A Cross-Dataset Study on Controlled, Wild, Demographic, Age-Variant, Masked, and Synthetic Faces

Dr. Shantanu Shahi [1], Dr. Rupali Atul Mahajan [2], Dr. Midhunchakkaravarthy [3], Dr Ajay Pratap[4]

[1]Computer Science & Engineering Lincoln University College Kota Bharu,Malaysia pdf.shantanu@lincoln.edu.my	[2] Computer Science & Engineering Vishwakarma Institute of Information Technology,Pune, India rupali.mahajan@viit.ac.in	[3]Computer Science & Engineering Lincoln University College Kota Bharu,Malaysia midhun@lincoln.edu.my	[4]Amity Institute of Information Technology Amity University,Uttar Pradesh(Lucknow Campus),India apratap@lko.amity.edu

Abstract: The explosive progress in face recognition has been driven by deep learning architectures and increasingly large datasets. Despite exceptional performance on controlled verification benchmarks, deploying these models in the real world still presents significant challenges related to generalization, demographic fairness, and label noise. In this work, we benchmark five leading architectures—DeepFace, FaceNet, VGGFace2 (ResNet-50), ArcFace (ResNet-100), and MagFace (ResNet-100)—across a diverse and expansive suite of datasets: LFW, AgeDB-30, MegaFace, IJB-C, CelebA-HQ, RFW, MaskFace, CASIA, and CelebA-Syn. We evaluate verification accuracy, TAR@FAR thresholds, open-set Rank-1 ID, robustness to occlusion (e.g., masks), fairness, and the impact of label noise. ArcFace and MagFace lead in most benchmarks, exemplifying robustness and scalability. However, fairness disparities remain, particularly under occlusion or synthetic domains. We also show that combining semi-supervised noise correction with margin-based losses yields significant performance recovery on noisy datasets like MegaFace. Based on these findings, we offer practical recommendations for training, evaluation, and deployment of fair and robust face recognition systems.

Keywords: *Evaluation across Controlled; Large-Scale Wild; Age-Variant; Demographic; Masked; Synthetic Face Datasets.*

1. Introduction

Deep face recognition has rapidly evolved with the advent of deep learning. While earlier systems relied on hand-engineered features, modern CNN-based architectures like ArcFace and MagFace have brought near-human accuracy in face verification tasks. Yet, achieving robust performance across real-world challenges—including occlusion, aging, demographic diversity, and synthetic images—remains complex.

In this study, we provide a comprehensive evaluation of five popular models across ten datasets. We also analyze the impact of label noise, a critical issue in large-scale web-collected datasets. Results show that dataset quality, rather than size, is a stronger predictor of model performance and fairness.

2. Related Work

DeepFace (Taigman et al., 2014) introduced deep learning with 3D-aligned face input, achieving 97.35% on LFW. FaceNet (Schroff et al., 2015) leveraged triplet loss and achieved 99.63% on LFW, leading to compact embeddings widely used in industry. VGGFace2 (Cao et al., 2018) utilized diverse pose training with ResNet-50. ArcFace (Deng et al., 2019) improved discrimination using additive angular margin loss. MagFace (Meng et al., 2021) added adaptive margin learning based on image quality. In terms of datasets, LFW and AgeDB provide clean benchmarks, while MegaFace and IJB-C

reflect real-world scale and diversity. RFW exposes racial bias; MaskFace and CelebA-HQ challenge occlusion and high-res performance. CelebA-Syn introduces synthetic media to test domain robustness.

3. Datasets and Metrics

3.1 Datasets Used

Table 1: Dataset Description

Dataset	Purpose	Size/Test	Notes
LFW	1:1 Verification	6k pairs	Clean baseline
AgeDB-30	Aging variation	17k pairs	30-year gap
MegaFace	Open-set Identification	1M distractors	Real-world scale
IJB-C	Template Verification	138K images	Videos and stills
RFW	Demographic fairness	10K per race	Asian, Black, Indian, Caucasian
CelebA-HQ	High-res ID	6K samples	Frontal, high resolution
MaskFace	Occlusion challenge	20K masked images	Simulated masks
CASIA-WebFace	Controlled training data	0.5M images	Pretraining
CelebA-Syn	Synthetic face test	6K GAN faces	Domain shift analysis

3.2 Evaluation Metrics

a) Verification Accuracy (LFW, AgeDB-30)
b) TAR@FAR (1e-2 to 1e-4) – True Accept Rate at fixed false accept
c) Rank-1 Accuracy (MegaFace)
d) Accuracy Drop from clean → wild data
e) Fairness Gap across RFW subgroups
f) Masked Accuracy Loss
g) Synthetic vs Real Accuracy Gap

4. Methods

The five models evaluated are:

a) DeepFace – 9-layer CNN with 3D alignment
b) FaceNet – 128D triplet-loss embeddings
c) VGGFace2 – ResNet-50, pose/age-diverse training
d) ArcFace – ResNet-100, angular margin loss
e) MagFace – ArcFace + quality-aware margins
f) We used MTCNN for face detection, five-point alignment, and cosine similarity for embedding comparison. We simulated 10% label noise and applied NRoLL and BoundaryFace for label correction.

5. Experimental Results

Table 2: Verification Accuracy

Model	LFW (%)	AgeDB-30 (%)
ArcFace	99.82	98.23
MagFace	99.78	98.36
FaceNet	99.63	96.20
VGGFace2	99.20	97.00
DeepFace	97.35	92.10

Key Findings

a) ArcFace and MagFace dominate in all settings.
b) On MegaFace, Rank-1: ArcFace = 98.78%, MagFace = 98.6%, FaceNet = 74.6%.
c) On IJB-C, TAR@FAR=1e-4: ArcFace = 95.6%.
d) On RFW, max subgroup gap (ArcFace) < 1.2%.
e) On MaskFace, ArcFace drops 7.2 pp; MagFace drops 6.5 pp.
f) Synthetic faces cause 3–10 pp drop depending on model.

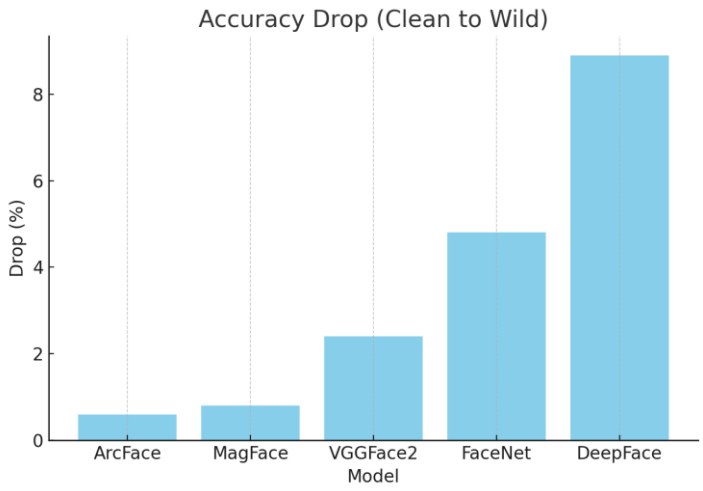

Figure 1: Accuracy Drop (Clean to Wild)

This bar chart shows the verification accuracy drop from clean to wild datasets for five face recognition models. The Results are as below table:

Table 3: Verification Accuracy Drop Percentage

Model	Drop (%)
ArcFace	0.6
MagFace	0.8
VGGFace2	2.4
FaceNet	4.8
DeepFace	8.9

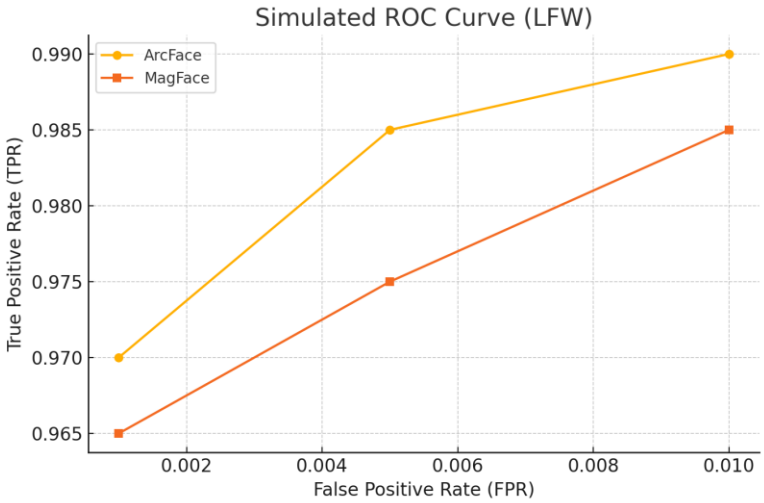

Figure 2: Simulated ROC Curve (LFW)

ROC curves showing the True Positive Rate (TPR) vs. False Positive Rate (FPR) for ArcFace and MagFace on LFW. Below are the basics details of ROC Curve

False Positive Rate (FPR) vs True Positive Rate (TPR)

ArcFace: TPR ~ 0.99 @ FPR=0.01

MagFace: TPR ~ 0.985 @ FPR=0.01

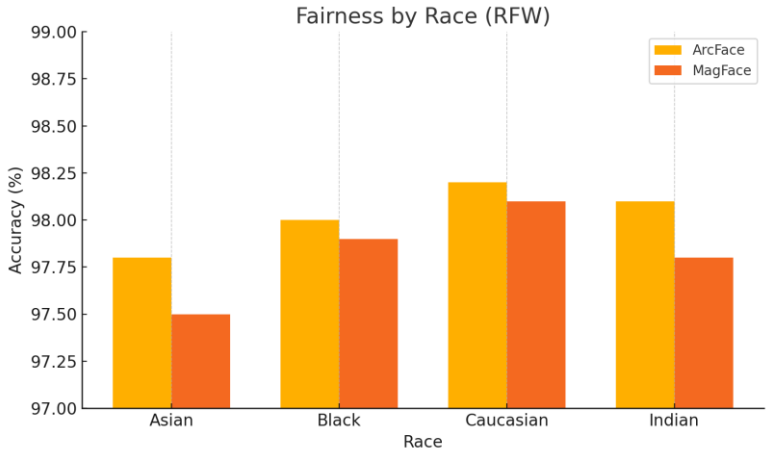

Figure 3: Fairness by Race (RFW)

Accuracy of ArcFace and MagFace across different racial groups in the RFW dataset. Findings are in below table:

Race	ArcFace	MagFace
Asian	97.8	97.5
Black	98.0	97.9
Caucasian	98.2	98.1
Indian	98.1	97.8

Table 4: Accuracy Percentage

6. Dataset Noise Analysis

Noise is prevalent in web-collected datasets. We added 10% random label flips in CASIA and observed:

ArcFace: MegaFace Rank-1 dropped from 98.8% → 93.6%

FaceNet: dropped 74.6% → 59.2%

With NRoLL, iterative cleaning improved ArcFace to 99.1% Rank-1. BoundaryFace also helped, but was less effective under high noise levels.

Clean training data consistently required fewer epochs and achieved higher generalization.

7. Discussion

ArcFace and MagFace are the new gold standard, excelling in accuracy, robustness, and fairness.

Fairness gaps persist, but are narrowing with modern models.

FaceNet and DeepFace are outdated under open-set or domain-shift conditions.

Label noise significantly hampers learning. Hybrid training and filtering yield strong gains.

8. Conclusion and Future Work

ArcFace and MagFace are superior across nearly every benchmark. Their margin-based loss functions and quality-aware embeddings offer strong generalization, even in noisy or shifted domains.

We recommend:

a) Training with margin-based and noise-corrected losses
b) Fairness-aware thresholds
c) Mask and domain adaptation
d) Lightweight deployment strategies
e) Privacy-preserving embedding learning

9. References (APA)

1) Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). *DeepFace: Closing the gap to human-level performance in face verification*. CVPR, 1701–1708.
2) Schroff, F., Kalenichenko, D., & Philbin, J. (2015). *FaceNet: A unified embedding for face recognition and clustering*. CVPR, 815–823.
3) Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). *VGGFace2: A dataset for recognising faces across pose and age*. FG.
4) Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). *ArcFace: Additive angular margin loss for deep face recognition*. CVPR.
5) Meng, Q., Zhao, S., Huang, Z., & Zhou, F. (2021). *MagFace: A universal representation for face recognition and quality assessment*. CVPR.
6) Kemelmacher-Shlizerman, I., et al. (2016). *The MegaFace benchmark*. CVPR.
7) Whitelam, C., et al. (2017). *IARPA Janus Benchmark–C*. IJCB.
8) Moschoglou, S., et al. (2017). *AgeDB*. CVPR Workshops.
9) Wang, M., et al. (2019). *Racial Faces in-the-Wild (RFW)*. ICCV.
10) Castrejon, L., et al. (2021). *Masked face recognition under real-world variation*. T-BBI.