in Sections 4.1–4.6 are implemented in an open Python toolkit (module fveval) providing cosine pair scoring with explicit failure handling; FMR, FNMR, EER and TAR@FMR; calibration and transfer-gap computation; the identity-level bootstrap and paired-difference intervals; Holm correction; the Table 4 degradation grid; ONNX-Runtime CPU embedding and benchmarking; and Markdown table generation. The toolkit was first verified on synthetic score distributions with known analytic error rates (for two Gaussians with means 0.6 and 0.1 and common standard deviation 0.15, the expected EER is $\Phi(-1.67) \approx 4.8\%$, and the implementation returned 4.6% on 4,000 pairs per class) to confirm correctness of the metric code; these synthetic checks are software tests and not experimental results. Random seeds, configuration files, checkpoint hashes and the exact pair lists are archived with each run.

5. Experimental Results

Scope of this version. MFN = MobileFaceNet; ref. = non-lightweight reference. Results come from executed runs of the accompanying toolkit (B = 1000 identity-level bootstrap resamples, seed 0, target FMR α = 1%) on three benchmarks with four models: three lightweight candidates — MobileFaceNet-w600k (A; InsightFace w600k_mbf.onnx), SFace (B; OpenCV Zoo face_recognition_sface_2021dec.onnx) and MobileFaceNet-QC (C; Qualcomm AI Hub MobileFaceNet, ONNX float) — and one non-lightweight reference, ArcFace-R50 (InsightFace w600k_r50.onnx). LFW was evaluated on its official 10-fold protocol (6,000 pairs; 7,701 images; 4,281 identities) with identity labels from the benchmark. CFP-FP (7,000 pairs) and AgeDB-30 (6,000 pairs) were taken from a third-party redistribution (Hugging Face dataset cat-claws/face-verification, Parquet format; not an official release) without identity labels; identities were approximated by grouping images linked through genuine pairs, which gives many more groups (2,448 for CFP-FP and 2,338 for AgeDB-30) than the benchmarks contain, so identity-level intervals for these two datasets are approximate (a fold-level bootstrap gave near-identical intervals for Models A and R50, see Section 5.1). No target-domain data have been run. The condition-shift analysis in Section 5.6 is within LFW; cross-dataset transfer is reported in Sections 5.3 and 5.4. All values are reproduced from archived output files; no value has been estimated or interpolated.

5.1 Within-dataset performance across benchmarks (T1)

Table 6. EER (95% bootstrap CI) and 10-fold accuracy (mean ± SD, %) on the three benchmarks. Failure rate was 0 in every cell (images were supplied pre-aligned). CFP-FP and AgeDB-30 intervals are approximate (identities approximated).

Model	LFW EER (CI)	CFP-FP EER (CI)	AgeDB-30 EER (CI)	Accuracy LFW / CFP-FP / AgeDB-30
MFN-w600k (A)	0.43% (0.22–0.68)	5.51% (4.60–6.61)	3.70% (3.12–4.38)	99.57 ± 0.29 / 95.79 ± 1.24 / 96.62 ± 0.85
SFace (B)	0.57% (0.30–0.87)	5.80% (4.98–6.68)	5.57% (4.77–6.31)	99.47 ± 0.36 / 94.67 ± 1.27 / 94.92 ± 1.22
MFN-QC (C)	0.67% (0.41–0.95)	14.60% (13.44–15.85)	5.70% (4.93–6.50)	99.38 ± 0.28 / 86.79 ± 1.94 / 94.20 ± 1.29
ArcFace-R50 (ref.)	0.30% (0.09–0.48)	4.14% (3.31–5.08)	2.30% (1.80–2.89)	99.83 ± 0.17 / 97.31 ± 1.13 / 98.08 ± 0.68

For every model, EER on CFP-FP and AgeDB-30 is far higher than on LFW (RQ1): the lightweight models' EER rises from 0.43%, 0.57% and 0.67% on LFW to 5.51%, 5.80% and 14.60% on CFP-FP and 3.70%, 5.57% and 5.70% on AgeDB-30 (models A, B, C in order); the reference rises from 0.30% to 4.14% and 2.30%. Clean LFW does not predict the size of this drop. On LFW the four models span 0.30%–0.67% (a 2.2-fold range, with overlapping intervals for the three lightweight models), whereas on CFP-FP they span 4.14%–14.60%; in particular MobileFaceNet-QC has a clean LFW EER (0.67%) close to the other lightweight models but a CFP-FP EER of 14.60% and accuracy of 86.79 ± 1.94%, i.e. a frontal-to-profile failure that LFW does not reveal. On CFP-FP, the EER difference between Model A and MobileFaceNet-QC (A − C) is -9.09 points (95% CI -10.18 to -7.94; Holm $p < 0.001$), i.e. MobileFaceNet-QC is higher. Rankings by clean LFW EER, CFP-FP EER, AgeDB-30 EER and worst LFW-degradation EER are: R50 < MFN-w600k < SFace < MFN-QC; R50 < MFN-w600k < SFace < MFN-QC; R50 < MFN-w600k < SFace < MFN-QC; R50 < SFace < MFN-w600k < MFN-QC. A fold-level bootstrap (resampling the ten folds) gave near-identical intervals to the approximate-identity bootstrap for Model A and the reference (for example Model A on CFP-FP 4.54–6.59% versus 4.60–6.61%). TAR at FMR = 10^{-3} is not interpreted because it rests on about three to four impostor pairs in total. Figure 1 shows the same comparison graphically.

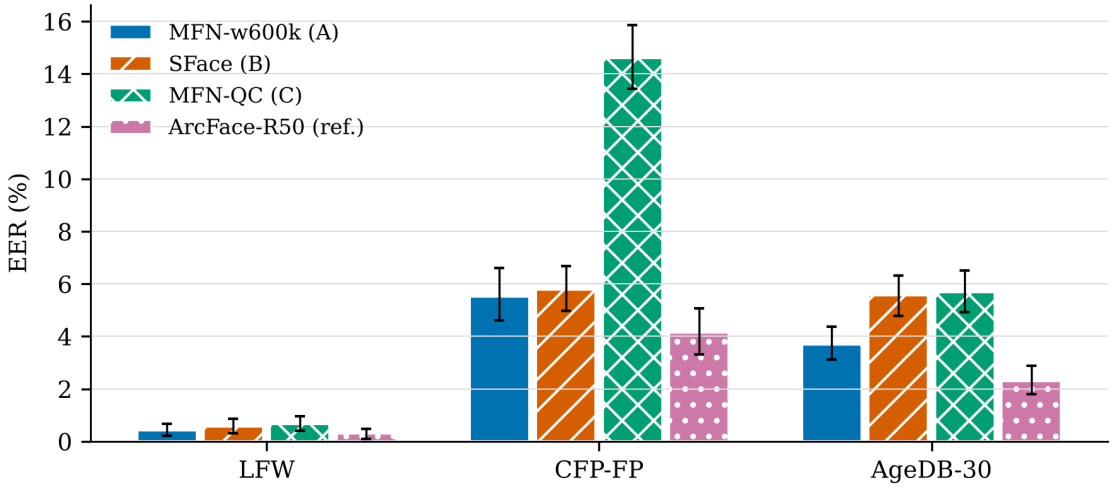


Figure 1. EER (%) of the four models on LFW, CFP-FP and AgeDB-30 with 95% bootstrap intervals (CFP-FP and AgeDB-30 intervals approximate). Bars are also distinguished by hatching.

5.2 Paired model comparison

Table 7. Paired difference in EER (percentage points) for each pair of models on identical pairs and identical identity resamples; negative = first model has lower EER. * = Holm-adjusted p < 0.05 over all 150 comparisons (6 model pairs × 25 settings); p-values use a normal approximation to the bootstrap distribution (bootstrap SE = 95% CI width ÷ 3.92) because 1,000 resamples cannot resolve p below about 0.002. Full table with intervals in Appendix D.

Setting	R50 − MFN-w600k	R50 − SFace	R50 − MFN-QC	MFN-w600k − SFace	MFN-w600k − MFN-QC	SFace − MFN-QC
LFW clean	-0.13	-0.27	-0.37	-0.13	-0.23	-0.10
CFP-FP	-1.37*	-1.66*	-10.46*	-0.29	-9.09*	-8.80*
AgeDB-30	-1.40*	-3.27*	-3.40*	-1.87*	-2.00*	-0.13
LFW blur σ=3	-1.60*	-2.50*	-10.00*	-0.90	-8.40*	-7.50*
LFW 20 px	-2.23*	-3.10*	-11.50*	-0.87	-9.27*	-8.40*
LFW JPEG 10	-0.97*	-1.77*	-2.20*	-0.80	-1.23*	-0.43
LFW band 0.3	-2.07*	-1.97*	-7.20*	+0.10	-5.13*	-5.23*

Of the 150 paired comparisons, 57 remain distinguishable after Holm correction. Among the lightweight models on clean LFW no pair is distinguishable (A − B -0.13, A − C -0.23, B − C -0.10, in points), whereas on CFP-FP MobileFaceNet-QC differs from both other lightweight models (-9.09* and -8.80*). The reference ArcFace-R50 is distinguishable from each lightweight model on CFP-FP and AgeDB-30 and in the severe LFW degradations (Table 7), but not on clean LFW for the three lightweight models taken one at a time.

5.3 Fixed-threshold cross-dataset transfer (T2)

Table 8. Threshold calibrated at FMR = 1% on the whole source dataset, frozen, applied to the target. Each cell: target FMR % / target FNMR %. Column header: source → target.

Model	LFW → CFP-FP	LFW → AgeDB-30	CFP-FP → LFW	CFP-FP → AgeDB-30	AgeDB-30 → LFW	AgeDB-30 → CFP-FP
MFN-w600k (A)	1.03 / 7.37	2.67 / 4.53	0.97 / 0.40	2.63 / 4.63	0.37 / 0.43	0.26 / 9.49
SFace (B)	0.14 / 17.11	1.83 / 9.23	4.77 / 0.27	6.17 / 5.20	0.63 / 0.57	0.06 / 18.71
MFN-QC (C)	0.94 / 34.09	3.23 / 7.70	1.10 / 0.57	3.30 / 7.57	0.27 / 0.87	0.09 / 43.51
ArcFace-R50 (ref.)	1.20 / 4.46	3.23 / 2.20	0.70 / 0.23	2.63 / 2.30	0.23 / 0.30	0.20 / 4.97

A threshold calibrated on LFW at FMR = 1% gave an FMR on AgeDB-30 above the target for all four models: 2.67% (95% CI 1.89%–3.51%), 1.83%, 3.23% and 3.23% (models A, B, C, reference), each with an interval above 1% (RQ3). On CFP-FP the same threshold gave FMR of 1.03%, 0.14%, 0.94% and 1.20%: close to the target for three models but far below it for SFace, whose FNMR rose to 17.11%. A matching FMR can also hide a large loss of usability: MobileFaceNet-QC met the FMR target on CFP-FP (0.94%) at an FNMR of 34.09%. In the opposite direction, thresholds calibrated on AgeDB-30 were conservative on the other datasets (FMR 0.06%–0.63%) with FNMR on CFP-FP between 4.97% and 43.51%. No single frozen threshold held both error rates near target across the three benchmarks for any

model, and the direction and size of the shift differed by model. The data do not identify the cause; age-gap and pose variation are plausible contributors but were not tested.

5.4 Target recalibration and its data cost (T3)

Table 9. Source = LFW (threshold at FMR = 1%). Target thresholds were recalibrated at FMR = 1% on calibration pairs from folds 1–5 of the target (stratified, half genuine; median of 50 random draws for n = 200 and 1,000; 'all' = every calibration pair) and tested on folds 6–10. 'Frozen' = LFW threshold on the same test folds. FMR / FNMR in %.

Model	Target	Frozen FMR / FNMR	Recal. FMR, n = 200	Recal. FMR, n = 1,000	Recal. FMR / FNMR, all
MFN-w600k (A)	CFP-FP	0.91 / 6.00	0.57	0.86	0.74 / 6.17
MFN-w600k (A)	AgeDB-30	2.67 / 5.13	0.60	1.07	1.13 / 6.33
SFace (B)	CFP-FP	0.11 / 15.89	1.09	1.20	1.37 / 9.09
SFace (B)	AgeDB-30	1.87 / 9.93	0.77	1.13	1.13 / 11.53
MFN-QC (C)	CFP-FP	0.74 / 33.09	0.63	0.63	0.63 / 34.11
MFN-QC (C)	AgeDB-30	3.67 / 8.07	0.60	0.87	0.87 / 13.93
ArcFace-R50 (ref.)	CFP-FP	1.26 / 3.37	0.77	0.80	0.91 / 3.43
ArcFace-R50 (ref.)	AgeDB-30	3.40 / 2.80	0.50	1.30	1.27 / 3.60

Recalibration brought the target FMR closer to 1% than the frozen LFW threshold in 6 of 8 model–dataset cases when all calibration pairs were used. On AgeDB-30, where the frozen threshold overshot for every model, recalibration with 1,000 pairs gave FMR of 1.07%, 1.13%, 0.87% and 1.30% (A, B, C, reference), compared with 2.67%, 1.87%, 3.67% and 3.40% frozen, usually at the cost of higher FNMR. For SFace on CFP-FP, where the frozen threshold was too strict (FMR 0.11%, FNMR 15.89%), recalibration moved FMR to 1.37% and reduced FNMR to 9.09%. With only 200 calibration pairs (about 100 impostor pairs) the recalibrated FMR is unstable, because an FMR of 1% corresponds to roughly one impostor pair; stable values appeared from about 1,000 pairs (RQ4). Calibration and test folds were separated by fold, but because identities are approximated for these datasets, disjointness of true identities between calibration and test could not be verified, so recalibration results may be somewhat optimistic.

5.5 Condition robustness on LFW (T4)

Table 10. LFW EER (95% identity-level bootstrap CI) under the predeclared degradation grid (both images of each pair transformed; failure rate 0 in every cell). Lower-face band: black band over the bottom 2×level of image height (levels 0.1/0.2/0.3 = 20%/40%/60%).

Condition / level	MFN-w600k (A)	SFace (B)	MFN-QC (C)	ArcFace-R50 (ref.)
Clean baseline	0.43% (0.22–0.68)	0.57% (0.30–0.87)	0.67% (0.41–0.95)	0.30% (0.09–0.48)
Blur σ 0.5	0.43% (0.22–0.69)	0.57% (0.32–0.83)	0.70% (0.42–1.02)	0.30% (0.09–0.50)
Blur σ 1.0	0.53% (0.29–0.75)	0.60% (0.37–0.85)	0.83% (0.56–1.17)	0.27% (0.08–0.49)
Blur σ 2.0	1.03% (0.70–1.32)	1.37% (0.95–1.81)	3.33% (2.79–3.94)	0.40% (0.19–0.56)
Blur σ 3.0	2.63% (2.17–3.19)	3.53% (2.98–4.17)	11.03% (10.09–12.13)	1.03% (0.70–1.37)
Downsample to px 56	0.50% (0.29–0.76)	0.63% (0.37–0.88)	1.07% (0.74–1.39)	0.27% (0.08–0.48)
Downsample to px 40	0.63% (0.38–0.91)	0.97% (0.64–1.27)	1.87% (1.46–2.35)	0.30% (0.14–0.48)
Downsample to px 28	1.37% (1.04–1.76)	2.20% (1.76–2.69)	5.90% (5.24–6.79)	0.53% (0.28–0.72)
Downsample to px 20	3.83% (3.31–4.59)	4.70% (4.11–5.41)	13.10% (12.06–14.16)	1.60% (1.26–2.05)
JPEG quality 90	0.47% (0.22–0.70)	0.60% (0.34–0.85)	0.67% (0.39–0.94)	0.30% (0.09–0.50)
JPEG quality 50	0.50% (0.29–0.77)	0.57% (0.33–0.84)	0.73% (0.43–1.03)	0.27% (0.08–0.49)
JPEG quality 30	0.53% (0.29–0.77)	0.73% (0.45–0.97)	1.03% (0.70–1.39)	0.30% (0.09–0.49)
JPEG quality 10	1.57% (1.17–2.02)	2.37% (1.96–2.91)	2.80% (2.31–3.38)	0.60% (0.36–0.88)
Brightness Δ -60	0.47% (0.29–0.74)	0.70% (0.48–1.00)	0.70% (0.47–1.01)	0.27% (0.08–0.49)
Brightness Δ -30	0.43% (0.22–0.69)	0.67% (0.41–1.02)	0.70% (0.42–1.02)	0.30% (0.09–0.48)
Brightness Δ 30	0.47% (0.23–0.70)	0.57% (0.28–0.84)	0.67% (0.44–1.00)	0.30% (0.09–0.48)
Brightness Δ 60	0.47% (0.26–0.72)	0.63% (0.34–0.87)	0.63% (0.42–0.97)	0.30% (0.09–0.50)
Contrast × 0.5	0.40% (0.21–0.66)	0.63% (0.34–0.84)	0.70% (0.46–1.00)	0.30% (0.09–0.50)
Contrast × 0.75	0.40% (0.20–0.68)	0.50% (0.28–0.79)	0.70% (0.43–1.00)	0.30% (0.09–0.50)
Contrast × 1.5	0.47% (0.27–0.69)	0.67% (0.36–0.98)	0.73% (0.47–1.07)	0.27% (0.08–0.44)
Lower-face band, level 0.1	0.57% (0.34–0.82)	0.77% (0.48–1.02)	1.00% (0.72–1.34)	0.30% (0.13–0.49)
Lower-face band, level 0.2	1.27% (0.95–1.70)	1.10% (0.77–1.49)	3.00% (2.50–3.58)	0.80% (0.49–1.05)
Lower-face band, level 0.3	5.77% (5.00–6.65)	5.67% (4.97–6.55)	10.90% (9.92–11.93)	3.70% (3.10–4.41)

Brightness and contrast changes, mild blur ($\sigma \leq 1$), moderate downsampling (≥ 40 px) and JPEG at quality ≥ 50 leave EER near the clean baseline for all four models. Differences emerge in four conditions. Under blur $\sigma = 3$, EER rises to 2.63%, 3.53%, 11.03% and 1.03%; at 20 px to 3.83%, 4.70%, 13.10% and 1.60%; under the 60%-height band to 5.77%, 5.67%, 10.90% and 3.70%; and at JPEG quality 10 to 1.57%, 2.37%, 2.80% and 0.60% (models A, B, C, reference). Relative to its clean baseline, EER at 20 px rose by +785%, +729%, +1865% and +433%, and at $\sigma = 3$ by +508%, +524%, +1555% and +244%. MobileFaceNet-QC is the most sensitive to blur and low resolution (worst-condition EER 13.10% at downsampling to 20 px), while for the other three models the lower-face band is the most damaging condition. Small point differences from the clean baseline, including slightly negative values,

lie within sampling uncertainty and should not be interpreted. Figure 2 plots these EER values against degradation level.

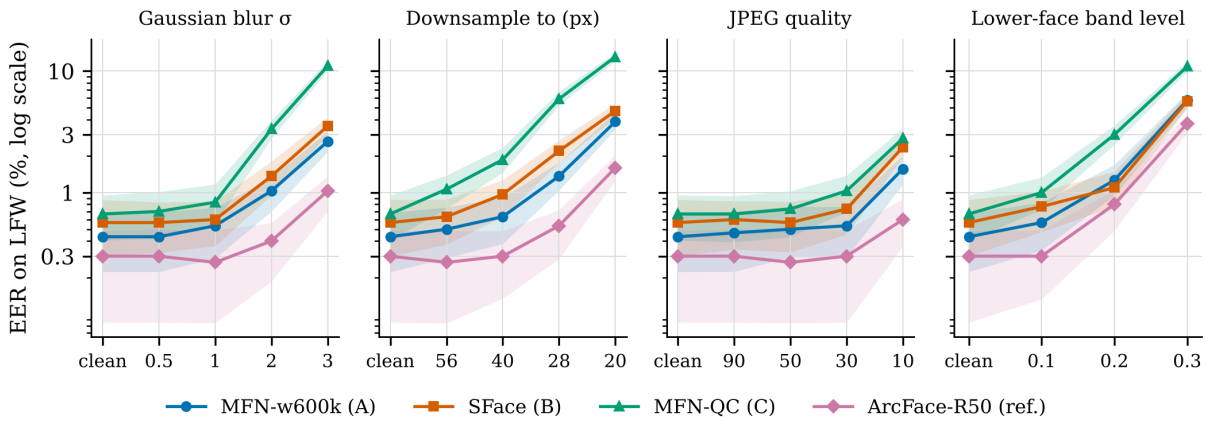


Figure 2. LFW EER (%, logarithmic axis) versus degradation level for the four models; shaded bands are 95% identity-level bootstrap intervals. 'Clean' is the undegraded baseline. Markers distinguish the models in addition to colour.

5.6 Threshold transfer under condition shift on LFW (clean → degraded)

A threshold was calibrated on the first five LFW folds (clean images) at α = 1% and frozen; it was applied to the last five folds under each degradation. The recalibrated threshold was calibrated at α = 1% on the first five folds under the same degradation. Each configuration uses 3,000 calibration pairs and 3,000 test pairs (1,500 impostor), so single FMR values carry sampling error of roughly ±0.5 points around 1%.

Table 11. Frozen clean threshold versus threshold recalibrated under each degradation (22 conditions, target FMR = 1%). Column 2: mean absolute deviation of FMR from 1% (points), frozen → recalibrated. Column 3: number of conditions with frozen FMR ≥ 2%. Columns 4–5: strongest lower-face band (level 0.3). Full tables in Appendix C.

Model	Mean |FMR − α|, frozen → recal.	Conditions with frozen FMR ≥ 2%	Band 0.3: frozen FMR	Band 0.3: recal. FMR / FNMR
MFN-w600k (A)	2.26 → 0.16	6	32.5%	1.5% / 9.9%
SFace (B)	0.92 → 0.27	6	4.2%	1.1% / 10.4%
MFN-QC (C)	12.09 → 0.23	9	93.1%	0.5% / 54.3%
ArcFace-R50 (ref.)	1.41 → 0.25	5	18.7%	1.4% / 4.9%

Across the 88 model–condition cases, the recalibrated FMR was closer to 1% than the frozen one in 57, equal in 19, and farther in 12; the cases where the frozen threshold was closer are all mild conditions (JPEG ≥ 30, brightness, contrast) with frozen FMR between 0.93% and 1.4%, i.e. within sampling noise. Recalibration restores FMR under severe degradation but moves the error to FNMR, so the choice between recapture, human review and automatic rejection remains an application decision. A threshold that does not transfer here is evidence of score-distribution shift under these synthetic conditions; no causal claim is made. Figure 3 contrasts the two thresholds under the most severe level of each degradation family.

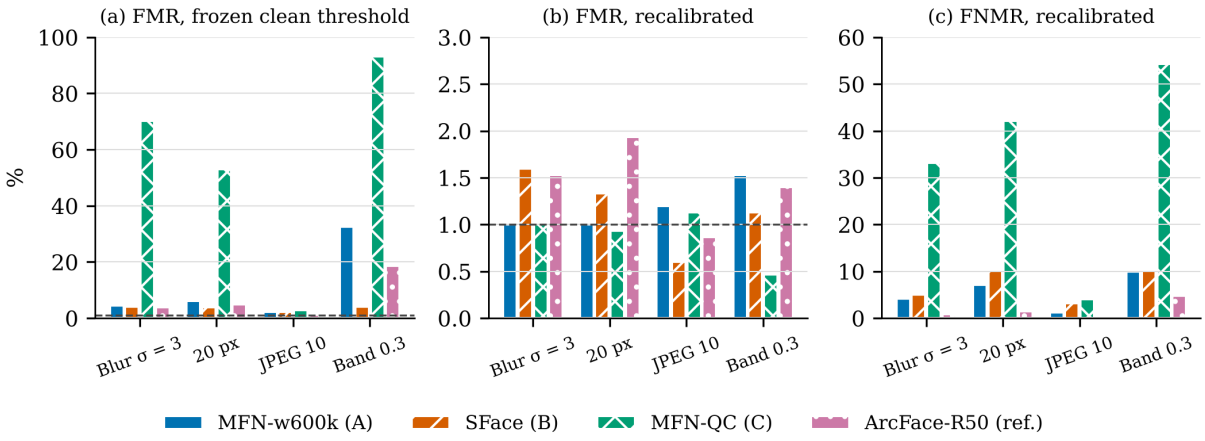


Figure 3. Threshold transfer under the most severe level of each degradation (target FMR = 1%, dashed line in panels a and b). (a) FMR with the frozen clean threshold; (b) FMR and (c) FNMR with a threshold recalibrated under the same degradation. Note the different vertical scales.

5.7 CPU performance (T5)

Table 12. CPU inference, embedding-only (normalisation + ONNX inference, no detection), batch 1, one thread, 112×112 input; three independent process launches per model on one machine (Intel Core i3-6100, 3.7 GHz, 2 physical / 4 logical cores, 31.9 GB RAM, Windows 10, Python 3.12.10, onnxruntime 1.31.0). Latency in ms (per image; MobileFaceNet-QC embeds two images per call and the call time is halved). 20 warm-up and 200 timed calls per launch.

Model	Median ms, 3 launches	Median of medians	p95 ms	p99 ms	Throughput /s	Peak RAM MB	ONNX size MiB	Latency vs Model A
MFN-w600k (A)	13.6, 13.7, 13.7	13.7	14.8	17.0	71.9	80	13.0	1.00×
SFace (B)	24.9, 25.1, 24.9	24.9	26.4	29.7	39.4	117	36.9	1.82×
MFN-QC (C)	11.4, 11.3, 11.3	11.3	12.0	12.9	87.0	88	4.13 (graph 0.37 + external weights 3.76)	0.83×
ArcFace-R50 (ref.)	287.7, 287.6, 287.6	287.6	289.4	290.6	3.5	369	166.3	21.03×

The three launches agree to within about 1% for each model, so on this machine the timings are repeatable within one session. MobileFaceNet-QC and Model A are the fastest (11.3 ms and 13.7 ms per image), SFace takes 24.9 ms and ArcFace-R50 287.6 ms, i.e. 21 times Model A. Earlier single-launch timings on the same machine had given ArcFace-R50 345 ms and 154 ms and MobileFaceNet 20.4 ms and 8.0 ms in two separate sessions, so absolute latency depends on system state and the table should be read as relative cost under one controlled session. Speed does not track robustness: MobileFaceNet-QC is faster than Model A and SFace but is far less robust on CFP-FP and under blur and low resolution (Sections 5.1 and 5.5), and the reference is most accurate but about 12–25 times slower (RQ5). End-to-end timing including detection has not been measured.

5.8 Assessment of the pre-specified hypotheses

Table 13. Status of H1–H4 on the data available in this version (three benchmarks; four models).

Hypothesis	Status	Basis
H1 non-transfer across datasets	Supported (AgeDB-30; not CFP-FP)	LFW-calibrated threshold at FMR = 1% gave FMR above 1% on AgeDB-30 for all four models, with 95% intervals excluding 1% (lower bounds 1.89%, 1.16%, 2.35%, 2.37%); on CFP-FP the FMR was close to 1% for three models and far below it for SFace (0.14%) (Table 8).
H2 recalibration reduces |FMR − α|	Supported (descriptive); effect depends on the shift	With all calibration pairs, recalibration reduced |FMR − α| in 6 of 8 model–dataset cases (AgeDB-30 for every model; CFP-FP for SFace and ArcFace-R50 only); stable from about 1,000 pairs (Table 9). Condition shift on LFW: reduced in 57 of 88 cases and equal in 19 (Table 11). FNMR rose where FMR fell. No formal interval was computed for this comparison.
H3 low resolution and strong blur degrade more than mild JPEG	Supported	Relative EER change at JPEG quality 90–50 stayed within the clean interval for all models; at 20 px: +785%, +729%, +1865%, +433%; at σ = 3: +508%, +524%, +1555%, +244% (A, B, C, reference). The 60%-height band, which H3 did not anticipate, was the most damaging condition for three of four models.
H4 ranking by clean EER differs from ranking by worst-condition EER	Not supported as a rank reversal beyond noise	Rankings (lowest EER first): clean LFW R50 < MFN-w600k < SFace < MFN-QC; CFP-FP R50 < MFN-w600k < SFace < MFN-QC; AgeDB-30 R50 < MFN-w600k < SFace < MFN-QC; worst LFW condition R50 < SFace < MFN-w600k < MFN-QC. The only point-estimate reversal is between Model A and SFace in the worst-condition ranking versus clean LFW, and these two are not distinguishable in either setting (Holm p = 1.00 on clean LFW, 1.00 under the band). The large discrepancy is instead in magnitude: models with near-identical clean LFW EER differ several-fold on CFP-FP and under blur and low resolution.

6. Discussion

Four observations follow for the models and benchmarks tested. First, clean LFW accuracy is a poor guide to robustness: three lightweight models with overlapping clean LFW intervals (0.43%, 0.57%, 0.67%) range from 5.51% to 14.60% EER on CFP-FP, and under blur σ = 3 from 2.63% to 11.03%.

Second, the operating point is more fragile than EER: a threshold set at FMR = 1% on LFW gave FMR from 1.83% to 3.23% on AgeDB-30, and under the strongest lower-face band a clean-calibrated threshold gave FMR of 32.53%, 4.20%, 93.13% and 18.73% (A, B, C, reference), which in an access-control setting is a large increase in false accepts even though EER under that band was at most 10.90%. Third, matching FMR is not enough: a model can meet its FMR target on a new dataset while FNMR becomes unusable (MobileFaceNet-QC on CFP-FP). Fourth, recalibration restores FMR but shifts the error to FNMR and needs on the order of a thousand target pairs, so the choice between recapture, human review and automatic rejection remains an application decision.

Cost and robustness did not move together: the fastest model in this set was the least robust to pose and blur, while the reference was the most accurate and robust but more than an order of magnitude slower. Which trade-off is right depends on the deployment's capture conditions; for clean, frontal, adequately resolved images the lightweight models were not distinguishable from one another or, in most cases, from the reference on LFW, whereas for profile, low-resolution or blurred captures they differed substantially.

These conclusions are limited to the conditions actually tested. The transformations are synthetic; the band occlusion is a crude stand-in for a mask; LFW is near saturation; and no claim is made about other lightweight architectures, other datasets or any particular population. Controlled degradations isolate sensitivity but do not reproduce real camera processing, motion blur, sensor noise or user behavior. For operational systems, error rates must be read alongside failure-to-process rates; here the failure rate is zero only because the images were pre-aligned, so this study is silent on detection failures.

7. Limitations and Threats to Validity

- Coverage: three lightweight models (MobileFaceNet-w600k, SFace, MobileFaceNet-QC) and one non-lightweight reference (ArcFace-R50) were run on LFW, CFP-FP and AgeDB-30. Two of the lightweight models are MobileFaceNet-family; no target-domain dataset was evaluated, and other architecture families (for example transformer or EdgeFace-type models) were not included.
- Data provenance: CFP-FP and AgeDB-30 came from an unofficial redistribution whose pair order, alignment and licence status are not documented on its dataset card. Consistency checks passed (balanced genuine/impostor pairs and equal genuine counts in all ten folds), but results should be re-run on the official releases before publication.
- Identity labels: for CFP-FP and AgeDB-30 identities were approximated from genuine-pair links (2,448 and 2,338 groups, far more than the benchmarks' subject counts), so identity-level intervals and the identity-disjointness of calibration and test folds in the recalibration analysis are approximate; a fold-level bootstrap gave near-identical intervals for two of the models.
- Benchmark ceiling: with EER of 0.3–0.7% on LFW and 3,000 pairs per class, clean-condition intervals are wide relative to the differences of interest, and one pair changes EER by about 0.03 points.
- Pre-aligned images: the images were aligned before scoring (LFW) or supplied at 112×112 (CFP-FP, AgeDB-30; alignment quality not independently verified), so detection and alignment failures were not measured (failure rate zero by construction).
- Training-data overlap: Models A and R50 are named w600k (understood to be trained on WebFace600K); MobileFaceNet-QC was trained on MS-Celeb-1M according to its model card, a dataset withdrawn by its publisher; for SFace, the paper experiments with CASIA-WebFace, VGGFace2 and MS1MV2 (a cleaned MS-Celeb-1M), but neither the OpenCV Zoo page nor the code repository states which data trained the released MobileFaceNet file, so its training set is unknown. Overlap between any model's training identities and the evaluation identities has not been checked and could inflate accuracy.

- Preprocessing selection: mean/std and channel order were chosen for Models B and C by a probe of eight fixed variants on 600 LFW pairs (fold 1); the probe tied for SFace between RGB and BGR channel order (identical probe EER 0.67%), RGB was used. A full-data check with BGR order (all 6,000 LFW, 7,000 CFP-FP and 6,000 AgeDB-30 pairs) gave higher EER on every benchmark (LFW 0.70% versus 0.57%, CFP-FP 7.69% versus 5.80%, AgeDB-30 5.97% versus 5.57%), so RGB order, as in the reported results, was retained; the pipeline's image-loading channel convention was not independently audited, so this should be read as an empirical choice rather than a statement about the model's native convention. Models A and R50 used mean = std = 127.5.
- Synthetic degradations and a simple lower-face band approximate, but do not replace, real capture variation; transformations were applied to both images of each pair, so enrolment–probe quality mismatch (one-sided degradation) was not studied.
- Statistical: transfer comparisons use 3,000 calibration and 3,000 test pairs (1,500 impostor), so FMR estimates near 1% carry noticeable sampling error; no interval was computed for the fixed-versus-recalibrated comparison; FMR at 10^{-3} is not supported by the number of impostor pairs.
- CPU timing: single hardware platform (Windows 10, Intel Core i3-6100, 2 physical / 4 logical cores), single thread, embedding only; MobileFaceNet-QC embeds two images per call and its time is halved; earlier single-launch timings differed by 2–3× in absolute latency from the final three-launch session, indicating sensitivity to system state.
- Ethical and legal: only public benchmarks were used; the licences of the CFP-FP, AgeDB-30 and model files must be checked before release (EdgeFace-type models, for example, carry non-commercial terms); any local dataset would require an appropriate legal basis, consent, privacy and access-control safeguards, and no demographic claims are made.

8. Conclusion

This paper defines a reproducible framework for measuring robustness and generalization of lightweight face-verification models and reports a first executed study of three lightweight models and one reference on LFW, CFP-FP and AgeDB-30. The three lightweight models had overlapping clean LFW EER (0.43%–0.67%) but differed several-fold on CFP-FP (5.51%–14.60%) and under blur and low resolution; a threshold calibrated on LFW overshot its FMR target on AgeDB-30 for every model, and a matching FMR on CFP-FP could coincide with a very high FNMR. Recalibration restored FMR at the price of higher FNMR and needed about a thousand target pairs. Speed did not predict robustness: the reference was most robust but 25 times slower than the fastest lightweight model on a single CPU thread. These findings rest on two benchmarks with approximated identities from an unofficial source, synthetic degradations and one hardware platform, so official-release replication, additional architecture families and real-capture data are required before deployment-level conclusions are drawn.

Declarations

- Funding: This research received no external funding.
- Conflicts of interest: The author declares no conflicts of interest.

- Ethics and consent: Public benchmarks are used under their licences; no new human-subject data are collected in this protocol version.
- Model licences: MobileFaceNet-w600k and ArcFace-R50 (InsightFace model zoo): available for non-commercial research purposes only, per the model-zoo README, which does not state terms for the WebFace600K training data; SFace (OpenCV Zoo): Apache-2.0, training data not stated on the model page; MobileFaceNet-QC (Qualcomm AI Hub): Apache-2.0 per the model card.
- Data availability: LFW, CFP-FP and AgeDB-30 are available from their original providers under their respective terms; pair lists follow the official protocols. CFP-FP and AgeDB-30 pair data in the

reported run came from the Hugging Face redistribution cat-claws/face-verification (official releases: AgeDB is distributed by its authors at Imperial College London (iBUG group) for non-commercial research use, with the archive password provided on request by email, and annotations may not be redistributed; CFP-FP is distributed by its authors (Sengupta et al., WACV 2016) and its terms of use could not be accessed during preparation and should be checked with the dataset providers before release). Any target-domain data are not public.

- Code availability: Evaluation toolkit (fveval) and the archived output files of the reported run (summary.json, section5.json, section5_new.json, section5_extra.json, paired4.json, run_log.json, per-pair score CSVs for all models and datasets, bootstrap cache, CPU benchmark JSON files); [add repository URL and commit hash upon release].
- Author contributions: MH Rahman conceived the study, designed the protocol, ran the experiments, analysed the results and wrote the manuscript.

References

1. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823.
2. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4690–4699.
3. Chen, S., Liu, Y., Gao, X., & Han, Z. (2018). MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. Chinese Conference on Biometric Recognition (CCBR), 428–438.
4. Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. University of Massachusetts Amherst, Technical Report 07-49.
5. Sengupta, S., Chen, J.-C., Castillo, C., Patel, V. M., Chellappa, R., & Jacobs, D. W. (2016). Frontal to profile face verification in the wild. IEEE Winter Conference on Applications of Computer Vision (WACV).
6. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., & Zafeiriou, S. (2017). AgeDB: The first manually collected, in-the-wild age database. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
7. Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10), 1090–1104.
8. International Organization for Standardization / International Electrotechnical Commission. ISO/IEC 19795 series, Information technology — Biometric performance testing and reporting. Part 1: Principles and framework (ISO/IEC 19795-1:2021).
9. National Institute of Standards and Technology. Face Recognition Technology Evaluation (FRTE, formerly FRVT) reports. FRTE 1:1 Verification, https://pages.nist.gov/frvt/ (accessed 27 October 2026).
10. Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman & Hall.
11. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

12. Zhong, Y., Deng, W., Hu, J., Zhao, D., Li, X., & Wen, D. SFace: Sigmoid-constrained hypersphere loss for robust face recognition. arXiv:2205.12010.

13. OpenCV Zoo. face_recognition_sface model page, https://github.com/opencv/opencv_zoo (accessed 27 October 2026). InsightFace model zoo README, https://github.com/deepinsight/insightface (accessed 27 October 2026). AgeDB resource page, https://ibug.doc.ic.ac.uk/resources/agedb/ (accessed 27 October 2026).

Appendix A. Reproducibility Checklist

- Record exact model architecture, checkpoint hash, licence and inference framework.
- Record dataset versions, official protocols, pair lists, exclusions and sample counts.
- Keep identity, session and calibration/test partitions separate where applicable; check cross-dataset identity overlap.
- Predefine transformations and all parameter values before examining final test results.
- Freeze source-calibrated thresholds for cross-dataset fixed-threshold evaluation.
- Count detection, alignment and embedding failures explicitly; report failure-inclusive sensitivity analysis.
- Report confidence intervals and the method used (identity-level bootstrap, B, seed).
- Record CPU, RAM, OS, runtime version, thread settings, warm-up, repetitions and timing boundaries.
- Avoid unsupported demographic claims and document privacy safeguards for local data.
- Publish code, configuration and result tables where licences and privacy rules permit.

Appendix B. Suggested Experiment Sequence

1. Select models and verify that checkpoints and preprocessing can be reproduced (hash check, embedding sanity check on a few pairs).
2. Run each official benchmark protocol independently (T1).
3. Create calibration splits and freeze thresholds before final testing.
4. Run fixed-threshold cross-dataset transfer tests (T2) and recalibration with a ladder of calibration sizes (T3).
5. Apply the predeclared degradation grid to held-out test images (T4).
6. Run CPU timing tests under documented conditions (T5).
7. Compute identity-level confidence intervals, paired differences and Holm-adjusted tests; fill Tables 6–13.
8. Rewrite the abstract, discussion and conclusion using only measured evidence, and mark each of H1–H4 as supported or not supported (done for the four-model run; repeat on official CFP-FP/AgeDB-30 releases and target-domain data).

Appendix C. Full clean-to-degraded threshold transfer (target FMR = 1%)

Table C1. MFN-w600k (A): all 22 degraded conditions. Fixed = frozen clean threshold; Recal = recalibrated under the same degradation.

Condition	FMR fixed	FNMR fixed	FMR recal	FNMR recal
Blur σ 0.5	1.00%	0.33%	1.00%	0.33%
Blur σ 1.0	1.33%	0.40%	1.13%	0.40%
Blur σ 2.0	1.47%	0.53%	0.87%	1.07%
Blur σ 3.0	4.47%	1.60%	1.00%	4.20%
Downsample to px 56	1.20%	0.33%	1.20%	0.33%
Downsample to px 40	1.27%	0.47%	0.87%	0.47%
Downsample to px 28	2.53%	0.73%	1.00%	1.33%
Downsample to px 20	6.13%	1.87%	1.00%	7.20%
JPEG quality 90	1.07%	0.33%	0.93%	0.33%
JPEG quality 50	1.40%	0.33%	1.60%	0.33%
JPEG quality 30	1.20%	0.33%	1.07%	0.33%
JPEG quality 10	2.20%	1.07%	1.20%	1.27%
Brightness Δ -60	1.13%	0.33%	1.40%	0.27%
Brightness Δ -30	1.00%	0.33%	1.13%	0.33%
Brightness Δ 30	1.00%	0.33%	0.93%	0.33%
Brightness Δ 60	1.07%	0.27%	0.87%	0.27%
Contrast × 0.5	1.13%	0.33%	1.07%	0.33%
Contrast × 0.75	0.93%	0.33%	0.93%	0.33%
Contrast × 1.5	0.93%	0.27%	0.87%	0.27%
Lower-face band, level 0.1	1.53%	0.13%	1.07%	0.27%
Lower-face band, level 0.2	4.93%	0.33%	1.33%	1.13%
Lower-face band, level 0.3	32.53%	1.20%	1.53%	9.93%

Table C2. SFace (B): all 22 degraded conditions. Fixed = frozen clean threshold; Recal = recalibrated under the same degradation.

Condition	FMR fixed	FNMR fixed	FMR recal	FNMR recal
Blur σ 0.5	1.33%	0.40%	1.20%	0.40%
Blur σ 1.0	1.53%	0.40%	1.27%	0.40%
Blur σ 2.0	2.27%	0.73%	1.13%	1.20%
Blur σ 3.0	4.20%	2.67%	1.60%	5.07%
Downsample to px 56	1.33%	0.33%	1.13%	0.40%
Downsample to px 40	1.80%	0.53%	0.73%	0.80%
Downsample to px 28	3.07%	1.27%	1.47%	2.67%
Downsample to px 20	4.00%	4.93%	1.33%	10.20%
JPEG quality 90	1.40%	0.40%	1.40%	0.40%
JPEG quality 50	1.27%	0.40%	1.20%	0.40%
JPEG quality 30	1.40%	0.33%	1.53%	0.33%
JPEG quality 10	2.27%	2.33%	0.60%	3.20%
Brightness Δ -60	0.40%	0.87%	1.33%	0.53%
Brightness Δ -30	0.80%	0.47%	1.13%	0.47%
Brightness Δ 30	1.40%	0.40%	1.13%	0.40%
Brightness Δ 60	1.40%	0.33%	1.60%	0.33%
Contrast × 0.5	1.27%	0.40%	1.27%	0.40%
Contrast × 0.75	1.27%	0.40%	1.27%	0.40%
Contrast × 1.5	0.80%	0.40%	1.00%	0.40%
Lower-face band, level 0.1	0.73%	0.47%	0.93%	0.40%
Lower-face band, level 0.2	0.47%	1.13%	1.00%	0.80%
Lower-face band, level 0.3	4.20%	5.40%	1.13%	10.40%

Table C3. MFN-QC (C): all 22 degraded conditions. Fixed = frozen clean threshold; Recal = recalibrated under the same degradation.

Condition	FMR fixed	FNMR fixed	FMR recal	FNMR recal
Blur σ 0.5	1.40%	0.47%	1.40%	0.47%
Blur σ 1.0	1.33%	0.40%	1.13%	0.47%
Blur σ 2.0	9.33%	0.80%	0.93%	5.13%
Blur σ 3.0	70.33%	0.33%	1.00%	33.20%
Downsample to px 56	1.47%	0.60%	1.07%	0.87%
Downsample to px 40	3.60%	1.07%	0.93%	2.60%
Downsample to px 28	16.60%	1.33%	0.73%	15.53%
Downsample to px 20	53.00%	1.67%	0.93%	42.20%
JPEG quality 90	1.40%	0.47%	1.40%	0.47%
JPEG quality 50	1.53%	0.40%	1.40%	0.53%
JPEG quality 30	1.33%	0.73%	1.33%	0.73%
JPEG quality 10	2.93%	2.33%	1.13%	4.07%
Brightness Δ -60	1.13%	0.60%	1.00%	0.60%
Brightness Δ -30	1.47%	0.47%	1.27%	0.53%
Brightness Δ 30	1.33%	0.47%	1.27%	0.47%
Brightness Δ 60	1.20%	0.40%	1.20%	0.40%
Contrast × 0.5	1.27%	0.47%	1.13%	0.47%
Contrast × 0.75	1.40%	0.47%	1.40%	0.47%
Contrast × 1.5	1.33%	0.53%	1.27%	0.53%
Lower-face band, level 0.1	2.00%	0.27%	0.93%	0.73%
Lower-face band, level 0.2	19.47%	0.60%	0.47%	5.07%
Lower-face band, level 0.3	93.13%	0.07%	0.47%	54.33%

Table C4. ArcFace-R50 (ref.): all 22 degraded conditions. Fixed = frozen clean threshold; Recal = recalibrated under the same degradation.

Condition	FMR fixed	FNMR fixed	FMR recal	FNMR recal
Blur σ 0.5	1.20%	0.20%	1.20%	0.20%
Blur σ 1.0	1.00%	0.13%	1.00%	0.13%
Blur σ 2.0	1.47%	0.07%	1.47%	0.07%
Blur σ 3.0	3.93%	0.40%	1.53%	0.87%
Downsample to px 56	0.93%	0.13%	0.93%	0.13%
Downsample to px 40	1.33%	0.13%	0.93%	0.13%
Downsample to px 28	2.33%	0.20%	1.53%	0.20%
Downsample to px 20	4.87%	0.87%	1.93%	1.47%
JPEG quality 90	1.13%	0.20%	1.47%	0.20%
JPEG quality 50	1.27%	0.20%	1.33%	0.20%
JPEG quality 30	1.27%	0.13%	1.13%	0.20%
JPEG quality 10	1.27%	0.27%	0.87%	0.47%
Brightness Δ -60	1.20%	0.20%	1.07%	0.20%
Brightness Δ -30	1.07%	0.20%	1.20%	0.20%
Brightness Δ 30	0.93%	0.20%	1.00%	0.20%
Brightness Δ 60	1.07%	0.20%	1.13%	0.20%
Contrast × 0.5	1.00%	0.13%	1.00%	0.13%
Contrast × 0.75	1.00%	0.13%	1.00%	0.13%
Contrast × 1.5	0.87%	0.20%	0.87%	0.20%
Lower-face band, level 0.1	1.47%	0.20%	1.20%	0.20%
Lower-face band, level 0.2	3.20%	0.27%	0.60%	0.60%
Lower-face band, level 0.3	18.73%	1.13%	1.40%	4.87%

Appendix D. Full paired EER differences with Holm-adjusted significance

Table D1. Paired ΔEER in percentage points [95% bootstrap CI]; * = Holm-adjusted p < 0.05 over 150 tests. Column pairs as in Table 7.

Setting	R50 − MFN-w600k	R50 − SFace	R50 − MFN-QC	MFN-w600k − SFace	MFN-w600k − MFN-QC	SFace − MFN-QC
LFW clean	-0.13 [-0.35, +0.00]	-0.27 [-0.55, -0.09]	-0.37 [-0.64, -0.15]	-0.13 [-0.38, +0.12]	-0.23 [-0.45, -0.02]	-0.10 [-0.38, +0.20]
LFW: Blur σ 0.5	-0.13 [-0.35, +0.00]	-0.27 [-0.50, -0.08]	-0.40 [-0.69, -0.17]	-0.13 [-0.33, +0.09]	-0.27 [-0.47, -0.03]	-0.13 [-0.41, +0.12]
LFW: Blur σ 1.0	-0.27 [-0.43, -0.07]	-0.33 [-0.52, -0.12]	-0.57* [-0.86, -0.29]	-0.07 [-0.26, +0.10]	-0.30 [-0.56, -0.10]	-0.23 [-0.51, +0.02]
LFW: Blur σ 2.0	-0.63* [-0.95, -0.34]	-0.97* [-1.37, -0.64]	-2.93* [-3.55, -2.50]	-0.33 [-0.71, +0.00]	-2.30* [-2.93, -1.87]	-1.97* [-2.58, -1.49]
LFW: Blur σ 3.0	-1.60* [-2.09, -1.15]	-2.50* [-3.15, -1.94]	-10.00* [-11.03, -9.10]	-0.90 [-1.45, -0.37]	-8.40* [-9.41, -7.59]	-7.50* [-8.49, -6.63]
LFW: Brightness Δ -30	-0.13 [-0.36, -0.03]	-0.37 [-0.68, -0.17]	-0.40 [-0.67, -0.19]	-0.23 [-0.49, +0.00]	-0.27 [-0.45, -0.02]	-0.03 [-0.30, +0.28]
LFW: Brightness Δ -60	-0.20 [-0.46, -0.02]	-0.43* [-0.70, -0.24]	-0.43* [-0.70, -0.22]	-0.23 [-0.50, +0.00]	-0.23 [-0.46, +0.00]	+0.00 [-0.29, +0.28]
LFW: Brightness Δ 30	-0.17 [-0.36, -0.02]	-0.27 [-0.49, -0.07]	-0.37 [-0.66, -0.19]	-0.10 [-0.30, +0.12]	-0.20 [-0.46, -0.05]	-0.10 [-0.42, +0.13]
LFW: Brightness Δ 60	-0.17 [-0.39, +0.00]	-0.33 [-0.53, -0.12]	-0.33 [-0.63, -0.14]	-0.17 [-0.33, +0.07]	-0.17 [-0.42, +0.02]	+0.00 [-0.33, +0.19]
LFW: Contrast × 0.5	-0.10 [-0.31, +0.00]	-0.33 [-0.51, -0.09]	-0.40 [-0.69, -0.20]	-0.23 [-0.35, +0.02]	-0.30 [-0.53, -0.10]	-0.07 [-0.39, +0.14]
LFW: Contrast × 0.75	-0.10 [-0.33, +0.00]	-0.20 [-0.42, -0.09]	-0.40 [-0.66, -0.17]	-0.10 [-0.29, +0.07]	-0.30 [-0.50, -0.07]	-0.20 [-0.43, +0.07]
LFW: Contrast × 1.5	-0.20 [-0.41, -0.03]	-0.40 [-0.68, -0.16]	-0.47* [-0.77, -0.24]	-0.20 [-0.45, +0.07]	-0.27 [-0.52, -0.05]	-0.07 [-0.41, +0.22]
LFW: Downsample to px 20	-2.23* [-2.84, -1.74]	-3.10* [-3.80, -2.49]	-11.50* [-12.53, -10.43]	-0.87 [-1.51, -0.24]	-9.27* [-10.31, -8.11]	-8.40* [-9.37, -7.24]
LFW: Downsample to px 28	-0.83* [-1.22, -0.57]	-1.67* [-2.20, -1.24]	-5.37* [-6.25, -4.78]	-0.83* [-1.28, -0.37]	-4.53* [-5.32, -3.93]	-3.70* [-4.55, -3.07]
LFW: Downsample to px 40	-0.33 [-0.58, -0.14]	-0.67* [-0.97, -0.37]	-1.57* [-2.01, -1.19]	-0.33 [-0.61, +0.00]	-1.23* [-1.62, -0.86]	-0.90* [-1.37, -0.55]
LFW: Downsample to px 56	-0.23 [-0.44, -0.05]	-0.37 [-0.59, -0.13]	-0.80* [-1.09, -0.47]	-0.13 [-0.32, +0.08]	-0.57* [-0.83, -0.25]	-0.43 [-0.71, -0.12]
LFW: JPEG quality 10	-0.97* [-1.32, -0.63]	-1.77* [-2.26, -1.34]	-2.20* [-2.70, -1.69]	-0.80 [-1.31, -0.40]	-1.23* [-1.74, -0.79]	-0.43 [-0.92, +0.12]
LFW: JPEG quality 30	-0.23 [-0.43, -0.05]	-0.43* [-0.65, -0.17]	-0.73* [-1.07, -0.42]	-0.20 [-0.42, +0.03]	-0.50 [-0.79, -0.22]	-0.30 [-0.66, +0.02]
LFW: JPEG quality 50	-0.23 [-0.47, -0.07]	-0.30 [-0.52, -0.12]	-0.47* [-0.69, -0.20]	-0.07 [-0.24, +0.14]	-0.23 [-0.41, +0.02]	-0.17 [-0.37, +0.08]

LFW: JPEG quality 90	-0.17 [-0.36, +0.00]	-0.30 [-0.51, -0.09]	-0.37 [-0.62, -0.14]	-0.13 [-0.33, +0.09]	-0.20 [-0.39, +0.00]	-0.07 [-0.30, +0.19]
LFW: Lower-face band, level 0.1	-0.27 [-0.50, -0.07]	-0.47 [-0.74, -0.20]	-0.70* [-1.03, -0.42]	-0.20 [-0.41, +0.07]	-0.43 [-0.73, -0.19]	-0.23 [-0.61, +0.02]
LFW: Lower-face band, level 0.2	-0.47 [-0.94, -0.27]	-0.30 [-0.67, -0.08]	-2.20* [-2.78, -1.77]	+0.17 [-0.14, +0.56]	-1.73* [-2.18, -1.19]	-1.90* [-2.40, -1.42]
LFW: Lower-face band, level 0.3	-2.07* [-2.75, -1.51]	-1.97* [-2.61, -1.40]	-7.20* [-8.05, -6.26]	+0.10 [-0.41, +0.68]	-5.13* [-5.94, -4.15]	-5.23* [-6.00, -4.25]
CFP-FP	-1.37* [-1.93, -0.93]	-1.66* [-2.25, -1.05]	-10.46* [-11.65, -9.39]	-0.29 [-0.85, +0.37]	-9.09* [-10.18, -7.94]	-8.80* [-10.00, -7.76]
AgeDB-30	-1.40* [-1.96, -0.93]	-3.27* [-3.87, -2.53]	-3.40* [-4.02, -2.68]	-1.87* [-2.41, -1.12]	-2.00* [-2.57, -1.24]	-0.13 [-0.85, +0.58]