International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(19s) 2026 pp. 1221-1235
https://doi.org/10.70917/ijcisim-2026-5120
Article

FaceBench: A Locked-Pipeline Comparative Evaluation of Five Off-the-Shelf Face Recognizers across Pose, Age, and Resolution

Nehalkumar Patel[1], Dr. Bharat Patel[1]

[1] Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India
Email: iamnehalpatel@gmail.com, patelbharat99@yahoo.com

Abstract: Fair comparison of pretrained face recognition models is often confounded by vendor-specific detectors, incompatible crop geometries, and accuracy-only reporting at an unstated decision threshold. This paper presents FaceBench, a reproducible evaluation framework, and reports a locked comparative study of five off-the-shelf recognizers — FaceNet, Dlib, InsightFace Buffalo-L, AdaFace, and MagFace — under one shared RetinaFace/SCRFD detect-and-crop protocol (Baseline B: bounding-box margin 0.35, cosine operating point 0.40). Primary evidence spans pairwise verification on four high-quality (HQ) benchmarks — LFW, CFP-FP, CPLFW, and AgeDB-30 — covering frontal, cross-pose, and cross-age conditions, together with dedicated uncached embedding throughput, pairwise McNemar significance tests against the leading model, LFW closed-set Rank-1 identification as a function of gallery size $N \in \{10, 100, 500, 1000\}$, and an accuracy–efficiency synthesis. Buffalo-L attains the highest mean verification accuracy across the four HQ sets (0.9045) at 27.63 frames/s on the reported workstation; MagFace and AdaFace are the fastest embedders (97.60 and 92.83 frames/s) but measurably weaker on hard-pose accuracy at the shared cosine threshold; Dlib retains a high ROC-AUC and low equal error rate (EER) yet its fixed-threshold accuracy collapses to chance level because the fixed cosine threshold of 0.40 is poorly calibrated for Dlib's similarity distribution. Every accuracy gap against Buffalo-L is statistically significant under McNemar's test (all $p < 0.001$; closest is CPLFW versus FaceNet, $p = 4.2\times10^{-5}$). A supplementary TinyFace Rank-1 probe (Gallery_Match versus Probe, $N \leq 500$) is reported separately from the HQ means and is explicitly not official TinyFace mean average precision. Source code and experiment manifests are released under an MIT license; face images are used under each dataset's original terms and are not redistributed.

Keywords: Face recognition; benchmarking; reproducibility; RetinaFace; FaceNet; Dlib; ArcFace; Buffalo-L; AdaFace; MagFace; McNemar's test; equal error rate; closed-set identification

1. Introduction

Deep face recognizers are compared constantly across research papers, vendor benchmarks, and internal engineering evaluations, yet the experimental practice behind those comparisons is frequently loose. Studies mix vendor-specific detectors, use incompatible crop geometries and input sizes, and report accuracy at a single unstated or arbitrarily chosen similarity threshold, without also disclosing the false-accept rate, false-reject rate, equal error rate (EER), latency, or memory footprint that a deployment engineer would actually need [1]–[4]. When two recognizers are evaluated through two different preprocessing pipelines, an apparent accuracy gap could be a genuine backbone effect, a preprocessing artifact, or some blend of both, and there is usually no way for a reader to tell which.

Canonical high-quality (HQ) verification protocols — LFW [5], CFP-FP [6], CPLFW [7], and AgeDB-30 [8] — are the standard reporting surface in training-loss papers such as MagFace [3] and AdaFace [4]. Those tables evaluate new loss functions under each method's own native ArcFace-style 112×112 alignment, tuned specifically to the model being introduced; they are not an independent, apples-to-apples re-evaluation of frozen, publicly distributed FaceNet [1], Dlib [10], and Buffalo-L [2], [24] packs run through one shared detector and one shared decision threshold. General-purpose toolkits such as DeepFace [11], face.evoLVe [12], and FaceX-Zoo [13] partially address this gap, but each either varies the detector per model, is oriented around training rather than head-to-head frozen-

model comparison, or omits this specific five-model mix of triplet, ArcFace-family, quality-adaptive, and classical-descriptor recognizers.

This paper makes four contributions, extending an earlier LFW-only Baseline B pilot into a multi-protocol study with statistical testing, throughput measurement, and closed-set identification:

- FaceBench, a YAML-driven Python framework that fixes detection and alignment once per image and reuses experiment manifests, weight checksums, and cached crops across every recognizer evaluated in a run, for comparative evaluation of frozen, pretrained recognizers.
- A locked Baseline B campaign of five public models on LFW, CFP-FP, CPLFW, and AgeDB-30 at a single shared cosine threshold of 0.40, reporting accuracy, AUC, and EER for every model on every dataset, together with pairwise McNemar significance tests against the leading model.
- Dedicated, uncached embedding throughput measured on the same crops used for accuracy, and LFW closed-set Rank-1 identification accuracy as a function of gallery size $N \in \{10, 100, 500, 1000\}$, to characterize how verification-style accuracy degrades under identification-style search.
- An accuracy–efficiency Pareto reading of the four HQ sets, together with a supplementary TinyFace [19] Rank-1 probe that is deliberately kept out of the HQ means and is not presented as official TinyFace mean average precision (mAP).

The remainder of the paper proceeds as follows. Section II reviews the training objectives behind the five recognizers and prior benchmarking tooling. Section III describes the FaceBench pipeline, the Baseline B preprocessing configuration, and the full set of metrics and statistical tests used, with explicit formulas. Section IV reports results across seven axes: verification accuracy, ROC-AUC and EER, statistical significance, computational cost, identification versus gallery size, the supplementary TinyFace probe, and an accuracy–efficiency synthesis. Section V discusses limitations and threats to validity, and Section VI concludes.

2. Related Work

A. Recognition backbones and training objectives

FaceNet learns embeddings directly with a triplet loss, optimizing relative distances between anchor, positive, and negative images rather than a classification surrogate [1]. ArcFace reformulates the softmax classifier with an additive angular margin so that the training objective matches the cosine-similarity comparison used at test time [2]; InsightFace's Buffalo-L is a widely used, large-scale ArcFace-style pack distributed as an ONNX model [2], [24]. MagFace couples an angular margin with a magnitude-dependent regularizer so that embedding norm becomes a usable proxy for perceived face quality [3]. AdaFace instead scales its margin adaptively using a feature-norm-derived quality estimate, down-weighting low-quality training samples [4]. Dlib provides a compact ResNet descriptor together with five-point landmark alignment, trained with a different, less publicly documented pipeline than the four deep-margin models [10].

Because Acc@0.40 is measured here under one shared bounding-box-margin crop rather than under each model's native alignment recipe, it should be read as a different operating condition from the native-align state-of-the-art tables reported in the MagFace and AdaFace papers themselves [3], [4] — a distinction this study returns to repeatedly in Sections IV and V, since it is the single most important methodological caveat behind the headline numbers.

B. Benchmarking toolkits and the shared-detector gap

DeepFace publishes convenient LFW-style comparison matrices across many backbones but varies the detector per model, which reintroduces exactly the preprocessing confound this study is designed to remove [11]. face.evoLVe and FaceX-Zoo are broader training-and-evaluation toolboxes, oriented more toward reproducing training pipelines than toward auditing frozen, off-the-shelf models under one locked detector [12], [13]. Independent multi-architecture comparisons such as Shahi et al. [22] evaluate several face-recognition heads across datasets but omit this locked five-model mix under a shared detector. Sowan et al. compare two recognizers on a single LFW subset without the multi-protocol, multi-metric scope pursued here [14]. Friedman et al. and the NIST Face Recognition Vendor Test (FRVT) Part 2 address how accuracy scales with gallery size at an operational, vendor-submission scale that differs from the single-workstation, five-model academic comparison in Section IV-E [15], [16]. Pairwise McNemar-style significance testing for face recognition has precedent in classical FERET/PCA evaluations [17], [18], which this study revisits with modern deep embeddings.

To our knowledge, this is the first joint off-the-shelf comparison of FaceNet, Dlib, Buffalo-L, AdaFace, and MagFace under a shared RetinaFace/SCRFD Baseline B protocol that simultaneously reports HQ verification accuracy across four datasets, dedicated uncached embedding throughput, pairwise McNemar significance tests, and LFW Rank-1 identification versus gallery size N ∈ {10, 100, 500, 1000}. That claim concerns experimental scope and methodology, not a new recognition backbone or loss function.

3. FaceBench and Experimental Protocol

A. Framework overview

FaceBench runs an evaluation as a linear pipeline: configuration → dataset integrity checking → identity-pair or probe/gallery construction → shared detection and alignment → per-model embedding extraction → similarity matching or CMC ranking → metric computation and manifest generation. Every stage writes to a run-specific experiment directory (see Data and Code Availability), so the resolved configuration, the pairs actually scored, the cached crops, the per-model embeddings, and the final metrics are all inspectable after the fact. AdaFace and MagFace are evaluated with their official backbones (IR-50 and iResNet-50 respectively, both trained on MS1MV2 [20]) through paper-local input wrappers that convert FaceBench's cached crop to each model's documented convention — BGR scaled to [−1, 1] for AdaFace, BGR scaled to [0, 1] for MagFace — without editing the shared preprocessing utilities that every other model continues to use unchanged. FaceNet uses an InceptionResNetV1 backbone trained on VGGFace2 [21].

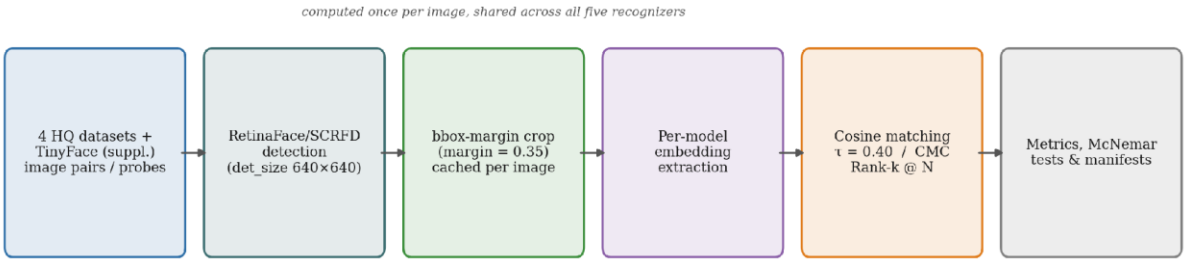

Figure 1: The FaceBench Baseline B pipeline. Detection and alignment are computed once per image and cached, then reused across all five recognizers for both verification (cosine matching at τ = 0.40) and identification (CMC Rank-k at gallery size N).

B. Detection, alignment, and matching

Detection uses InsightFace's RetinaFace/SCRFD detector at a fixed input size of 640×640 [9], [23]. Detected faces are cropped with a bounding-box margin of 0.35, expanding each detected box by 35% on every side before cropping — a deliberate compromise that preserves context for models which re-detect or re-align internally (Buffalo-L, Dlib) while remaining directly consumable by resize-based embedders (FaceNet, AdaFace, MagFace). Cosine matching is locked at $\tau = 0.40$ for all five models on all four HQ verification sets; EER is reported separately as a threshold-independent complement. Pairs or probes whose detection fails are skipped rather than aborting the run, which is why per-model scored counts vary slightly across datasets (Section IV). All experiments ran on a single NVIDIA GeForce RTX 3050 (cuda:0); uncached embedding throughput (Section IV-D) is measured over N = 200 timed crops after a warm-up pass.

The cosine similarity between two L2-normalized embeddings e_1 and e_2, and the shared decision rule applied at every threshold-dependent evaluation, are

$$\mathrm{sim}_{\cos}(\mathbf{e}_1, \mathbf{e}_2) = \frac{\mathbf{e}_1 \cdot \mathbf{e}_2}{\|\mathbf{e}_1\| \, \|\mathbf{e}_2\|}$$

$$\hat{y} = 1 \text{ (same) if } \mathrm{sim}_{\cos}(\mathbf{e}_1, \mathbf{e}_2) \geq \tau \,, \quad \hat{y} = 0 \text{ (different) otherwise} \,, \quad \tau = 0.40$$

C. Datasets and scale

Table 1: summarizes the five protocols used. LFW, CFP-FP, CPLFW, and AgeDB-30 form the primary, high-quality (HQ) verification evidence and are combined into the Table 3 mean; TinyFace is reported separately as a supplementary, low-resolution identification probe and is explicitly excluded from every HQ aggregate. Occlusion, surveillance-camera, and video-based protocols are out of scope for this study.

Dataset	Task / scale	Primary variation	Reference
LFW View-2	6,000 verification pairs	Mostly frontal, in-the-wild	[5]
CFP-FP	7,000 verification pairs	Frontal-to-profile pose	[6]
CPLFW	6,000 verification pairs	Cross-pose, in-the-wild	[7]
AgeDB-30	6,000 verification pairs (≈30-yr gap)	Cross-age, manually verified	[8]
TinyFace (supplementary)	CMC Rank-1, Gallery_Match vs. Probe	Native low resolution	[19]

Table 1. Verification and identification protocols used in this study. TinyFace uses FaceBench's own CMC Rank-1 computation on Gallery_Match vs. Probe only; it is not the official TinyFace MATLAB mean average precision (mAP) protocol, and Training_Set / Gallery_Distractor splits are unused.

D. Recognizer architectures

Model	Backbone	Training objective / data	Embedding dim.
Buffalo-L	ResNet-50 / w600k_r50 (InsightFace ONNX; WebFace600K)	ArcFace additive angular margin [2]	512
FaceNet	InceptionResNetV1	Triplet loss; VGGFace2 [21]	512
MagFace	iResNet-50	Magnitude-aware angular margin; MS1MV2 [20]	512
AdaFace	IR-50	Quality-adaptive angular margin; MS1MV2 [20]	512
Dlib	Compact ResNet + 5-point landmarks	Metric-learning descriptor [10]	128

Table 2. Recognizer backbones and training objectives evaluated under the shared Baseline B pipeline.

E. Evaluation metrics

At the fixed threshold τ = 0.40, the false accept and false reject rates are

$$\mathrm{FAR}(\tau) = \frac{FP(\tau)}{N} \,, \qquad \mathrm{FRR}(\tau) = \frac{FN(\tau)}{P}$$

Accuracy, precision, and recall follow the standard confusion-matrix definitions, and F1 is their harmonic mean, Section IV reports Acc@0.40, AUC, and EER; precision, recall, and F1 are not tabulated separately.

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \,, \quad \mathrm{Prec} = \frac{TP}{TP + FP} \,, \quad \mathrm{Rec} = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{\mathrm{Prec} \cdot \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}$$

Threshold-independent ranking quality is summarized by the area under the ROC curve and by the equal error rate, the operating point at which FAR and FRR coincide,

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(x))\,dx \;=\; P(\mathrm{sim}_{\cos}(\mathrm{pos}) > \mathrm{sim}_{\cos}(\mathrm{neg}))$$

$$\mathrm{EER} = \mathrm{FAR}(\tau^*) = \mathrm{FRR}(\tau^*)\ , \quad \tau^* = \arg\min_{\tau} |\,\mathrm{FAR}(\tau) - \mathrm{FRR}(\tau)\,|$$

For the closed-set identification experiments in Section IV-E, Rank-k accuracy over a probe set Q with gallery size N is

$$\text{Rank-}k\ \mathrm{Acc} = \frac{1}{|Q|}\sum_{q \in Q} \mathbb{1}\left[\mathrm{rank}(q) \le k\right]$$

where rank(q) is the position of the true gallery match for probe q in the similarity-sorted candidate list; this study reports Rank-1 (k = 1) at N ∈ {10, 100, 500, 1000} for LFW and N ∈ {10, 100, 500} for the supplementary TinyFace probe. Uncached embedding throughput is reported as

$$\mathrm{FPS} = \frac{N_{\mathrm{frames}}}{T_{\mathrm{embed}}}$$

measured on the same RetinaFace/SCRFD crops used for verification, over N_frames = 200 timed images after a warm-up pass, so that all five models are timed on an identical input distribution.

F. Statistical testing

To assess whether the accuracy gap between Buffalo-L (the leading model on every HQ set; Section IV-A) and each other model is statistically distinguishable rather than an artifact of a single run, this study applies McNemar's test [17] to the paired per-pair correct/incorrect outcomes of Buffalo-L against each of the other four models, on every HQ dataset. With continuity correction, the test statistic is

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

where n_{01} is the number of pairs Buffalo-L classifies correctly and the comparison model classifies incorrectly, and n_{10} is the reverse; χ^2 is compared against a chi-squared distribution with one degree of freedom to obtain a p-value. McNemar's test is the appropriate paired test here because both models are scored on the identical set of pairs under the identical Baseline B crop, so the comparison of interest is within-pair disagreement rather than a comparison of two independent accuracy samples. The reported McNemar n on each dataset is the five-model common scored subset (the intersection of pairs successfully embedded by all five models), which on pose-heavy CFP-FP and CPLFW is limited by Dlib's landmark-alignment skips (n = 3,191 and n = 2,983 respectively) and is therefore not the same per-model scored count underlying Table 3.

4. Results

A. Verification accuracy

Table 3: reports Acc@0.40 for all five models on each of the four HQ verification sets and their mean.

Model	LFW	CFP-FP	CPLFW	AgeDB-30	Mean
Buffalo-L	0.9828	0.9094	0.8605	0.8653	0.9045
FaceNet	0.9758	0.8822	0.7939	0.7767	0.8572
MagFace	0.8864	0.5125	0.5479	0.7232	0.6675
AdaFace	0.8098	0.5007	0.5381	0.5462	0.5987
Dlib	0.5009	0.5080	0.4861	0.4993	0.4986

Table 3. Accuracy at cosine 0.40 on four HQ verification sets (TinyFace excluded; see Section IV-F).

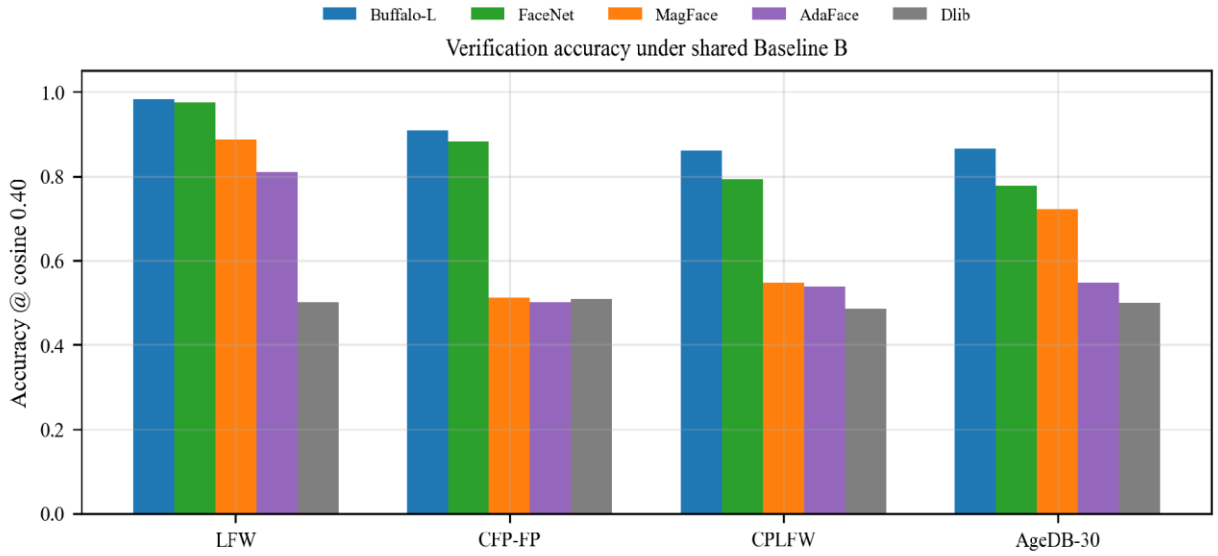

Figure 2: Verification accuracy under shared Baseline B, grouped by dataset.

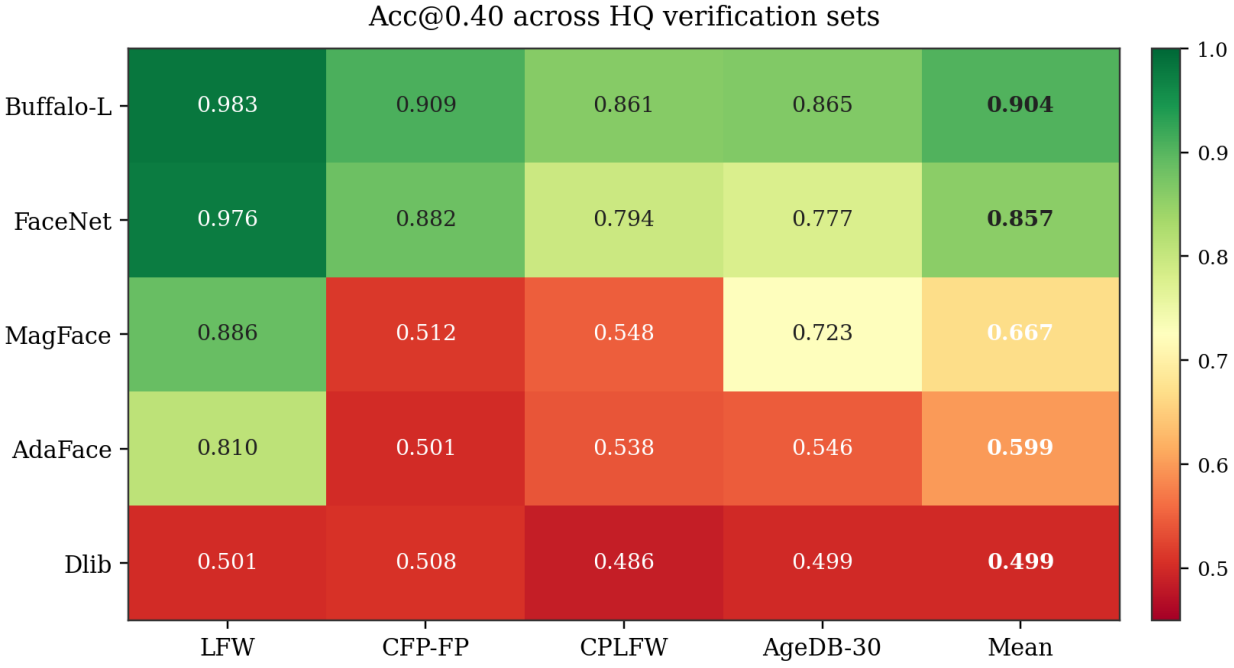

Figure 3: Acc@0.40 across the four HQ sets and their mean, as a heat map. Buffalo-L and FaceNet form a clearly separated top tier; MagFace and AdaFace show markedly lower accuracy specifically on the pose-heavy CFP-FP and CPLFW sets.

Dlib's skip rate is elevated on the pose-heavy sets — 3,809 of 7,000 CFP-FP pairs and 3,017 of 6,000 CPLFW pairs are skipped — because its five-point landmark aligner fails on a large fraction of profile crops. Buffalo-L leads Acc@0.40 on every HQ set without exception. The AdaFace and MagFace accuracy drops on CFP-FP and CPLFW are Baseline B pipeline effects specific to the shared bounding-box-margin crop, not a reproduction of the native-align state-of-the-art numbers reported for these methods in their original papers [3], [4]. Dlib's Acc@0.40 ≈ 0.50 on LFW, with a false-accept rate near 1, occurs despite an AUC of 0.9922 and an EER of 0.0234 — the clearest single illustration in this study that fixed-threshold accuracy and ranking quality are not the same property of a recognizer.

B. ROC-AUC and equal error rate

Table 4: extends the accuracy comparison with threshold-independent ranking quality on every HQ dataset.

Model	LFW AUC / EER	CFP-FP AUC / EER	CPLFW AUC / EER	AgeDB-30 AUC / EER
Buffalo-L	0.9899 / 0.0216	0.9977 / 0.0073	0.9644 / 0.0641	0.9904 / 0.0245
FaceNet	0.9933 / 0.0275	0.9800 / 0.0673	0.8653 / 0.2098	0.9154 / 0.1623
Dlib	0.9922 / 0.0234	0.9661 / 0.0924	0.8759 / 0.1931	0.9760 / 0.0752
AdaFace	0.9711 / 0.0730	0.7553 / 0.3187	0.6254 / 0.4097	0.8786 / 0.2000
MagFace	0.9515 / 0.1079	0.6943 / 0.3617	0.5637 / 0.4533	0.9062 / 0.1750

Table 4. ROC-AUC / equal error rate (EER) on the four HQ sets.

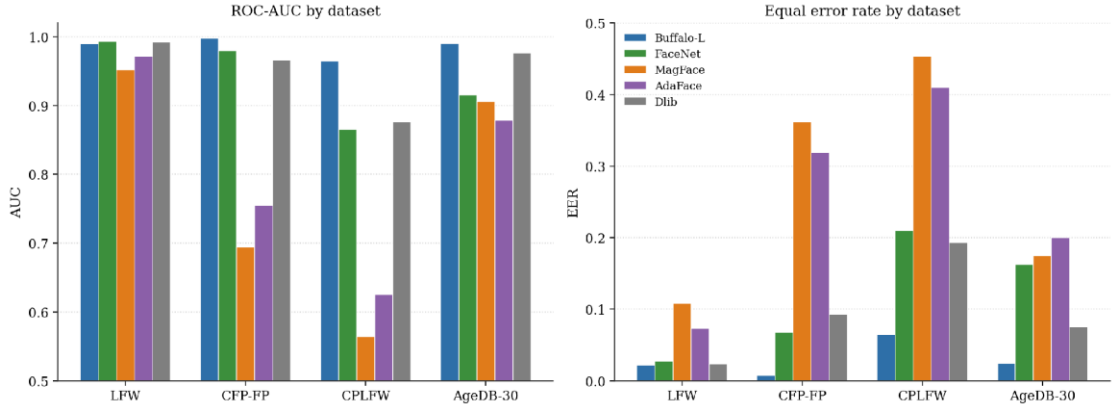

Figure 4: AUC (left) and EER (right) by dataset. FaceNet and Dlib are competitive with, or exceed, Buffalo-L in pure ranking quality on LFW despite trailing badly in Table 3's fixed-threshold accuracy — confirming that the accuracy gap for these two models is substantially a calibration effect rather than an embedding-quality effect, while AdaFace and MagFace show a genuine ranking-quality drop on CFP-FP and CPLFW, not merely a threshold problem.

Buffalo-L's AUC leads on three of four datasets and is a close second on LFW (0.9899 versus FaceNet's 0.9933 and Dlib's 0.9922), and it holds the lowest EER on every HQ dataset, including CFP-FP (0.0073 versus FaceNet's 0.0673). AdaFace and MagFace show the steepest AUC decline moving from LFW to CPLFW (0.9711→0.6254 and 0.9515→0.5637 respectively), which is consistent with the alignment-sensitivity argument in Section V: cross-pose faces amplify the mismatch between the shared bbox-margin crop and each model's native, quality-adaptive alignment target.

C. Statistical significance

Table 5: Reports McNemar p-values comparing Buffalo-L against each other model, per dataset, using the paired per-pair correct/incorrect outcomes described in Section III-F.

Dataset (n scored)	vs. FaceNet	vs. Dlib	vs. AdaFace	vs. MagFace
LFW (n = 5,893)	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
AgeDB-30 (n = 5,878)	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
CFP-FP (n = 3,191)	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
CPLFW (n = 2,983)	$p = 4.2\times10^{-5}$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table 5. McNemar p-values at cosine 0.40, Buffalo-L versus each model, on the five-model common scored subset per dataset (Dlib-limited on CFP-FP and CPLFW; not the per-model n behind Table 3).

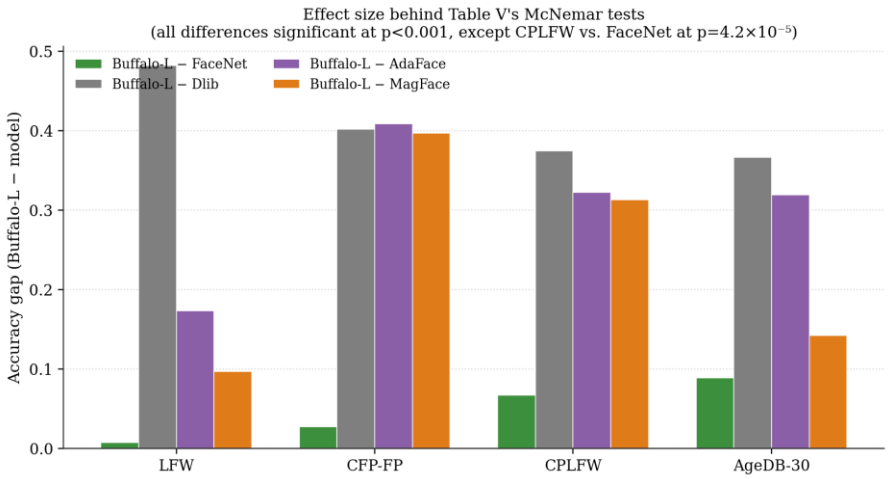

Figure 5: Accuracy gap (Buffalo-L minus each comparison model) underlying Table 5. Every gap is statistically significant at $p < 0.001$; the closest is CPLFW versus FaceNet ($p = 4.2\times10^{-5}$; accuracy gap of 0.0666).

Every comparison in Table 5 is significant at conventional thresholds, including the numerically smallest gap in the entire study (Buffalo-L vs. FaceNet on LFW, an accuracy gap of only 0.0070, still significant at $p < 0.001$ given $n = 5{,}893$ paired trials). The CPLFW-vs-FaceNet result ($p = 4.2\times10^{-5}$) is reported to full precision specifically because it is the tightest contest observed on any dataset, and a reader comparing Buffalo-L to FaceNet specifically for pose-robust deployment should weigh that smaller margin of statistical confidence against the larger, more decisive gaps Buffalo-L holds over Dlib, AdaFace, and MagFace on the same dataset.

D. Computational cost

Model	Load (s)	Embed (s)	FPS	Memory (MiB)
MagFace	≈0	0.0102	97.60	374.8
AdaFace	≈0	0.0108	92.83	667.8
FaceNet	1.31	0.0200	49.98	—
Buffalo-L	1.56	0.0362	27.63	600.7
Dlib	0.34	0.2633	3.80	56.0

Table 6: Uncached embedding cost (NVIDIA RTX 3050, N = 200 timed crops after warm-up).

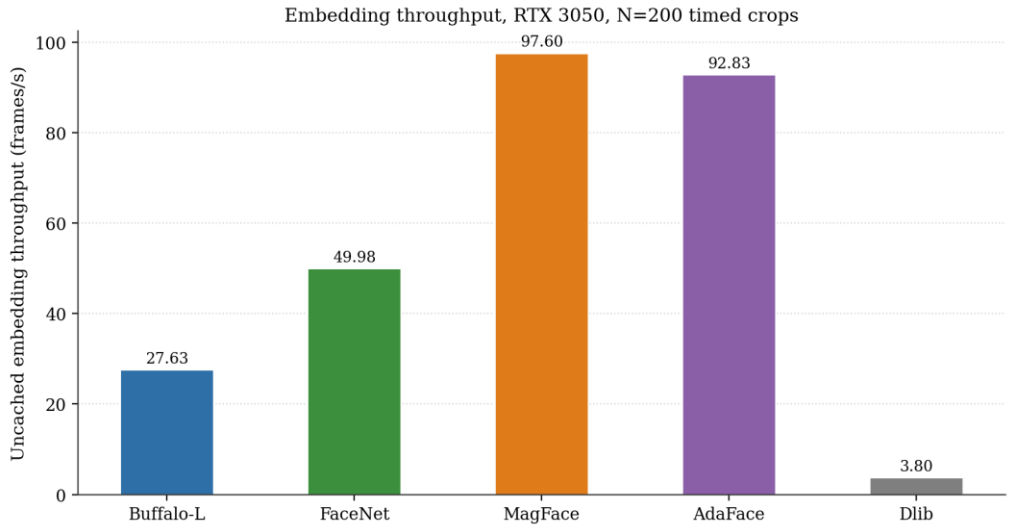

Figure 6: Uncached embedding throughput by model. MagFace and AdaFace are roughly 25× and 24× faster than Dlib, and about 3.5× faster than Buffalo-L, on identical RetinaFace/SCRFD crops.

Dlib's throughput (3.80 FPS) is substantially below every deep-learning-based model, consistent with its landmark-heavy CPU-oriented pipeline and its elevated detection-skip rate on pose-heavy sets (Section IV-A). MagFace's speed advantage coincides with the lowest recorded model size among the ONNX/PyTorch models with a reported value (374.8 MiB versus 600.7 MiB for Buffalo-L and 667.8 MiB for AdaFace). AdaFace and MagFace load times of approximately 0 s in Table 6 are not comparable cold starts: those backbones were pre-injected already built, whereas FaceNet (1.31 s) and Buffalo-L (1.56 s) include pack load. Together with Section IV-A's accuracy figures, these costs set up the accuracy–efficiency trade-off formalized in Section IV-G.

E. Identification versus gallery size

Table 7: reports LFW closed-set Rank-1 identification accuracy as gallery size N grows from 10 to 1,000 identities, isolating how verification-style accuracy (Section IV-A, a fixed pair-count task) degrades under an identification-style search whose difficulty scales with N.

Model	**N = 10**	**N = 100**	**N = 500**	**N = 1000**
Buffalo-L	1.000	0.965	0.988	0.986
Dlib	1.000	0.930	0.940	0.913
FaceNet	1.000	0.930	0.913	0.882
AdaFace	1.000	0.789	0.678	0.691
MagFace	1.000	0.737	0.483	0.531

Table 7. LFW Rank-1 identification accuracy versus gallery size N (not NIST FRVT operational scale).

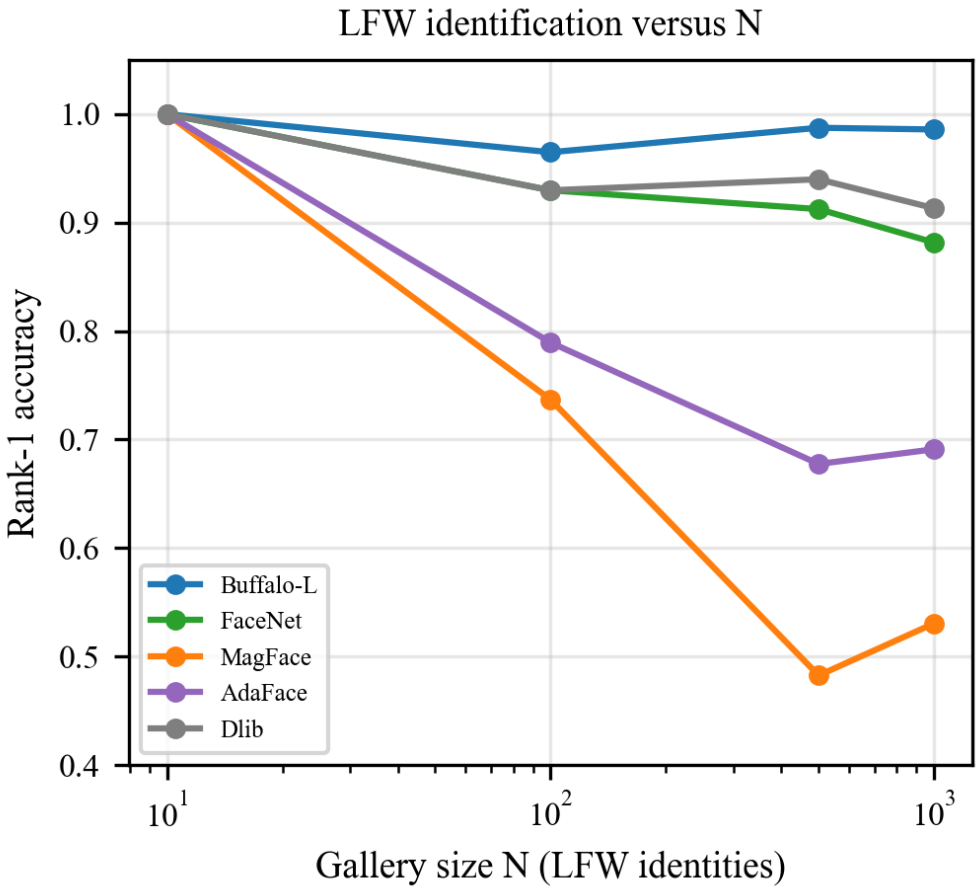

Figure 7: LFW Rank-1 accuracy versus gallery size N (log scale). Buffalo-L is the only model that stays above 0.96 across the full N = 10–1,000 range; MagFace and AdaFace lose more than 30 and 25 percentage points respectively between N = 10 and N = 500.

All five models identify correctly at N = 10, which is expected given the small candidate pool. The ranking that emerges by N = 500 (Buffalo-L > Dlib ≈ FaceNet > AdaFace > MagFace) is a substantially different ordering from the pure verification ranking in Table 3, where Dlib places last; a model with a poorly calibrated fixed-threshold operating point (Dlib) can still separate genuine top-1 matches from a large gallery of impostors well, because identification depends only on relative similarity ranking against the gallery, not on a global decision threshold. MagFace's non-monotonic dip at N = 500 followed by a partial recovery at N = 1,000 (0.483 → 0.531) falls within the range plausibly attributable to which specific identities enter the gallery at each sampled N, and should not be over-interpreted as a real non-monotonic capability.

F. Supplementary TinyFace Rank-1

TinyFace results are reported separately and are not mixed into the Table 3 HQ means or the Table 9 Pareto accuracy figures. The protocol used is FaceBench's own CMC computation on Gallery_Match versus Probe only; N = 1,000 was skipped because, of 3,134 unique images drawn for an N = 1,000 gallery attempt, 3,042 aligned and 92 failed; after RetinaFace filtering, only 990 gallery identities remained, leaving an insufficient gallery for a meaningful N = 1,000 draw. This is explicitly not the official TinyFace MATLAB mAP protocol [19].

Model	N = 10	N = 100	N = 500
Buffalo-L	1.000	0.889	0.819
FaceNet	1.000	0.810	0.675
Dlib	0.778	0.776	0.644

Model	N = 10	N = 100	N = 500
AdaFace	0.769	0.587	0.382
MagFace	0.615	0.508	0.337

Table 8: Supplementary TinyFace Rank-1 (FaceBench CMC on Gallery_Match vs. Probe; not official TinyFace mAP).

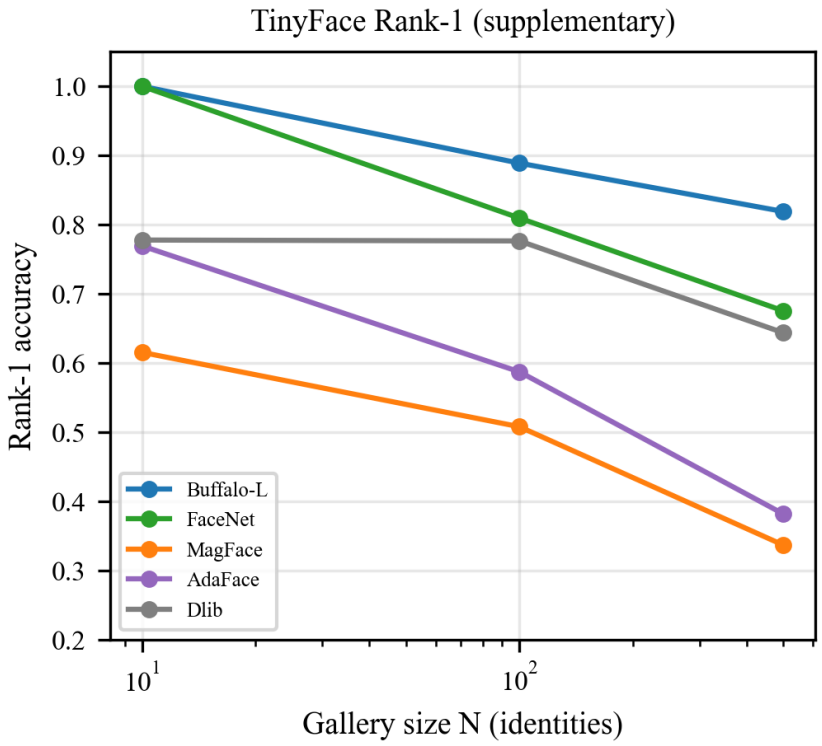

Figure 8: Supplementary TinyFace Rank-1 versus gallery size N. The five-model ordering at N = 500 (Buffalo-L > FaceNet > Dlib > AdaFace > MagFace) largely mirrors Table 7's LFW ordering, with Dlib performing relatively better here than on LFW verification, plausibly because native low-resolution probes are less dependent on the fine landmark alignment that fails on CFP-FP/CPLFW pose extremes.

Unlike the HQ sets, Dlib does not lead or trail as sharply on TinyFace: at N = 10, Dlib (0.778) already trails Buffalo-L and FaceNet (both 1.000) but clearly outperforms AdaFace (0.769, comparable) and MagFace (0.615), a relative position it roughly holds through N = 500. This is consistent with TinyFace probes being uniformly low-resolution — removing the profile-pose alignment failures that specifically hurt Dlib on CFP-FP and CPLFW — rather than with any change in Dlib's underlying embedding quality.

G. Accuracy–efficiency synthesis

Model	Mean Acc (HQ)	Worst-case Acc (HQ)	FPS
Buffalo-L	0.9045	0.8605	27.63
FaceNet	0.8572	0.7767	49.98
MagFace	0.6675	0.5125	97.60
AdaFace	0.5987	0.5007	92.83
Dlib	0.4986	0.4861	3.80

Table 9: HQ mean and worst-case Acc@0.40 versus uncached embedding throughput.

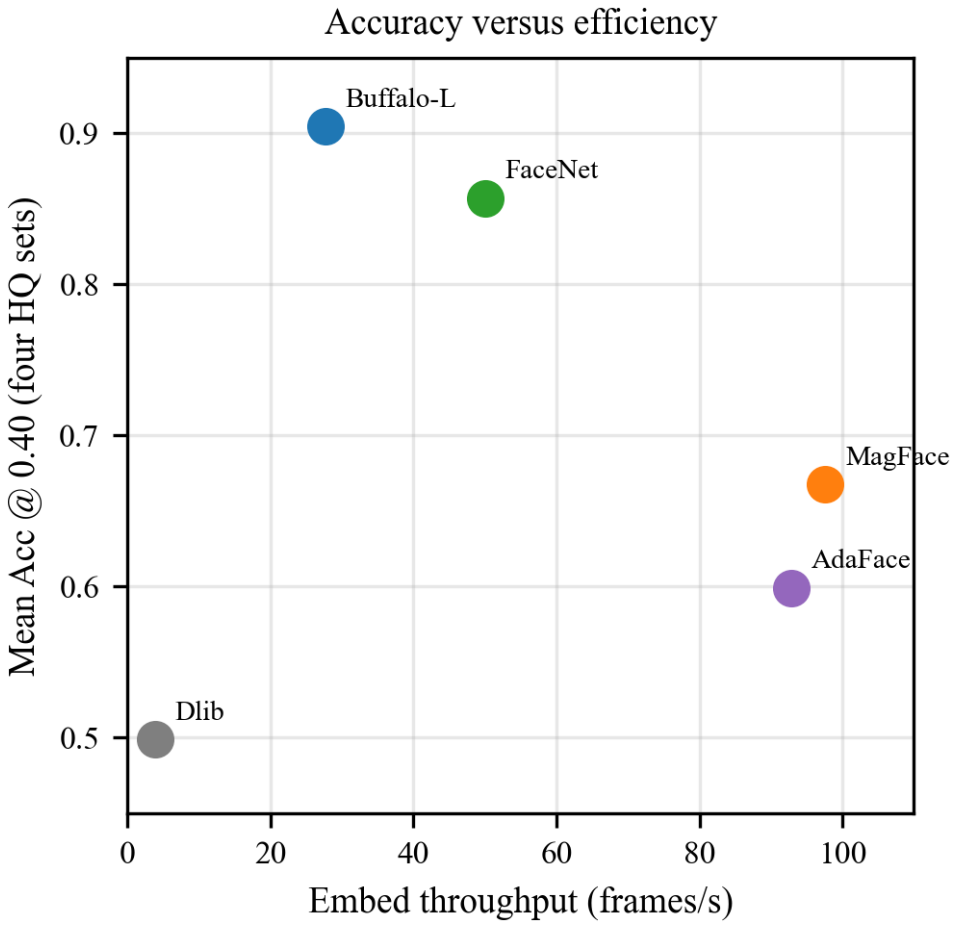

Figure 9: Mean HQ Acc@0.40 versus embedding throughput. Buffalo-L and FaceNet dominate the upper-left accuracy-leaning region; MagFace and AdaFace dominate the lower-right efficiency-leaning region; Dlib is dominated on both axes simultaneously by every other model in this study.

Buffalo-L is accuracy-leaning: it holds both the highest mean and the highest worst-case HQ accuracy, at roughly 3.5× the throughput cost of MagFace. MagFace and AdaFace are efficiency-leaning, trading about 24 (MagFace) and 31 (AdaFace) mean HQ accuracy percentage points versus Buffalo-L for 3.4–3.5× the throughput of Buffalo-L, but their worst-case accuracy (0.5125 and 0.5007) sits close to chance, which matters more for a deployment that cannot guarantee frontal, HQ-only input. FaceNet occupies an intermediate position on both axes. Dlib is dominated on both axes simultaneously by every other model in this comparison and is retained in the study specifically as the clearest illustration of the threshold-versus-ranking distinction developed in Sections IV-A, IV-B, and V, not as a competitive candidate on this Pareto frontier. All throughput figures reflect a single RTX 3050 workstation, not edge or embedded hardware, and this paper does not recommend a specific deployment model from the accuracy–efficiency numbers alone.

5. Discussion and Limitations

A. What the shared crop isolates, and what it does not

Locking RetinaFace/SCRFD detection and the bounding-box-margin crop once per image, and reusing it across all five recognizers, controls detector choice and crop geometry across models as experimental factors when interpreting the accuracy gaps in Table 3: no model benefits from a friendlier detector than any other. What the shared crop does not remove is each model's sensitivity to deviating from its own native alignment recipe. AdaFace and MagFace were trained and originally reported under ArcFace-style, tightly cropped 112×112 alignment; the wider,

less tightly centered Baseline B crop is a harder input for both, and the CFP-FP/CPLFW accuracy drops in Table 3, together with the matching AUC drops in Table 4, are best read as alignment-sensitivity rather than as a deficiency in either method's underlying loss function.

B. Ranking quality versus fixed-threshold accuracy

Dlib is the clearest demonstration in this study that a single accuracy number at one shared threshold can misrepresent a recognizer. Its AUC and EER (Table 4) are competitive with, or better than, three of the other four models on most datasets, yet its Acc@0.40 (Table 3) is at or near chance level everywhere. Because the fixed threshold of 0.40 is poorly calibrated for Dlib's similarity distribution, the fixed threshold used throughout this study accepts nearly every pair it is shown, positive or negative. Table 7's identification results reinforce the same point from a different angle: Dlib's Rank-1 accuracy at N = 100–500 is competitive with FaceNet's, because identification depends on relative ranking within a gallery rather than on a single global threshold. Any benchmarking exercise that reports only fixed-threshold accuracy for multiple recognizers, without also reporting AUC, EER, or an identification-style ranking metric, risks the same misranking a naive reading of Table 3 alone would produce for Dlib.

C. Statistical versus practical significance

Table 5 shows that every accuracy gap against Buffalo-L, however small, reaches conventional statistical significance, which is a direct consequence of the paired sample sizes involved (thousands of scored pairs per dataset) rather than evidence that every gap is practically meaningful for a deployment decision. The smallest gap in the study, Buffalo-L versus FaceNet on LFW (0.0070, Table 3), is statistically significant but may not justify a backbone switch on its own once factors such as licensing, embedding dimensionality (512 for both, Table 2), and downstream integration cost are considered; the larger gaps against Dlib, AdaFace, and MagFace (0.30–0.48 on at least one dataset each) are both statistically and practically decisive.

D. Accuracy–efficiency trade-offs are workload-dependent

Section IV-G's Pareto reading should not be reduced to “Buffalo-L is best.” For a workload dominated by frontal, HQ input and where accuracy dominates the deployment cost function, Buffalo-L's position is close to strictly dominant. For a workload with a hard latency or memory budget and a tolerance for lower worst-case accuracy — for example, a coarse pre-filtering stage ahead of a more expensive verifier — MagFace or AdaFace's throughput advantage (Table 6) may outweigh their accuracy cost. FaceNet's intermediate position on both axes makes it a reasonable default when neither extreme is required. These are workload-dependent judgments that this paper deliberately leaves to the reader rather than resolving with a single recommended model, consistent with the single-workstation scope noted in Section IV-G.

E. Limitations and threats to validity

- Dataset scope: this study covers LFW, CFP-FP, CPLFW, and AgeDB-30 as HQ evidence, plus a supplementary TinyFace probe; occlusion, surveillance-camera, and video-based protocols were not run, and the ranking observed here may not transfer to those conditions.
- Fixed, non-model-specific threshold: $\tau = 0.40$ was locked once for the entire study rather than calibrated per model. Section V-B argues this is a deliberate and informative design choice, but a per-model EER-calibrated threshold would likely narrow several of Table 3's accuracy gaps, most visibly Dlib's.
- Shared crop versus native alignment: the bbox-margin crop (margin = 0.35) is one compromise geometry, not any model's native alignment target, which plausibly disadvantages AdaFace and MagFace specifically on pose-heavy sets (Section V-A).
- TinyFace protocol divergence: the supplementary TinyFace probe uses FaceBench's own CMC Rank-1 computation on Gallery_Match versus Probe only, with N = 1,000 skipped after RetinaFace filtering of 3,134 unique images (3,042 aligned / 92 failed), leaving 990 gallery identities; it is not the official TinyFace MATLAB mAP protocol and is not combined with the HQ means at any point.
- CPLFW fold interpretation: CPLFW AUC after common-N truncation with Dlib's elevated skip rate should not be over-interpreted as a like-for-like comparison across models, since the effective scored subset differs by model on this dataset specifically.
- Hardware and run scope: absolute FPS figures (Table 6) reflect a single NVIDIA RTX 3050 workstation and a single run per configuration; they are not edge-device or multi-run averaged figures, and this paper does not recommend a specific deployment model from throughput numbers alone.

6. Conclusion

Under locked Baseline B preprocessing, Buffalo-L delivers the strongest Acc@0.40 across LFW, CFP-FP, CPLFW, and AgeDB-30 among the five models studied, with FaceNet second and MagFace/AdaFace trading accuracy for substantially higher embedding throughput. Every accuracy gap against Buffalo-L is statistically significant under McNemar's test, though the smallest gaps (notably against FaceNet on LFW and CPLFW) are closer contests than the headline accuracy ranking alone suggests. Dlib demonstrates concretely that ranking-quality metrics (AUC, EER) and fixed-threshold accuracy can diverge sharply for the same model and the same embeddings, and that identification-style Rank-k accuracy can partially recover a model whose verification-style accuracy looks weak at one shared threshold. A supplementary TinyFace Rank-1 probe follows a broadly similar model ordering at N = 500 but is explicitly not official TinyFace mAP and is not mixed into the HQ evidence. Future work includes per-model EER-calibrated thresholds reported alongside the shared threshold used here, a native-alignment companion arm for AdaFace and MagFace, broader coverage of occlusion and video protocols, and multi-run replication to characterize run-to-run variability beyond the single-workstation, single-seed scope of this study.

Data and Code Availability

Face images are used under each dataset's original terms and are not redistributed. Source code and experiment manifests are released under the MIT License at https://github.com/NehalPatel/FaceBenchX. HQ experiment identifiers referenced in this study: 20260806T192706Z_..._f86260f8 (LFW), 20260812T194624Z_..._0e72f05d (AgeDB-30), 20260812T210334Z_..._e8c4a652 (CFP-FP), and 20260812T222132Z_..._61e1df5c (CPLFW).

Reference

1. F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE CVPR, 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.
2. J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE/CVF CVPR, 2019, pp. 4690–4699, doi: 10.1109/CVPR.2019.00482.
3. Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “MagFace: A universal representation for face recognition and quality assessment,” in Proc. IEEE/CVF CVPR, 2021, pp. 14225–14234, doi: 10.1109/CVPR46437.2021.01400.
4. M. Kim, A. K. Jain, and X. Liu, “AdaFace: Quality adaptive margin for face recognition,” in Proc. IEEE/CVF CVPR, 2022, pp. 18750–18759, doi: 10.1109/CVPR52688.2022.01819.
5. G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled Faces in the Wild,” Univ. Massachusetts, Amherst, Tech. Rep. 07-49, 2007.
6. S. Sengupta et al., “Frontal to profile face verification in the wild,” in Proc. IEEE WACV, 2016, doi: 10.1109/WACV.2016.7477558.
7. T. Zheng and W. Deng, “Cross-Pose LFW,” Beijing Univ. Posts and Telecommun., 2018.
8. S. Moschoglou et al., “AgeDB: The first manually collected, in-the-wild age database,” in Proc. IEEE CVPR Workshops, 2017, doi: 10.1109/CVPRW.2017.250.
9. J. Deng et al., “RetinaFace: Single-shot multi-level face localisation in the wild,” in Proc. IEEE/CVF CVPR, 2020, doi: 10.1109/CVPR42600.2020.00525.
10. D. E. King, “Dlib-ml: A machine learning toolkit,” J. Mach. Learn. Res., vol. 10, pp. 1755–1758, 2009.
11. S. I. Serengil and A. Ozpinar, “LightFace: A hybrid deep face recognition framework,” in Proc. ASYU, 2020, doi: 10.1109/ASYU50717.2020.9259802.
12. Q. Zhao et al., “face.evoLVe,” arXiv:2107.08621, 2021.
13. J. Wang et al., “FaceX-Zoo: A PyTorch toolbox for face recognition,” in Proc. ACM MM, 2021, doi: 10.1145/3474085.3478324.
14. B. Sowan, A. Ibrahim, and A. Keshlaf, “Facial recognition accuracy and verification time comparison between ArcFace and Facenet models on DeepFace framework,” in Proc. IEEE MI-STA, 2026, doi: 10.1109/mi-sta68962.2026.11511304.
15. L. Friedman et al., “Biometric performance as a function of gallery size,” Appl. Sci., vol. 12, no. 21, p. 11144, 2022, doi: 10.3390/app122111144.
16. P. Grother, M. Ngan, and K. Hanaoka, “Ongoing FRVT Part 2: Identification,” NISTIR 8238, 2018, doi: 10.6028/NIST.IR.8238.
17. Q. McNemar, Psychometrika, vol. 12, no. 2, pp. 153–157, 1947, doi: 10.1007/BF02295996.
18. W. S. Yambor, B. A. Draper, and J. R. Beveridge, “Analyzing PCA-based face recognition algorithms: Eigenvector selection and distance measures,” in Empirical Evaluation Methods in Computer Vision, 2002.
19. Z. Cheng, X. Zhu, and S. Gong, “Low-resolution face recognition,” in Proc. ACCV, 2019, doi: 10.1007/978-3-030-20887-5_38.
20. Y. Guo et al., “MS-Celeb-1M,” arXiv:1607.08221, 2016.
21. Q. Cao et al., “VGGFace2,” in Proc. IEEE FG, 2018, doi: 10.1109/FG.2018.00020.
22. S. Shahi et al., “A Cross-Dataset Study on Controlled, Wild, Demographic, Age-Variant, Masked, and Synthetic Faces,” SGS — Engineering & Sciences, vol. 1, no. 4, 2025.

23. J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, “Sample and computation redistribution for efficient face detection,” in Proc. ICLR, 2022.
24. J. Deng et al., “Lightweight face recognition challenge,” in Proc. IEEE/CVF ICCVW, 2019, doi: 10.1109/ICCVW.2019.00327.