Benchmarking Face Recognition without Real Faces

Paweł Borsukiewicz
University of Luxembourg
Luxembourg
pawel.borsukiewicz@uni.lu

Daniele Lunghi*
University of Luxembourg
Luxembourg
daniele.lunghi@uni.lu

Wendkûuni C. Ouédraogo
University of Luxembourg
Luxembourg
wendkuuni.ouedraogo@uni.lu

Jacques Klein
University of Luxembourg
Luxembourg
jacques.klein@uni.lu

Tegawendé F. Bissyandé
University of Luxembourg
Luxembourg
tegawende.bissyande@uni.lu

arXiv:2607.14932v1 [cs.CV] 16 Jul 2026

Abstract

Synthetic face datasets have become effective enough to train face recognition models with accuracy rivaling that of models trained on real photographs. This progress sidesteps the ethical and legal burdens of collecting real biometric data, yet evaluation has not kept pace. Even studies that train entirely on synthetic images still rely on real-face benchmarks to measure performance, leaving the privacy problem only half solved.

We ask whether synthetic datasets can replace real benchmarks for face recognition evaluation. We test 12 synthetic datasets against 7 established real benchmarks using 24 pre-trained models that span both convolutional and transformer architectures. Our evaluation covers biometric verification metrics, similarity score distributions, cross-model ranking consistency, and the underlying distributional properties of each dataset. Benchmarking fidelity varies widely across the synthetic candidates, but the two strongest, MorphFace and Vec2Face, reproduce the relative behavior of real benchmarks and reach agreement levels that fall within the natural disagreement already observed among the real benchmarks themselves.

These results establish that well-constructed synthetic datasets can support reliable comparative evaluation for face recognition, moving the field closer to a fully synthetic and privacy-preserving pipeline for both training and benchmarking.

1. Introduction

Biometric identifiers such as fingerprints, facial geometry, and vein patterns constitute sensitive personal data. Their collection and processing fall under strict legal scrutiny, including the EU's General Data Protection Regulation (GDPR) [14] and the Illinois Biometric Information Privacy Act (BIPA) [24]. Beyond legal compliance, ethical obligations further constrain the biometric data life cycle. Data acquisition typically requires explicit informed consent and clearance from an institutional ethics committee [8]. These constraints limit both development and reproducibility, because many biometric datasets cannot be redistributed once privacy or consent requirements change.

Synthetic data offers a more ethical and privacy-friendly alternative [4, 33, 41] to real biometrics, since no sensitive information is collected from real subjects and regulatory exposure is correspondingly reduced. Advances in generative models, from Generative Adversarial Networks (GANs) [18] to Diffusion Models (DMs) [22], have made synthetic training data increasingly competitive. Models trained entirely on artificial faces now achieve recognition rates comparable to those of models trained on real photographs [4]. Yet this is only half a solution. While the training side of the pipeline can now operate without real faces, evaluation still cannot. Even studies that propose new synthetic datasets evaluate them on real benchmarks. A proper assessment of whether synthetic data suffices for evaluation is, to the best of our knowledge, absent from the literature. Without this evidence, researchers remain forced to include real data in their pipelines, if only for benchmarking, and the privacy problem persists.

Research into synthetic benchmark datasets has been limited [2, 10, 34] and has not produced clear guidance on their practical applicability. In this work we provide a systematic answer to the question: *can synthetic benchmarks reliably reflect model performance on real data?* This question is not merely academic. With limited control over the generation process [30], synthetic data can

*Corresponding author

exhibit poor identity separability, insufficient intra-class diversity, and demographic biases. A distributional mismatch between synthetic and real images therefore cannot be excluded without thorough empirical evaluation.

The main contributions of this work are as follows:

- We conduct the first comprehensive evaluation of 12 synthetic face recognition datasets as benchmarks across 24 face recognition models, comparing them with 7 commonly used real benchmark datasets.
- We provide a detailed breakdown of each dataset's benchmarking performance across recognition accuracy and a set of standard biometric evaluation metrics.
- We show that MorphFace and Vec2Face achieve the strongest alignment with real benchmarks and can serve as reliable substitutes.
- We investigate the factors that influence benchmarking utility, including domain gap, identity separability, and intra-class diversity.
- To promote reproducible research, our code is publicly available[1].

2. Related Work

Face recognition benchmarking has historically relied on real-world data (2.1). The growing body of synthetic datasets (2.2) has recently opened the door to synthetic evaluation (2.3).

2.1. Real Benchmark Datasets

Labeled Faces in the Wild (LFW) [23], with 13,233 images of 5,749 identities, was long the standard unconstrained benchmark for face recognition. As model performance on LFW has saturated [45], however, with most recent systems approaching 100% accuracy, its ability to discriminate among models has diminished.

Several more challenging alternatives have since been introduced. Cross-Pose LFW (CPLFW) [49] and Cross-Age LFW (CALFW) [50] modify the original LFW pairing strategy. CPLFW increases pose variability, while CALFW narrows the age gap between mated and non-mated pairs so that age alone is less predictive. On LFW, the mean age difference is 4.94 years for mated pairs and 14.85 years for non-mated pairs; CALFW brings these to 16.61 and 16.14, respectively. AgeDB-30 [32] targets temporal variation more directly by selecting 16,488 images from 568 identities, each pair separated by approximately 30 years. Celebrities in Frontal-Profile (CFP-FP) [38] instead focuses on viewpoint, pairing frontal and profile images across 7,000 images of 500 identities.

[1] https://github.com/Pabilito/Synthetic-Facial-Recognition-Benchmark

Representativeness is an equally important concern. IJB-B [45] addresses this with 1,845 human-labeled identities captured under diverse acquisition conditions; its 21,798 still images and 55,026 video frames carry, among others, annotations for occlusion, skin tone, facial landmarks, and localization. IJB-C [29] extends IJB-B to 31,334 images of 3,531 identities, broadening occupational and racial diversity and introducing stronger occlusion variation.

2.2. Synthetic Training Datasets

Retractions of real face datasets [8] and potential legal consequences [40] have motivated the search for alternatives. Synthetic data emerged as a response to the mass-scraping practices behind many widely used datasets. The development of Generative Adversarial Networks [18] and Diffusion Models [22] made high-fidelity image generation feasible, and research on synthetic face datasets has accelerated accordingly. State-of-the-art generators [26, 37] can be conditioned [8] to preserve a specific identity across multiple images.

The central difficulty [4] in constructing a useful synthetic dataset is achieving high intra-class diversity, the ability to produce semi-hard [42] samples that challenge the model, without sacrificing identity separability, meaning that distinct identities must not overlap unnaturally in the embedding space. Early datasets such as Syn-1M [28] and SynMulti-PIE [10] struggled with this balance and yielded lower performance when used for training. More recent datasets have closed the gap considerably. MorphFace [31], CemiFace [42], Vec2Face [46], and VariFace [47] now support training pipelines whose recognition performance matches, and in some cases exceeds [47], that of models trained on real data.

2.3. Synthetic Benchmark Datasets

Only a handful of datasets have been proposed specifically for benchmarking. SynMulti-PIE [10] was the first to address this goal, emphasizing the balance [4] between identity separability (minimal overlap between the similarity distributions of genuine and impostor pairs) and intra-class diversity (sufficient variation to make evaluation challenging). The authors argue that generated samples preserve identity without revealing the real identities used to train the generator. However, their evaluation covers only 64 identities, which limits the robustness of the conclusions, particularly with respect to universality.

ControlFace10k [34] contains 10,008 images of 3,336 synthetic identities generated through the Synthetic Identity Generation (SIG) pipeline. Although it provides three images per identity (frontal, left profile, and right profile), the intra-class variation was reported to be insufficient [4]. Identity separability is also weak, making ControlFace10k a questionable choice for either training or benchmarking.

SDFD [2] consists of 1,000 high-quality diffusion-generated samples with one image per identity. It is useful for demographic attribute prediction but cannot support pairwise verification benchmarking since it lacks multiple images per identity.

No synthetic dataset, whether designed for training or evaluation, has been directly validated as a face recognition benchmark. Because synthetic images may differ from real images in distributional properties and generative biases, their suitability as evaluation tools cannot be assumed. Establishing this suitability through rigorous empirical comparison is the gap that our study fills.

3. Methodology

Our testbed comprises 7 real and 12 synthetic face recognition datasets evaluated across 24 pre-trained models.

3.1. Datasets

We selected 7 real datasets for evaluation: LFW, CPLFW, CALFW, AgeDB-30, CFP-FP, IJB-B, and IJB-C. Beyond being the most commonly used benchmarks in the literature, these datasets collectively cover a broad range of challenging scenarios including pose variation and age gaps.

For the synthetic side, we selected 12 datasets generated by methods ranging from pure GANs and pure diffusion-based approaches to hybrid pipelines combining DMs with 3D morphable models (see Table 1). Together, these 12 datasets constitute the majority of publicly available synthetic face data. Access to IDiff-Face and SFace2 was obtained through online request forms; access to MorphFace was obtained by contacting the authors directly.

Table 1. Overview of synthetic datasets used.

Dataset	Generation Method	First Release	Venue
SynMulti-PIE [10]	GAN	2021	IJCB'21
SynFace [36]	GAN	2021	ICCV'21
DCFace [27]	Diffusion	2022	CVPR'23
IDiff-Face [5]	Diffusion	2023	ICCV'23
SFace2 [6]	GAN	2024	T-BIOM
Langevin-DisCo [15]	GAN	2024	ICML'25
Vec2Face [46]	GAN	2024	ICLR'25
CemiFace [42]	Diffusion	2024	NeurIPS'24
ControlFace10k [34]	Diffusion	2024	ICPR'25
HyperFace [39]	GAN	2024	ICLR'25
MorphFace [31]	Diffusion + 3D	2025	CVPR'25
Digi2Real [17]	Diffusion + 3D	2025	WACV'25

3.2. Models

We selected 24 pre-trained models spanning 9 vision transformers (ViTs) [11, 16, 35] and 15 convolutional neural networks (CNNs) [3, 12, 43], trained on diverse datasets [1, 9, 13, 48]. Because this paper focuses on synthetic data, we specifically included 7 CNNs trained entirely on synthetic data [5, 7, 15, 39, 46], all using the ResNet50 backbone, which is the standard architecture for benchmarking synthetic face recognition. Table 2 lists all models used.

Table 2. Overview of models used.

Model	Training dataset
ResNet50	Langevin-DisCo/Dispersion, IDiff-Face, HSFace10/300k HyperFace, SFace
IResNet (18/50/100)	Glint360K, MS1MV2
FaceNet	VGGFace2, CASIA-WebFace
SwinFace	MS1MV2
TransFace (S, L)	Glint360K, MS1MV2
EdgeFace (base, S, XS, XXS)	MS1MV3

3.3. Evaluation Metrics

Face recognition systems typically report recognition accuracy, the percentage of correctly classified pairs, as a primary metric. Accuracy alone is insufficient for real-world biometric systems, however, because the costs of false positives and false negatives differ substantially [20]. Following ISO/IEC 19795-1:2021 [25], security-critical systems prioritize keeping the False Match Rate (FMR), also called False Acceptance Rate (FAR), below a fixed threshold, and then minimize the False Non-Match Rate (FNMR), the proportion of genuine samples incorrectly rejected.

The acceptable FMR is application-dependent. Systems protecting sensitive resources demand near-zero false positives, while consumer devices tolerate a less strict decision threshold. This distinction is captured by FMR100 and FMR1000, which report the FNMR when the decision threshold is set so that only 1 in 100 or 1 in 1,000 impostor comparisons results in a false match. ZeroFMR is the FNMR at FMR $= 0$. The Equal Error Rate (EER) is the operating point where FMR equals FNMR, balancing security against usability. Finally, TAR@FAR reports the percentage of genuine samples accepted at a given FAR threshold, complementing the FMR-based metrics.

3.4. Dataset Correlation

A synthetic dataset can serve as a benchmark only if the scores it produces correlate with those obtained on real data. Given n paired observations (x_i, y_i) representing the same models evaluated on two different datasets, we measure lin-

ear agreement with Pearson's coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where $\bar{x}$ and $\bar{y}$ are the sample means. We complement Pearson's r with the Mean Absolute Error (MAE), which quantifies the average absolute difference between the scores produced by two benchmarks:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \quad (2)$$

Even if a synthetic benchmark is not expected to replicate absolute accuracy scores, it must at minimum preserve the relative ordering of models: a model that ranks higher on real data should also rank higher on synthetic data. We measure this property with Spearman's rank correlation, computed from the rank differences d_i (difference between the positions, based on a sorted evaluation metric, assigned to the same model by two benchmarks) across n models:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (3)$$

Both r and ρ range from -1 (perfect negative alignment) to $+1$ (perfect positive alignment). Following established guidelines [19], we interpret intermediate values as weak (0–0.5), moderate (0.5–0.75), semi-strong (0.75–0.85), or strong (0.85–1), while noting that these thresholds are not universally fixed.

We also examine the distributional properties that may explain benchmarking performance, following prior analyses [4, 5]. To quantify the domain gap (statistical difference) between synthetic and real images, we use the Fréchet Inception Distance (FID) [21], formulated as:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\Big(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\Big) \quad (4)$$

where the subscripts r and g denote the real and generated distributions, respectively. The mean term measures distributions center shift, while the trace term evaluates mismatches in the shape and orientation of their covariance structures. To establish statistical significance, we report 95% confidence intervals obtained by bootstrapping (10,000 resamples with replacement), taking the 2.5th and 97.5th percentiles as interval lower and upper bounds.

3.5. Evaluation Protocols

We followed the evaluation protocol defined for each real dataset to preserve methodological consistency. LFW, CPLFW, CALFW, and AgeDB-30 all use the *Unrestricted with Labeled Outside Data* procedure, in which 3,000 mated pairs and 3,000 non-mated pairs are divided into 10 folds. In each fold, an optimal threshold is computed and then applied to compute accuracy; final results are the mean accuracy $\pm$ standard deviation across folds. CFP-FP follows the same procedure but increases the sample size to 3,500 mated and 3,500 non-mated pairs.

IJB-B and IJB-C use a template-based protocol in which multiple images or frames are aggregated per identity to approximate real-world conditions. We use the *1:1 Baseline Verification* protocol with 8,010,270 comparisons to compute TAR at FARs of 1% and 0.01%.

For the synthetic datasets, we adopted the *Unrestricted with Labeled Outside Data* protocol on a randomly selected set of 3,000 mated and 3,000 non-mated pairs, matching the majority of the real dataset protocols. The image-pair lists are included in our replication package.

In addition to the protocol-specified metrics, we computed EER, FMR100, FMR1000, and ZeroFMR for every model-dataset combination. This provides a more complete picture of how synthetic benchmarks behave across different security and usability requirements.

4. Experiments

For every model-dataset pair, we first compute cosine similarities and the resulting recognition accuracies. We then evaluate biometric performance metrics (EER, FMR100, FMR1000, and ZeroFMR) and measure Pearson's r and Spearman's ρ between datasets. IJB-B and IJB-C were excluded from the accuracy computation as the template-based protocol is incompatible with the pair-based synthetic setup. To shed light on the factors behind benchmarking quality, we also examine similarity score distributions and FID scores. We close with a focused analysis of the two strongest synthetic candidates.

4.1. Pearson's Correlation

Recognition accuracies assessed with Pearson's r (Figure 1) reveal that most dataset pairs exhibit high correlations, with values rarely falling below 0.9. The one clear outlier is ControlFace10k, whose low r indicates that it does not satisfy the criteria for a reliable benchmark. Analysis of 95% confidence intervals, however, exposes additional weaknesses that the point estimates conceal. Real-real pairs maintain consistently high lower bounds (0.77–0.97), whereas several real-synthetic pairs show substantially larger drops, with Langevin-DisCo and SynMulti-PIE reaching negative lower bounds. DCFace, Digi2Real, HyperFace, MorphFace, and Vec2Face, by contrast, maintain confidence bounds comparable to those of real-real pairs.

A finer-grained view by biometric metric (Table 3) shows that synthetic datasets achieve their strongest correlations under EER and their weakest under ZeroFMR. The ZeroFMR gap points to an unnatural distribution of outliers, specifically highly similar non-mated pairs as docu-

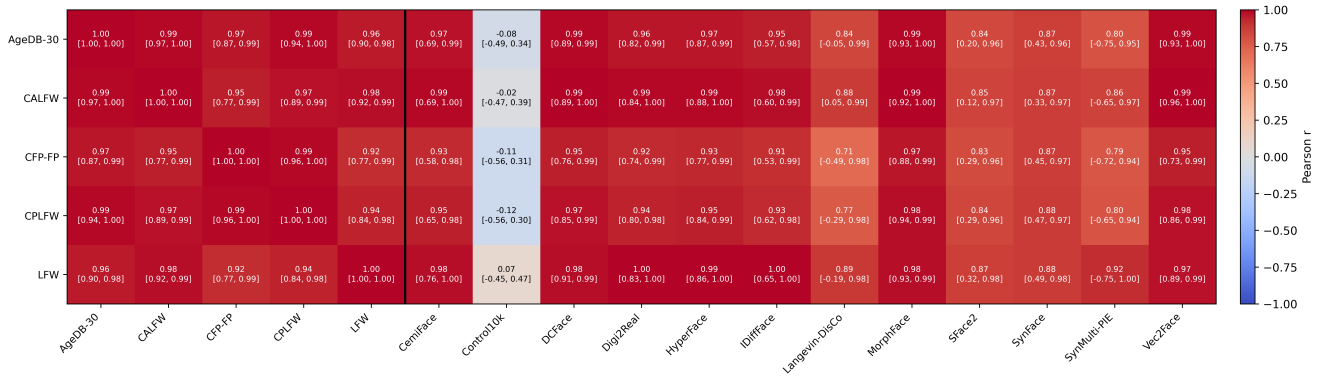

Figure 1. Recognition accuracy alignment between datasets across 24 models - Pearson's r with 95% CI

Table 3. Average Pearson's r per metric on 24 models across synthetic datasets when compared with 7 real datasets. Best in bold.

Dataset	EER	ZeroFMR	FMR100	FMR1000	Mean
CemiFace	0.97	0.72	0.88	0.89	0.87
ControlFace10k	-0.31	-0.44	-0.57	-0.54	-0.46
DCFace	0.98	0.78	0.89	0.91	0.89
Digi2Real	0.97	0.62	0.88	0.78	0.81
HyperFace	0.97	0.65	0.89	0.80	0.83
IDiff-Face	0.96	0.64	0.90	0.85	0.84
Langevin-DisCo	0.81	0.07	0.53	0.19	0.40
MorphFace	**0.98**	0.78	**0.93**	**0.92**	**0.90**
SFace2	0.84	0.30	0.78	0.51	0.61
SynFace	0.87	0.52	0.83	0.73	0.74
SynMulti-PIE	0.85	0.35	0.78	0.59	0.64
Vec2Face	0.98	**0.81**	0.90	0.92	0.90

mented by Borsukiewicz et al. [4], even though overall separability remains fairly consistent. In line with the accuracy results, MorphFace and Vec2Face rank as the best benchmarks, both averaging $r = 0.90$. The two are interchangeable for EER and FMR1000; MorphFace leads on FMR100 (0.93 vs. 0.90) while Vec2Face leads on ZeroFMR (0.81 vs. 0.78). The datasets designed explicitly for evaluation, ControlFace10k and SynMulti-PIE, are among the worst performers, with ControlFace10k being the only dataset that consistently shows negative correlations.

4.2. Spearman's Rank Correlation

Pearson's r assumes a linear relationship, which may not hold. Spearman's ρ (Figure 2) relaxes this assumption and directly measures how well model rankings are preserved across datasets. The results draw a clear line between synthetic datasets that can rank models and those that cannot. MorphFace and Vec2Face show the strongest alignment with real datasets ($\rho > 0.85$), only slightly below the agreement among real-real pairs, confirming them as the most suitable synthetic alternatives. The synthetic datasets that perform well do so consistently across all seven real benchmarks, including those focused on extreme poses or large age gaps, while poor synthetic benchmarks underperform just as uniformly.

Spearman's ρ for biometric metrics (Table 4) is broadly consistent with the Pearson results. MorphFace and Vec2Face again lead with a mean ρ of 0.87, while ControlFace10k (-0.52) and SynMulti-PIE (0.03) fall far behind. Most remaining datasets produce ρ values between 0.5 and 0.8, indicating a moderate-to-strong relationship that supports the feasibility of synthetic benchmarking for face recognition.

Table 4. Avg. Spearman's ρ per metric on 24 models across synthetic datasets when compared with 7 real datasets. Best in bold.

Dataset	EER	ZeroFMR	FMR100	FMR1000	Mean
CemiFace	0.65	0.59	0.64	0.71	0.65
ControlFace10k	-0.38	-0.56	-0.58	-0.55	-0.52
DCFace	0.77	0.67	0.78	0.86	0.77
Digi2Real	0.78	0.65	0.78	0.69	0.73
HyperFace	0.79	0.60	0.80	0.73	0.73
IDiff-Face	0.67	0.43	0.66	0.59	0.59
Langevin-DisCo	0.39	0.29	0.54	0.36	0.40
MorphFace	0.89	**0.79**	**0.91**	0.88	**0.87**
SFace2	0.47	0.49	0.52	0.43	0.48
SynFace	0.63	0.49	0.70	0.63	0.61
SynMulti-PIE	-0.10	0.09	0.08	0.04	0.03
Vec2Face	**0.89**	0.77	0.90	**0.90**	0.87

4.3. Fréchet Inception Distance

To quantify the domain gap, we computed FID (Table 5) using an IResNet18 model trained for face recognition, rather than the commonly used Inception-v3 [44]. Among real datasets, the highest alignment appears between LFW, CPLFW, and CALFW (FID at most 24.0), which share a substantial portion of their images. AgeDB-30 is the most distinct from the other real datasets, likely because of its high age variance, with FID scores in the range 52.1–73.2.

Comparing synthetic to real data reveals that a low

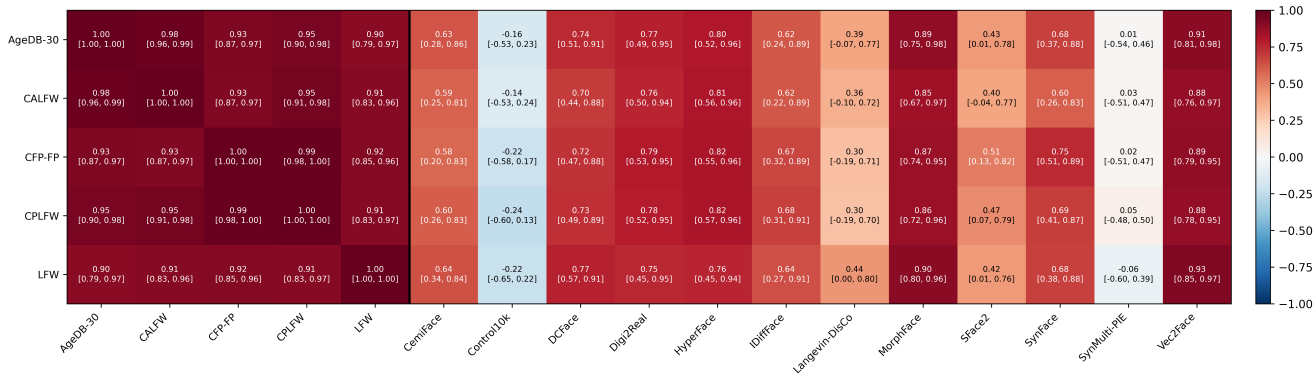

Figure 2. Spearman's ρ with 95% CI represents how strongly 24 models' accuracies ranking is preserved across datasets.

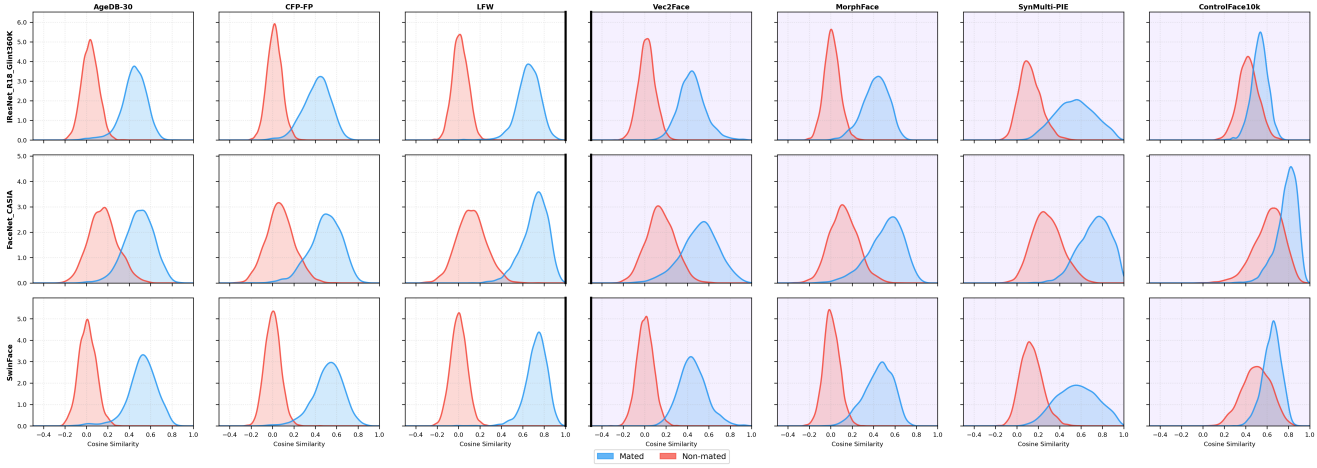

Figure 3. Mated vs. non-mated similarity distributions comparison across individual models and datasets.

FID is necessary but not sufficient for good benchmarking performance. Datasets with extremely high FID scores (ControlFace10k, SynFace, SynMulti-PIE, SFace2) consistently show poor correlation with real benchmarks, confirming that some baseline similarity to real data is required. Among the top-performing benchmarks, however, FID does not rank them correctly. MorphFace (FID 31.6–64.2) and Vec2Face (FID 51.9–92.0) both outperform CemiFace (FID 26.1–57.9) as benchmarks despite having higher FID scores, which indicates that distributional similarity alone does not determine benchmark reliability.

4.4. Similarity Distributions

To understand why some synthetic datasets succeed and others fail, we examined the cosine similarity distributions for mated and non-mated pairs (Figure 3). On real datasets, SwinFace and IResNet18 produce similar performance levels while FaceNet_CASIA clearly underperforms; Vec2Face and MorphFace reproduce this ordering, but SynMulti-PIE and ControlFace10k do not.

The shape of the distributions is also informative.

Table 5. Cross-dataset FID scores

Dataset	**AgeDB-30**	**CALFW**	**CPLFW**	**CFP-FP**	**LFW**
AgeDB-30	–	52.1	63.4	73.2	62.2
CALFW	52.1	–	24.0	44.6	11.1
CPLFW	63.4	24.0	–	29.3	22.3
CFP-FP	73.2	44.6	29.3	–	47.2
LFW	62.2	11.1	22.3	47.2	–
CemiFace	57.9	30.6	26.1	31.5	32.9
ControlFace10k	337.3	346.7	237.6	230.0	351.5
DCFace	67.6	41.5	36.6	39.6	44.9
Digi2Real	137.4	93.0	118.8	120.5	96.3
HyperFace	98.5	57.2	70.8	78.7	58.1
IDiff-Face	89.4	58.6	57.4	63.7	61.2
Langevin-DisCo	111.8	85.4	60.9	65.4	87.1
MorphFace	64.2	33.5	31.6	36.6	35.5
SFace2	176.0	171.5	108.4	107.4	174.8
SynFace	196.6	181.2	126.3	125.4	182.2
SynMulti-PIE	182.2	157.6	122.7	121.7	158.7
Vec2Face	92.0	51.9	68.3	78.6	53.5

Vec2Face and MorphFace resemble AgeDB-30 and CFP-FP, suggesting that they introduce more variability than LFW and therefore pose a greater challenge to models. This property makes them well suited for evaluating robustness to pose and age variation. SynMulti-PIE, by contrast, shows nearly identical overlap between mated and non-mated distributions regardless of model quality, indicating that some of its pairs are too diverse for any reasonable model to match, while others lack the diversity needed to challenge recognizers. ControlFace10k suffers from the opposite problem. High similarity scores among non-mated pairs signal low between-class variability and poor identity separability.

4.5. Analysis of MorphFace and Vec2Face

Given the results above, we focus on MorphFace and Vec2Face to assess their suitability as benchmarks more closely.

A per-dataset Spearman's ρ comparison (Table 6) shows that MorphFace and Vec2Face are often more aligned with real benchmarks than the real benchmarks are with one another. This is most visible for ZeroFMR, where MorphFace ($\rho = 0.79$) exceeds all real-real pairs (0.58–0.78) and Vec2Face (0.77) is close behind. The generally lower ρ values for ZeroFMR across the board confirm that measuring performance at zero false positives is inherently more variable, even among real benchmarks. The minimum 95% CI bounds show that alignment between individual dataset pairs can drop lower, but this is expected because certain real benchmarks serve specialized roles such as evaluating robustness to extreme pose or age differences. Overall, the evaluated metrics show no systematic gap between real and synthetic data for benchmarking purposes.

Pairwise comparisons (Figure 4) confirm that alignment levels are broadly consistent across metrics. MorphFace and Vec2Face position themselves as the most reliable proxies of LFW for recognition accuracy and FMR100, of CFP-FP for FMR1000, and of CALFW for EER. The only divergence appears for ZeroFMR, where Vec2Face tracks AgeDB-30 most closely while MorphFace corresponds best to CFP-FP. Comparing the two datasets head to head, Vec2Face has a slight edge for accuracy and FMR1000, while MorphFace is marginally better for FMR100. Differences for ZeroFMR and EER are inconsistent.

MAE estimates against real benchmarks (Table 7) show that MorphFace and Vec2Face align best on accuracy and EER and diverge more on FMR-based metrics. This pattern suggests that the two synthetic datasets serve well as direct accuracy proxies for the first two metrics and as ranking tools for the rest. The lowest error values appear most frequently against CALFW and AgeDB-30, indicating that MorphFace and Vec2Face are particularly well suited for evaluating models on age-invariant identity features. The two datasets are nearly indistinguishable from each other in MAE, confirming that either can serve as a benchmark.

Table 6. Per-metric and per-dataset Spearman's ρ coefficient on 24 models. Breakdown with respect to real datasets (excluding self-comparisons). MinLowCI represents the minimum observed value of the lower CI bound.

Metric		LFW	AgeDB-30	CALFW	CFP-FP	CPLFW	IJB-B	IJB-C	MorphFace	Vec2Face
EER	Mean ρ	0.93	0.94	0.89	0.94	0.93	0.94	0.95	0.89	0.89
	MinLowCI	0.80	0.75	0.65	0.81	0.80	0.80	0.88	0.70	0.67
ZeroFMR	Mean ρ	0.78	0.78	0.58	0.77	0.76	0.66	0.74	0.79	0.77
	MinLowCI	0.29	0.26	0.08	0.41	0.42	0.29	0.48	0.31	0.23
FMR100	Mean ρ	0.91	0.81	0.73	0.86	0.86	0.82	0.88	0.83	0.80
	MinLowCI	0.77	0.87	0.79	0.89	0.88	0.87	0.77	0.81	0.78
FMR1000	Mean ρ	0.89	0.96	0.95	0.97	0.96	0.95	0.95	0.95	0.95
	MinLowCI	0.70	0.76	0.72	0.79	0.74	0.80	0.79	0.70	0.70

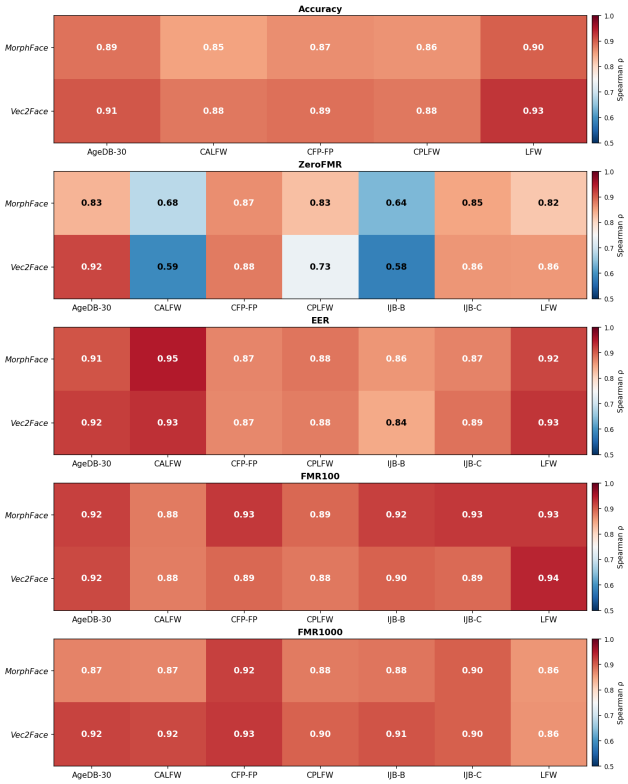

Figure 4. Per-metric Spearman's ρ on 24 models – MorphFace and Vec2Face vs. real datasets.

5. Discussion

Our experiments distinguish reliable from unreliable synthetic face datasets for the purpose of model benchmarking. Building on these results, we discuss the factors that

Table 7. MAE of MorphFace (M) and Vec2Face (V) vs. real datasets across 24 models.

Metric		LFW	CALFW	CFP-FP	CPLFW	AgeDB-30	IJB-B	IJB-C	Mean
Accuracy	M	0.045	0.015	0.024	0.051	0.015	–	–	0.030
	V	0.038	0.020	0.028	0.056	0.018	–	–	0.032
EER	M	0.048	0.024	0.024	0.063	0.016	0.035	0.039	0.036
	V	0.041	0.031	0.031	0.069	0.020	0.029	0.031	0.036
ZeroFMR	M	0.191	0.190	0.096	0.255	0.118	0.551	0.267	0.238
	V	0.182	0.189	0.097	0.272	0.090	0.560	0.276	0.238
FMR100	M	0.111	0.039	0.057	0.147	0.059	0.101	0.104	0.088
	V	0.130	0.072	0.091	0.128	0.076	0.084	0.094	0.096
FMR1000	M	0.195	0.094	0.063	0.136	0.099	0.131	0.141	0.123
	V	0.175	0.064	0.075	0.156	0.067	0.111	0.121	0.110

determine how well synthetic data aligns with real samples and propose criteria for validating future synthetic datasets.

5.1. Factors Affecting Benchmarking Quality

The benchmarking quality of a synthetic dataset appears to be governed primarily by data diversity. The similarity distributions and FID scores support this interpretation. MorphFace and Vec2Face, the two strongest synthetic benchmarks, exhibit well-separated mated and non-mated distributions and relatively low FID scores, and they achieve the highest agreement with real benchmarks. Weaker candidates such as Langevin-DisCo, SFace2, and especially ControlFace10k show limited separation [4, 30, 47], larger distributional distance from real data, and correspondingly poor benchmarking performance. ControlFace10k represents the extreme case of domain gap, combining near-absent identity separability and a large distribution gap, as exemplified by FID score, with the weakest correlations observed in this study.

A related observation connects benchmarking quality to training utility. The synthetic datasets that perform best as benchmarks also tend to perform best for training [4, 31, 46]. MorphFace and Vec2Face rank among the strongest candidates in both settings, while Langevin-DisCo and SFace2 remain among the weakest. This overlap suggests that the dataset properties supporting effective training, namely high intra-class diversity and clear identity separability, also contribute to reliable benchmarking. The relationship is not strictly one-to-one, however. SynFace, for instance, is relatively weak for training [47] but comparatively stronger for benchmarking, indicating that the connection between training utility and benchmarking utility deserves further investigation.

5.2. Criteria for Validating Future Synthetic Face Recognition Datasets

To confirm future synthetic dataset applicability for benchmarking, in line with our methodology, one should empirically confirm that it aligns with real benchmarks. Importantly, synthetic datasets should:

- not exceed the natural disagreement level observed between real datasets
- maintain agreement at operating points such as EER/FMR100/FMR1000

Our results across 24 models show that, when measuring model accuracy, we should strive for both Pearson's r and Spearman's ρ to exceed 0.9, with lower bounds of the 95% CIs > 0.75. However, specific values depend on the number and choice of the models to be evaluated.

6. Conclusions

We evaluated whether synthetic face datasets can serve as reliable benchmarks for face recognition. The results reveal substantial variability across candidates. Datasets designed explicitly for benchmarking do not necessarily produce the strongest agreement with real benchmarks; several synthetic training datasets prove far more effective in that role. Among all candidates, MorphFace and Vec2Face stand out, consistently achieving Pearson's r and Spearman's ρ values whose 95% confidence intervals overlap with the disagreement observed among real benchmarks themselves.

Beyond identifying the strongest candidates, our results show a consistent separation between high-fidelity and low-fidelity synthetic benchmarks across all biometric metrics, a separation that is also reflected in the underlying similarity distributions. Well-constructed synthetic datasets can therefore support not only training but also reliable comparative evaluation for face recognition, reducing reliance on privacy-sensitive real facial data.

A promising direction for future work is the construction of synthetic evaluation pairs under explicit diversity constraints spanning both subject attributes and acquisition conditions. Such a strategy could yield stronger alignment with specialized benchmarks and enable targeted evaluation under specific covariates such as pose, age, or occlusion.

Acknowledgments

- This research was funded by the Luxembourg Army.
- AI-based tools were used for fixing grammatical and typographical errors.

References

[1] X. An, X. Zhu, Y. Xiao, L. Wu, M. Zhang, Y. Gao, B. Qin, D. Zhang, and Y. Fu. Partial fc: Training 10 million identities on a single machine, 2021.

[2] G. Baltsou, I. Sarridis, C. Koutlis, and S. Papadopoulos. Sdfd: Building a versatile synthetic face image dataset with diverse attributes. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024.

[3] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen. Invertible residual networks. In *International conference on machine learning*, pages 573–582. PMLR, 2019.

[4] P. Borsukiewicz, F. Boutros, I. E. Olatunji, C. Beumier, W. C. Ouédraogo, J. Klein, and T. F. Bissyandé. Beyond real faces: synthetic datasets can achieve reliable recognition performance without privacy compromise. *npj Artificial Intelligence*, 2026.

[5] F. Boutros, J. H. Grebe, A. Kuijper, and N. Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023.

[6] F. Boutros, M. Huber, A. T. Luu, P. Siebke, and N. Damer. Sface2: Synthetic-based face recognition with w-space identity-driven sampling. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[7] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022.

[8] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, 135:104688, 2023.

[9] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2018.

[10] L. Colbois, T. de Freitas Pereira, and S. Marcel. On the use of automatically generated synthetic image datasets for benchmarking face recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[11] J. Dan, Y. Liu, H. Xie, J. Deng, H. Xie, X. Xie, and B. Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20642–20653, 2023.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[13] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2638–2646, 2019.

[14] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

[15] D. Geissbühler, H. O. Shahreza, and S. Marcel. Synthetic face datasets generation via latent space exploration from brownian identity diffusion. *arXiv preprint arXiv:2405.00228*, 2024.

[16] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel. Edgeface: Efficient face recognition model for edge devices. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(2):158–168, 2024.

[17] A. George and S. Marcel. Digi2real: Bridging the realism gap in synthetic data face recognition via foundation models. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1469–1478, 2025.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[19] G. E. Granato. Statistics for stochastic modeling of volume reduction, hydrograph extension, and water-quality treatment by structural stormwater runoff best management practices (bmps). Technical report, US Geological Survey, 2014.

[20] P. Grother, P. Grother, A. Hom, M. Ngan, and K. Hanaoka. *Face recognition vendor test (FRVT) Part 7: Identification for paperless travel and immigration*. US Department of Commerce, National Institute of Standards and Technology, 2021.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[24] Illinois General Assembly. Public act 095-0994, 2008.

[25] ISO/IEC JTC 1/SC 37 Biometrics. Information technology – biometric performance testing and reporting – part 1: Principles and framework, 2021.

[26] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021.

[27] M. Kim, F. Liu, A. Jain, and X. Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 12715–12725, 2023.

[28] A. Kortylewski, A. Schneider, T. Gerig, B. Egger, A. Morel-Forster, and T. Vetter. Training deep face recognition systems with synthetic data. *arXiv preprint arXiv:1802.05891*, 2018.

[29] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[30] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert. Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 3086–3095, 2023.

[31] Y. Mi, Z. Zhong, Y. Huang, Q. Yuan, X. Zhao, J. Xu, S. Ding, S. Wang, R. Guo, and S. Zhou. Data synthesis with diverse styles for face recognition via 3dmm-guided diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21203–21214, 2025.

[32] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

[33] M. Nisevic, D. Milojevic, and D. Spajic. Synthetic data in medicine: Legal and ethical considerations for patient profiling. *Computational and Structural Biotechnology Journal*, 28:190–198, 2025.

[34] K. Nzalasse, R. Raj, E. Laird, and C. Clark. Sig: A synthetic identity generation pipeline for generating evaluation datasets for face recognition. In *International Conference on Pattern Recognition*, pages 299–313. Springer, 2024.

[35] L. Qin, M. Wang, C. Deng, K. Wang, X. Chen, J. Hu, and W. Deng. Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2223–2234, 2023.

[36] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao. Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[38] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[39] H. O. Shahreza and S. Marcel. Hyperface: Generating synthetic face recognition datasets by exploring face embedding hypersphere. *arXiv preprint arXiv:2411.08470*, 2024.

[40] M. Smith and S. Miller. The ethical application of biometric facial recognition technology. *Ai & Society*, 37(1):167–175, 2022.

[41] Z. Sun and Z. Liu. Ensuring privacy in face recognition: a survey on data generation, inference and storage. *Discover Applied Sciences*, 7(5):441, 2025.

[42] Z. Sun, S. Song, I. Patras, and G. Tzimiropoulos. Cemiface: Center-based semi-hard synthetic face generation for face recognition. *Advances in Neural Information Processing Systems*, 37:35612–35638, 2024.

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[45] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.

[46] H. Wu, J. Singh, S. Tian, L. Zheng, and K. W. Bowyer. Vec2face: Scaling face dataset generation with loosely constrained vectors. *arXiv preprint arXiv:2409.02979*, 2024.

[47] M. Yeung, T. Teramoto, S. Wu, T. Fujiwara, K. Suzuki, and T. Kojima. Variface: Fair and diverse synthetic dataset generation for face recognition. *arXiv preprint arXiv:2412.06235*, 2024.

[48] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[49] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7):5, 2018.

[50] T. Zheng, W. Deng, and J. Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.