MR-FIQA: Face Image Quality Assessment with Multiple Reference Representations in Synthetic Data
— Supplementary Material —

Fu-Zhao Ou[1] Chongyi Li[2, 3] Shiqi Wang[1] Sam Kwong[4]
[1]City University of Hong Kong, Hong Kong SAR, China [2]Nankai University, Tianjin, China
[3]NKIARI (Shenzhen-Futian), Shenzhen, China [4]Lingnan University, Hong Kong SAR, China
fuzhao.ou@my.cityu.edu.hk lichongyi@nankai.edu.cn shiqwang@cityu.edu.hk samkwong@ln.edu.hk

1. Overview

In this supplementary material, we provide additional analysis of the proposed SynFIQA dataset and MR-FIQA method (Section 2), detailed experimental settings (Section 3), and additional experimental results (Section 4).

2. Additional Analysis

2.1. Synthetic Dataset – SynFIQA

Motivation Emphasis. In Table 1, we summarize the properties of the SynFIQA and other face synthetic datasets. It is worth noting that the original intentions of existing synthetic datasets and their evaluations are mainly for face recognition. Although they can be used to train synthetic-based FIQA models, they are not as tailored for FIQA as ours, especially since our dataset also provides additional reference and quality labels. For visual comparison, some degraded samples from previous and proposed synthetic datasets are illustrated in Fig. 1.

Customized Post-Processing. In our generation pipeline, we adopt a post-processing scheme to control blur degradation and downsampling. Our choice of this post-processing approach is based on the following considerations: 1) The native stable diffusion model possesses a strong capability to generate high-resolution and highly realistic images, while injecting blur and downsampling distortion as conditional information for generation opposes this inherent generative ability; 2) It ensures a better quality of the generated reference images; 3) There is a direct correlation between control parameters and distortion; 4) It guarantees that our generated data is fully synthetic, as no real data is used to simulate real distortions or style distributions.

Demographic Statistics. Fig. 2 presents demographic statistics encompassing all identities within our dataset. Contrary to the approach of GANDiffFace [20], we do not integrate demographic-specific transformations in the first phase of the generation pipeline, such as altering age, race, or gender. Consequently, the demographic composition within our dataset reflects the influence of the unconditional generator trained on FFHQ [16].

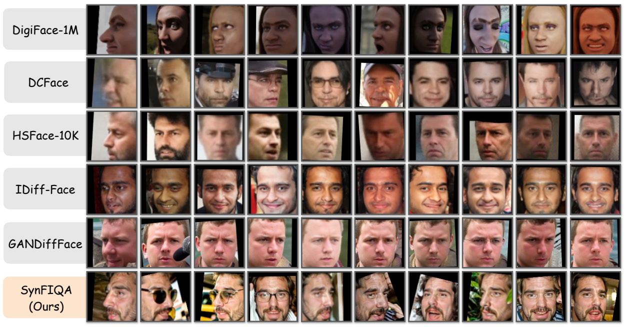

Figure 1. Visualization of various degraded samples from different synthetic datasets. The samples in each row originate from the same identity and are cropped and aligned via RetinaFace [12]. In our SynFIQA dataset, the samples display significant quality variations in pose, expression, and other quality factors, rendering it particularly well-suited for the FIQA task.

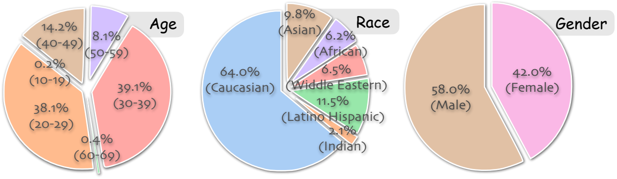

Figure 2. Proportions of age, race, and gender of identity in our SynFIQA dataset. These demographic attributes are obtained through the facial analysis via the deepface model [27].

Intensity Parameters of Degradation. For the generation of degraded samples, we define different selection probabilities for the control parameters of distortion within various parameter intervals, while the specific distortion parameter selection within each interval follows a uniform distribution. Specifically, the selection probabilities for the pose intensity parameter p_y in the intervals [5, 10], (10, 25], (25, 50], and (50, 75] are 15%, 30%, 50%, and 5%, respectively. The blur intensity parameter p_b has selection probabilities of 15%, 65%, 10%, 8%, and 2% for the value = 0, intervals (0, 1.6], (1.6, 2.5], (2.5, 4.0], and (4.0, 8.0], respectively. The downsampling intensity parameter p_d has selection probabilities of 15%, 65%, 10%, 8%, and 2% for the value = 0, intervals (1, 2.6], (2.6, 3.5], (3.5, 5.0], and (5.0,

Table 1. Property summary of different face datasets. Compared with previous face synthetic (Syn.) datasets, our proposed SynFIQA is dedicated to the face image quality assessment (FIQA) task instead of face recognition (FR). Meanwhile, our dataset offers a wealth of labels, encompassing identity, reference (Ref.), and quality annotations.

Datasets	Venue	Data generation	Data type	#Size	#Identity	Identity label	Ref. label	Quality label	Main task
CASIA-WebFace [13]	-	-	Real	0.49M	10.5K	✓	✗	✗	-
DigiFace-1M [4]	WACV'23	Digital Rendering	Fully Syn.	1.2M	10K+100K	✓	✗	✗	FR
SFace2 [8]	T-BIOM'24	GAN-Based	Fully Syn.	1.05M	10.5K	✓	✗	✗	FR
HSFace-10K [8]	ArXiv'24	GAN-Based	Fully Syn.	0.5M	10K	✓	✗	✗	FR
DCFace [17]	CVPR'23	Diffusion–Based	Syn.$^{\triangleright}$	0.5M	10K	✓	✗	✗	FR
IDiff-Face [7]	ICCV'23	Diffusion–Based	Fully Syn.	0.5M	10K	✓	✗	✗	FR
GANDiffFace [7]	ICCVW'23	GAN-Diffusion–Based	Fully Syn.	0.54M	10K	✓	✗	✗	FR
SynFIQA (Ours)	-	Diffusion-Based	Fully Syn.	0.5M	5K	✓	✓	✓	FIQA

▷: *Sampling with real data.*

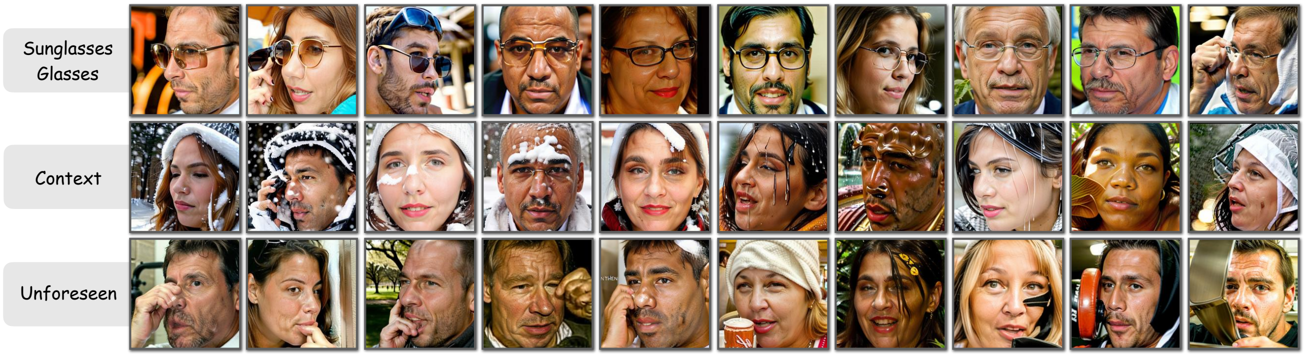

Figure 3. Visualization of degraded samples with different types of occlusions, including occlusion caused by controlling wearing glasses and sunglasses, and introducing a contextual environment by our customized positive text prompts. Moreover, there exist a few samples with unforeseen occlusion stemming from limbs, long hair, and objects. These diverse occluded samples collectively contribute to the quality variance within our SynFIQA dataset.

Table 2. pAUC and AUC results for different weighting factors λ.

Parameters	pAUC↓					
	CPLFW	XQLFW	AgeDB	Adience	TinyFace	Avg.
$\lambda = 0.1$	0.649	0.757	0.889	0.650	0.869	0.763
$\lambda = 0.3$	0.642	0.838	0.876	0.661	0.896	0.782
$\lambda = 0.5^{\dagger}$	0.632	0.800	0.887	0.608	0.808	0.747
$\lambda = 0.7$	0.605	0.819	0.885	0.599	0.823	0.746
$\lambda = 0.9$	0.576	0.765	0.908	0.642	0.812	0.741

Parameters	AUC↓					
	CPLFW	XQLFW	AgeDB	Adience	TinyFace	Avg.
$\lambda = 0.1$	0.363	0.375	0.685	0.445	0.519	0.477
$\lambda = 0.3$	0.351	0.439	0.665	0.439	0.549	0.489
$\lambda = 0.5^{\dagger}$	0.370	0.414	0.678	0.430	0.453	0.469
$\lambda = 0.7$	0.379	0.407	0.694	0.424	0.465	0.474
$\lambda = 0.9$	0.370	0.371	0.704	0.452	0.491	0.478

†: *This parameter is selected according to the overall Avg. pAUC and AUC.*

9.0], respectively. The expression, lighting, and position parameters are randomly selected within their respective control range intervals. The selection probabilities for each text prompt are uniform for positive text prompts used in occlusion control. It is important to highlight that while glasses and sunglasses introduce deterministic occlusion, all other prompts introduce uncertain occlusion influenced by the environmental context. Herein, various samples featuring different types of occlusions are illustrated in Fig. 3.

2.2. Quality Characterization Method – MR-FIQA

Recognition Embedding Domain. For the quality score in the recognition embedding domain, we use all reference representations in the intra-class embedding domain to compute the similarity between the target sample and reference representations as the quality score $\mathbf{f}_r$. In Fig. 5, we compare the results of adopting one and all reference representations under the AdaFace as the deployed recognition model. Herein, we only use the quality information of the embedding domain as quality labels to train the FIQA model, and the other settings are the same as our SynFIQA++. Clearly, the curve of our approach (denoted as All-RR) is consistently lower than one using only one reference representation, especially within the range of 30% to 70% ratio of unconsidered images. This validates the conclusion that All-RR has an advantage in predicting the quality of medium and high-quality samples.

Spatial Domain. In the spatial domain, we utilize the minimal rank among pose, blur, and downsampling to compute the quality score $\mathbf{s}_d$. This decision is based on the principle that the quality of samples is primarily influenced by the degradation factor with the highest intensity. In Fig. 4, we present the comparison results between using average rank and minimal rank. The minimal rank demonstrates stable performance across different test datasets compared to the average one. Thus, we adopt the minimal rank to calculate $\mathbf{s}_d$ in this domain.

Vision-Language Domain. The vision-language domain is introduced to address the issue of missing quality infor-

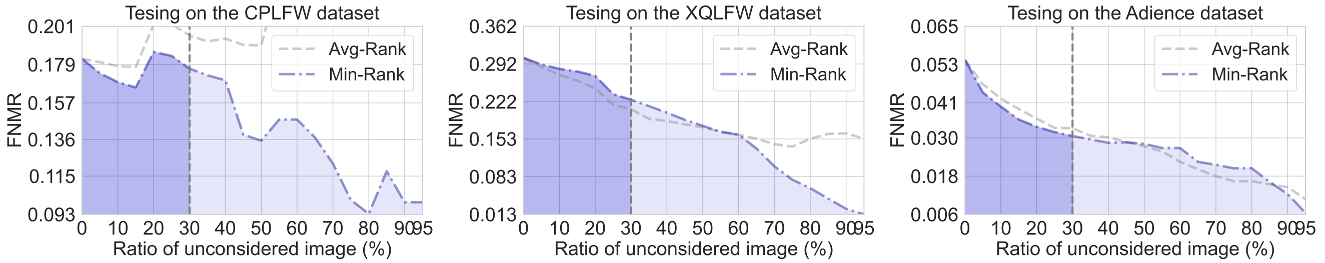

Figure 4. Comparison of using average rank (Avg-Rank) and minimal rank (Min-Rank) to compute quality scores in the spatial domain. The performance of Min-Rank is more stable than Avg-Rank on different test sets.

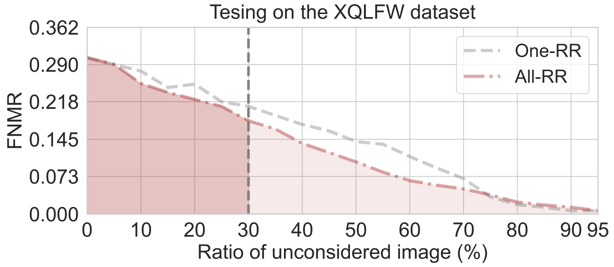

Figure 5. Comparison of using one and all Reference Representations (RR) as quality scores in the recognition embedding domain. Overall, All-RR yields superior results compared to One-RR.

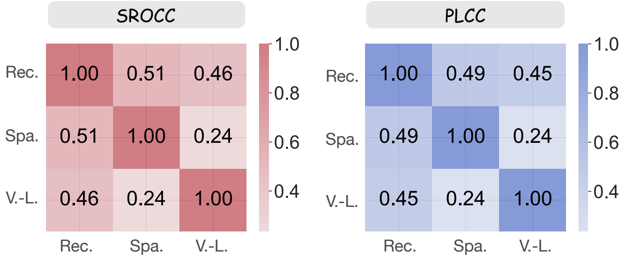

Figure 6. Correlation coefficients of quality scores for different domains, including recognition (Rec.), spatial (Spa.), and vision-language (V.-L.) domains. Quality scores from different domains have positive SROCC and PLCC results with each other.

mation in the other two domains. In the recognition embedding domain, $\mathbf{f}_r$ heavily relies on the recognition model used to compute similarity. However, the recognition model is robust to a certain degree of quality degradation, which is beneficial for recognition accuracy but can lead to insensitivity in quality assessment for FIQA, particularly evident in cross-recognition-model testing [28]. Furthermore, in the spatial domain, we only consider pose, blur, and downsampling factors that exhibit an absolute correlation with recognition utility to compute $\mathbf{s}_d$. To this end, $\mathbf{s}_d$ fails to capture the intricate quality mapping relationships among factors like lighting, expressions, and occlusions. To address these limitations, we leverage the powerful image-text matching capability of BLIP [19] to introduce the quality score $\mathbf{v}_l$ in the vision-language domain. Specifically, we use the text prompt from generating reference samples as the reference representation to compute the quality score for images. Although we only employ a pre-trained BLIP model in our work, according to the results of ablation experiments, the FIQA performance is satisfactory. In our future work, we plan to fine-tune the BLIP further to explore its superior potential in quality representation.

Design of Formula. Our final formula of MR-FIQA for calculating quality annotations is as follows:

$$Q(I_{|x_i}) = \varpi \left[\frac{\mathbf{f}_r + \lambda \cdot \mathbf{s}_d}{1 - \mathbf{v}_l} \right], \tag{1}$$

where λ is the weighting factor and $\varpi[\cdot]$ is the max-min normalization operator at the level of the whole dataset. The design of this formula is based on the following considerations: 1) $\mathbf{f}_r$ and $\mathbf{s}_d$ are able to reflect the recognition utility of samples directly, so we include them in the numerator. 2) The motivation behind introducing $\mathbf{v}_l$ is to compensate for the shortcomings of $\mathbf{f}_r$ and $\mathbf{s}_d$, hence we treat $\mathbf{v}_l$ in the denominator as a confidence-weighted design. In this scheme, $\mathbf{v}_l$ adjusts the overall calculation results. In Fig. 6, we present the Pearson Linear Correlation Coefficients (PLCC) and Spearman Rank-Order Correlation Coefficients (SROCC) between quality scores from different domains. It is evident that the highest correlation exists between $\mathbf{f}_r$ and $\mathbf{s}_d$. Additionally, the strong correlation between $\mathbf{v}_l$ and $\mathbf{f}_r$ further demonstrates the quality characterization's effectiveness via the vision-language domain.

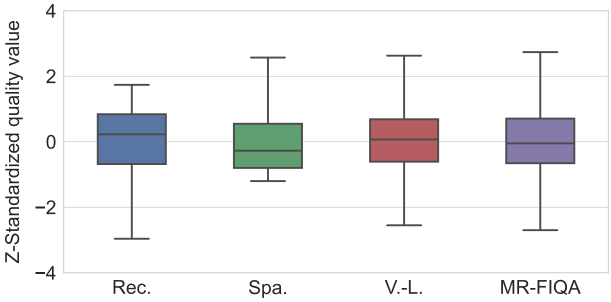

Figure 7. Quality distribution of Z-standardized scores across different quality characterization metrics. The box plots depict the median, interquartile range (IQR), and variability of quality value for the Rec., Spa., V.-L., and our proposed MR-FIQA metric.

Analysis of Quality Distribution. Here, we aim to pro-

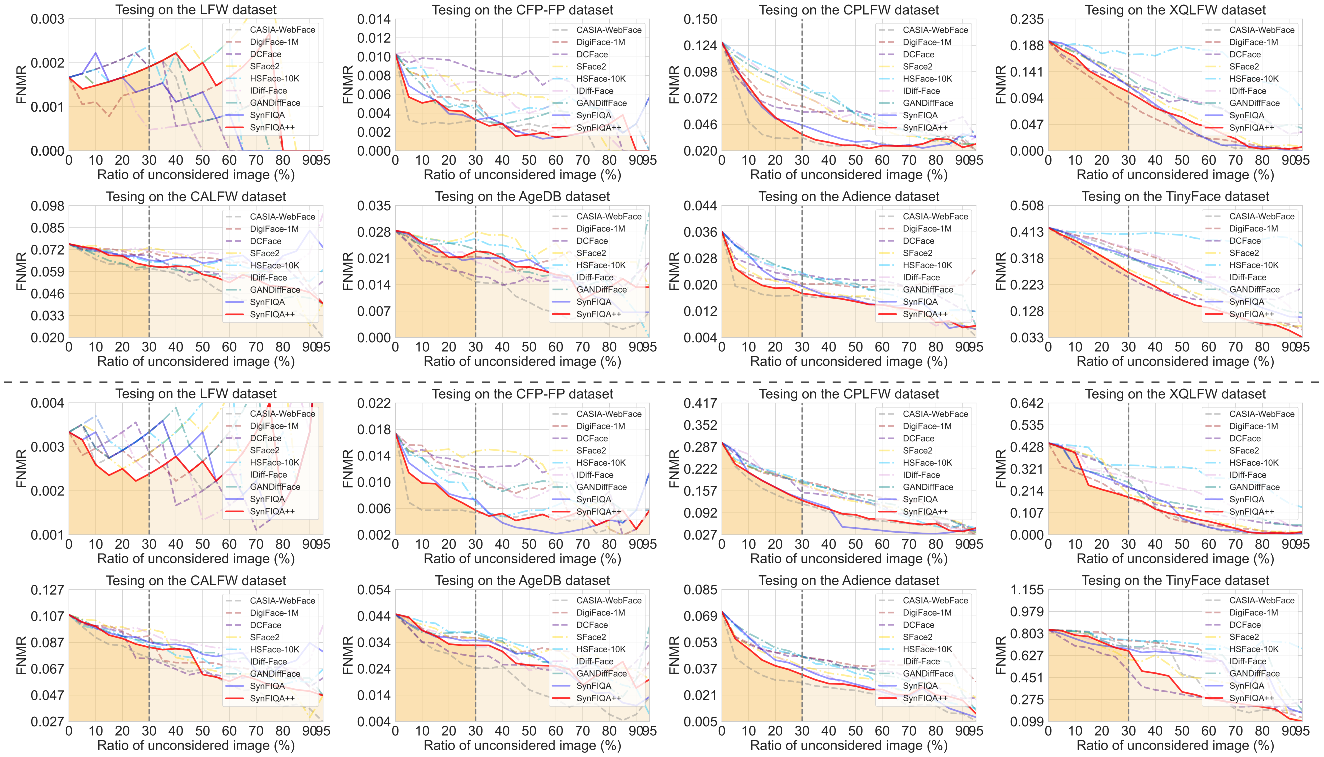

Figure 8. EDC results of various FIQA models trained on different training sets and tested on eight test datasets at FMR=1E-2 (Row#1-Row#2) and FMR=1E-4 (Row#3-Row#4) under the AdaFace deployed recognition model.

vide a comparative analysis of the distribution of each quality characterization metric under standardized conditions, to observe their relative consistency and variability. We show the quality distribution of these metrics and our proposed MR-FIQA in Fig. 7. Specifically, the distribution of $\mathbf{f}_r$ exhibits a symmetric and narrow box shape, with the median close to 0 and symmetric, short whiskers (approximately $\pm 1.5\sigma$), indicating fewer extreme values and reflecting its high stability. Due to the calculation based on the minimum ranking of multiple quality factors, the distribution of $\mathbf{s}_d$ is significantly right-skewed, with a median around -0.3 and an interquartile range (IQR) range of [-1.2, 0.8]. The distribution of $\mathbf{v}_l$ is symmetric, with an IQR range of [-0.8, 0.7] and a median close to 0, indicating balanced performance across most samples. For the proposed MR-FIQA, the box is highly compact (IQR [-0.4, 0.5]), with the median strictly aligned at 0 and symmetric whiskers distributed within $\pm 1.8\sigma$, significantly outperforming individual metrics. This demonstrates that MR-FIQA, through a weighted fusion strategy, achieves distribution centralization (IQR reduced by approximately 40%) by leveraging the discriminative power of $\mathbf{s}_d$ in spatial quality factors while inheriting the stability of $\mathbf{f}_r$ and $\mathbf{v}_l$, proving its effectiveness as a quality annotation method in our synthetic dataset.

Weighting Factor. In Table 2, we present the results of parameter sensitivity tests for the weighting factor λ in Eq. (1). In this study, we adjust λ to obtain varying quality annotations. Subsequently, we adopt the setting of SynFIQA++ under the MobileFaceNet [10] backbone to train different FIQA models for testing. The results in Table 2 indicate that as λ increases, there is a predominantly decreasing trend in pAUC and AUC, demonstrating the effectiveness of $\mathbf{s}_d$ in providing accurate quality annotations. Meanwhile, considering both pAUC and AUC comprehensively, we finally select $\lambda = 0.5$ for calculating quality annotations.

3. Experimental Settings

Error-versus-Discard Characteristic (EDC). In our experiments, we employ the EDC curve to assess the performance of different FIQA models. EDC is widely adopted in the evaluation of FIQA methods [2, 5, 20, 22, 24, 25, 28, 31]. Specifically, for a given target deployed face recognition model, the EDC reflects the accuracy of the FIQA model's prediction of recognition utility by measuring the variations of Ratios of Unconsidered Images (RUI) and False Non-Match Rate (FNMR) at a specific False Match Rate (FMR). During the computation of EDC, unconsidered images are eliminated as low-quality ones based on the ranking order of quality scores output by the FIQA model. Subsequently, the FNMR values of the remaining samples are calculated at certain RUI (*e.g.*, 0.05, 0.1, 0.2, etc.) under the deployed face recognition model. In essence, a faster

Table 3. pAUC and AUC results at FMR=1E-2, testing under the AdaFace as deployed recognition model.

Models	pAUC(↓)@FMR=1E-2								
	LFW	CFP-FP	CPLFW	XQLFW	CALFW	AgeDB	Adience	TinyFace	Avg.
CASIA-WebFace* [13]	1.166	0.321	0.414	0.829	0.870	0.733	0.517	0.901	0.719
DigiFace-1M‡ [4]	0.779	0.644	0.666	0.693	0.952	0.815	0.647	0.917	0.764
DCFace$^{\flat}$ [17]	1.206	0.925	0.603	0.768	0.935	0.723	0.709	0.795	0.833
SFace2 [8]	0.991	0.751	0.745	0.832	0.966	0.932	0.658	0.829	0.838
HSFace-10K [30]	1.105	0.692	0.827	0.944	0.930	0.918	0.778	0.960	0.894
IDiff-Face [7]	0.974	0.835	0.793	0.864	0.964	0.869	0.767	0.918	0.873
GANDiffFace [20]	1.006	0.617	0.790	0.852	**0.887**	0.891	0.789	0.871	0.838
SynFIQA (Ours)	1.035	0.542	0.578	0.836	0.939	**0.838**	0.722	0.882	0.797
SynFIQA++ (Ours)	**0.962**	**0.476**	**0.574**	**0.767**	0.925	0.854	**0.602**	**0.828**	**0.748**

Models	AUC(↓)@FMR=1E-2								
	LFW	CFP-FP	CPLFW	XQLFW	CALFW	AgeDB	Adience	TinyFace	Avg.
CASIA-WebFace* [13]	0.535	0.231	0.280	0.374	0.680	0.413	0.409	0.544	0.433
DigiFace-1M‡ [4]	0.617	0.390	0.445	0.313	0.872	0.715	0.584	0.621	0.570
DCFace$^{\flat}$ [17]	0.544	0.670	0.483	0.494	0.825	0.601	0.593	0.543	0.594
SFace2 [8]	1.145	0.467	0.461	0.428	0.827	0.840	0.466	0.531	0.646
HSFace-10K [30]	0.776	0.477	0.530	0.832	0.844	0.717	0.564	0.930	0.709
IDiff-Face [7]	**0.350**	0.491	0.489	0.551	0.932	0.636	0.482	0.689	0.578
GANDiffFace [20]	0.656	0.426	0.510	0.536	**0.768**	0.703	0.598	0.618	0.602
SynFIQA (Ours)	0.555	0.326	0.342	0.380	0.909	**0.624**	0.482	0.618	**0.530**
SynFIQA++ (Ours)	0.905	**0.295**	**0.323**	**0.376**	0.787	0.667	**0.434**	**0.481**	0.534

★: *Real data.* ‡: *Digital rendering.* ♭: *Sampling with real data.*

decrease in EDC indicates a more accurate prediction of recognition utility by the tested FIQA model.

FNMR@FMR=1E-3 in EDC. As suggested in [1, 2, 25, 28], we adopt FNMR@FMR=1E-3 to plot EDC curves and report AUC and pAUC results in our experiments. Because FMR=1E-3 is the threshold recommended by the best practice guidelines for automated border control systems at border inspection of Frontex [14]. Meanwhile, in the following Sec. 4, we also provide the results at FMR=1E-2 and FMR=1E-4 to reinforce our findings.

Cross-Recognition-Model Setting in FIQA Evaluation. The deployed recognition model indicates the recognition model employed in computing the EDC for the performance evaluation of FIQA models. Meanwhile, in the performance evaluation of FIQA models, in order to test the generalization capabilities of FIQA models across different recognition models and ensure fair comparisons between different FIQA models, we adopt the cross-recognition-model setting, which is widely embraced in existing FIQA methods [1, 2, 22, 24, 28]. This setting requires that the tested FIQA models use different recognition models from those involved in the training process and the deployed recognition model used in computing the EDC.

Area Under Curve (AUC) and partial AUC (pAUC). Drawing on [18, 22–24, 26], we present the AUC and pAUC outcomes in our experiments. It is worth noting that the AUC and pAUC are normalized as a proportion of the area of the EDC curve, which is determined using the formula:

$$\mathrm{AUC} = \frac{\int_a^b \mathcal{F}\left(r\right) dr}{\left(b-a\right) \times \mathcal{F}\left(a\right)}, \tag{2}$$

where $\mathcal{F}(r)$ represents the FNMR at a specific RUI r. For the AUC metric, the predefined lower and upper limits of RUI, a and b, are set at 0 and 0.95, respectively. The pAUC metric evaluates the FIQA performance at a reduced rejection threshold b, in order to provide an evaluation that aligns more closely with the practical deployment of FIQA models in real-world applications. In accordance with [2, 3, 23, 26], b is fixed at 0.3 for computing the pAUC metric.

Table 4. pAUC and AUC results at FMR=1E-4, testing under the AdaFace as deployed recognition model.

Models	pAUC(↓)@FMR=1E-4								
	LFW	CFP-FP	CPLFW	XQLFW	CALFW	AgeDB	Adience	TinyFace	Avg.
CASIA-WebFace* [13]	0.896	0.372	0.609	0.850	0.822	0.730	0.518	0.946	0.718
DigiFace-1M‡ [4]	0.873	0.787	0.732	0.625	0.907	0.865	0.716	0.972	0.810
DCFace$^{\flat}$ [17]	0.996	0.800	0.767	0.818	0.880	0.766	0.707	0.818	0.819
SFace2 [8]	0.879	0.845	0.753	0.826	0.930	0.865	0.691	0.912	0.838
HSFace-10K [30]	0.985	0.713	0.772	0.896	0.885	0.907	0.754	0.960	0.859
IDiff-Face [7]	0.966	0.780	0.786	0.807	0.922	0.855	0.730	0.921	0.846
GANDiffFace [20]	0.911	0.743	0.754	0.746	0.890	0.865	0.776	0.935	0.828
SynFIQA (Ours)	0.921	0.609	0.670	0.750	0.891	0.841	0.717	**0.894**	0.787
SynFIQA++ (Ours)	**0.779**	**0.535**	**0.655**	**0.694**	**0.877**	**0.834**	**0.640**	0.920	**0.742**

Models	AUC(↓)@FMR=1E-4								
	LFW	CFP-FP	CPLFW	XQLFW	CALFW	AgeDB	Adience	TinyFace	Avg.
CASIA-WebFace* [13]	0.833	0.288	0.348	0.368	0.600	0.441	0.344	0.610	0.479
DigiFace-1M‡ [4]	0.877	0.541	0.509	0.291	0.711	0.709	0.628	0.736	0.625
DCFace$^{\flat}$ [17]	0.764	0.628	0.479	0.415	0.677	0.633	0.541	0.495	0.579
SFace2 [8]	1.127	0.624	0.482	0.383	0.700	0.719	0.496	0.648	0.647
HSFace-10K [30]	1.009	0.439	0.476	0.700	0.706	0.691	0.505	0.891	0.677
IDiff-Face [7]	0.851	0.501	0.481	0.511	0.804	0.639	0.464	0.681	0.616
GANDiffFace [20]	0.966	0.567	0.499	0.418	0.742	0.664	0.583	0.781	0.652
SynFIQA (Ours)	0.848	**0.345**	**0.334**	0.337	0.771	**0.612**	0.446	0.707	0.550
SynFIQA++ (Ours)	**0.833**	0.357	0.372	**0.312**	**0.676**	0.642	**0.430**	**0.544**	**0.521**

★: *Real data.* ‡: *Digital rendering.* ♭: *Sampling with real data.*

Table 5. Comparison of different quality characterization methods, testing under the CosFace as deployed recognition model.

Models	pAUC↓				AUC↓				Avg.
	CFP-FP	CPLFW	XQLFW	Adience	CFP-FP	CPLFW	XQLFW	Adience	
FaceQnet [15]	0.583	0.741	0.843	0.659	0.334	0.481	0.454	0.441	0.567
SDD-FIQA [22]	0.603	0.729	0.850	0.697	0.417	0.477	0.445	0.434	0.581
CR-FIQA [6]	0.604	0.692	**0.812**	0.655	0.406	0.373	**0.419**	0.414	0.547
MR-FIQA (Ours)	**0.577**	**0.657**	0.828	**0.631**	**0.303**	**0.369**	0.450	**0.411**	**0.528**

TinyFace Dataset. Since the IJB datasets, including IJB-B and IJB-C, have been discontinued by NIST [21], we employ TinyFace [11] as an alternative in our experiments. It is important to note that, as TinyFace is tailored for 1:N recognition tests, it does not directly support performance evaluation based on the EDC metric of FIQA. To this end, we follow the practices outlined in [9, 26] to construct 19,478 positive mated comparisons and 24,513 negative mated comparisons from the Testing Set/Gallery Set and Testing Set/Probe Set. To ensure consistency in comparisons, the mated comparisons remain fixed across evaluations of different FIQA methods.

4. Additional Experiments

4.1. Additional Evaluation Results

Evaluation under different FMR. In our manuscript, we focused on EDC, pAUC, and AUC results, specifically at FMR=1E-3 due to space limitations. To provide a more comprehensive view, we extend this evaluation at FMR=1E-2 and FMR=1E-4, depicted in Fig. 8. At the same time, pAUC and AUC outcomes are presented in Table 3 and Table 4. Noteworthy is the consistent outperformance of our SynFIQA and SynFIQA++ against synthetic-based competitors across different FMRs.

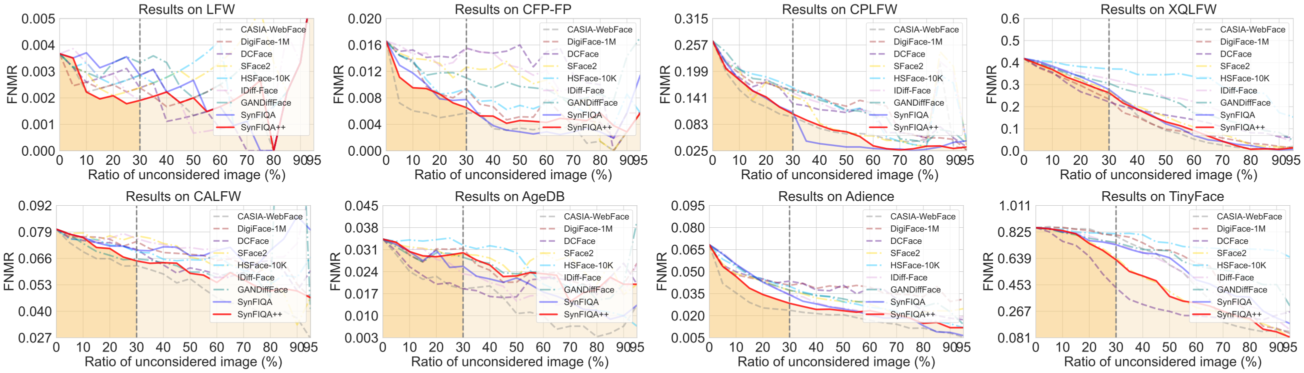

Figure 9. EDC results of various FIQA models trained on different training sets and tested on eight test datasets at FMR=1E-3 under the CosFace deployed recognition model.

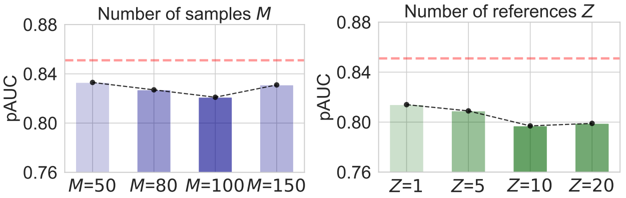

Figure 10. Parameter sensitivity of the sample number M and the reference number Z. The pAUC results are the average across eight test sets using ArcFace as the deployed recognition model. The optimal results are attained when $M = 100$ and $Z = 10$.

Evaluation under other Deployed Recognition Model. To further test the performance of synthetic-based FIQA models against other deployed recognition models. Here, we employ CosFace [29] trained on Glint360k dataset using the ResNet101 backbone. The EDC results at FMR=1E-3 under the CosFace recognition model are shown in Fig. 9. Moreover, we report the corresponding pAUC and AUC results in Table 6. Compared with other similar competitors (from SFace2 to GANDiffFace), our SynFIQA and SynFIQA++ also significantly outperform the others in terms of pAUC and AUC metrics. For the comparison of quality characterization methods under the CosFace deployed recognition model, we also report the results in Table 5. As shown in Table 5, except for the performance on XQLFW, our MR-FIQA still outperforms the other methods. And the average performance also surpasses CR-FIQA. This further indicates the effectiveness and robustness of the quality annotation in our SynFIQA dataset.

4.2. Ablation Study

Trade-off of Intra-Class and Inter-Class Samples. Here, we investigate the FIQA performance impact of the trade-off between the number of intra-class samples M and inter-class samples N. We explore different combinations of M and N while maintaining the total number of samples at $M \times N = 0.5$ million. FIQA models are trained under the CR-FIQA setting using the MobileFaceNet backbone. Considering that GANDiffFace [20] is a competitive competitor of the same fully synthetic type, we introduce its performance results as a reference indicated by the red dashed line in Fig. 10. From the experimental results in Fig. 10 (left), it is evident that the pAUC is lowest when $M = 100$, and there is an increasing trend in pAUC for M greater than 100. This suggests that the combination of $M = 100$ and $N = 5\text{K}$ is optimal within our comparison group. Moreover, for our dataset at $M = 50$, although this data level is slightly lower than GANDiffFace ($M = 54$), our synthetic data still significantly outperforms GANDiffFace.

Table 6. pAUC and AUC results at FMR=1E-3, testing under the CosFace as deployed recognition model.

Models	pAUC(↓)@FMR=1E-3								
	LFW	CFP-FP	CPLFW	XQLFW	CALFW	AgeDB	Adience	TinyFace	Avg.
CASIA-WebFace★ [13]	0.711	0.382	0.567	0.832	0.848	0.703	0.488	0.986	0.690
DigiFace-1M‡ [4]	0.709	0.665	0.715	0.779	0.943	0.899	0.698	0.983	0.799
DCFace♭ [17]	0.814	0.876	0.698	0.759	0.926	0.711	0.713	0.818	0.790
SFace2 [8]	0.707	0.815	0.714	0.832	0.959	0.899	0.700	**0.918**	0.818
HSFace-10K [30]	0.798	0.710	0.748	0.940	0.926	0.980	0.736	0.973	0.851
IDiff-Face [7]	0.807	0.884	0.715	0.886	0.960	0.871	0.736	0.948	0.851
GANDiffFace [20]	0.878	0.777	0.745	0.862	**0.885**	0.880	0.734	0.942	0.838
SynFIQA (Ours)	0.923	0.633	**0.629**	0.862	0.932	**0.856**	0.731	0.940	0.813
SynFIQA++ (Ours)	**0.643**	**0.571**	0.633	**0.818**	0.909	0.891	**0.614**	**0.918**	**0.750**

Models	AUC(↓)@FMR=1E-3								
	LFW	CFP-FP	CPLFW	XQLFW	CALFW	AgeDB	Adience	TinyFace	Avg.
CASIA-WebFace★ [13]	0.595	0.257	0.360	0.383	0.669	0.428	0.314	0.635	0.455
DigiFace-1M‡ [4]	0.636	0.436	0.485	0.382	0.847	0.733	0.594	0.714	0.603
DCFace♭ [17]	0.533	0.756	0.443	0.452	0.808	0.621	0.552	0.463	0.578
SFace2 [8]	0.863	0.598	0.453	0.438	0.829	0.796	0.490	0.549	0.627
HSFace-10K [30]	0.755	0.499	0.513	0.795	0.825	0.735	0.483	0.892	0.687
IDiff-Face [7]	0.622	0.671	0.527	0.607	0.908	0.622	0.458	0.665	0.635
GANDiffFace [20]	0.810	0.646	0.477	0.563	0.967	0.725	0.526	0.766	0.685
SynFIQA (Ours)	0.666	**0.355**	**0.295**	0.418	0.898	**0.604**	0.429	0.684	0.544
SynFIQA++ (Ours)	**0.606**	0.375	0.327	**0.406**	**0.767**	0.719	**0.401**	**0.548**	**0.519**

★: *Real data.* ‡: *Digital rendering.* ♭: *Sampling with real data.*

Number of Reference Samples. In Fig. 10 (right), we also present the pAUC results of FIQA models under different numbers of reference samples Z. In this study, our FIQA models are trained under the SynFIQA++ setting. It can be observed that as Z increases, there is a decreasing trend in pAUC results, which tends to converge around $Z = 10$. This not only underscores the importance of high-quality reference samples in our synthetic dataset but also further validates the effectiveness of leveraging integrated embeddings of reference samples as reference representations in the recognition domain to compute $\mathbf{f_r}$.

References

[1] Žiga Babnik, Peter Peer, and Vitomir Štruc. FaceQAN: Face image quality assessment through adversarial noise exploration. In *ICPR*, pages 748–754, 2022. 5

[2] Žiga Babnik, Peter Peer, and Vitomir Štruc. DifFIQA: Face image quality assessment using denoising diffusion probabilistic models. In *IJCB*, pages 1–10, 2023. 4, 5

[3] Žiga Babnik, Peter Peer, and Vitomir Štruc. eDifFIQA: Towards efficient face image quality assessment based on denoising diffusion probabilistic models. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pages 1–1, 2024. 5

[4] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. DigiFace-1M: 1 million digital face images for face recognition. In *WACV*, pages 3526–3535, 2023. 2, 5, 6

[5] Lacey Best-Rowden and Anil K Jain. Learning face image quality from human assessments. *IEEE Transactions on Information Forensics and Security*, 13(12):3064–3077, 2018. 4

[6] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. CR-FIQA: Face image quality assessment by learning sample relative classifiability. In *CVPR*, pages 5836–5845, 2023. 5

[7] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. IDiff-Face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *ICCV*, pages 19650–19661, 2023. 2, 5, 6

[8] Fadi Boutros, Marco Huber, Anh Thi Luu, Patrick Siebke, and Naser Damer. SFace2: Synthetic-based face recognition with w-space identity-driven sampling. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(3):290–303, 2024. 2, 5, 6

[9] Jacky Chen Long Chai, Tiong-Sik Ng, Cheng-Yaw Low, Jaewoo Park, and Andrew Beng Jin Teoh. Recognizability embedding enhancement for very low-resolution face recognition and quality estimation. In *CVPR*, pages 9957–9967, 2023. 5

[10] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR*, pages 428–438, 2018. 4

[11] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *ACCV*, pages 605–621, 2019. 5

[12] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212, 2020. 1

[13] Yi Dong, Lei Zhen, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2, 5, 6

[14] Frontex. Best practice technical guidelines for automated border control (abc) systems. https://www.frontex.europa.eu/assets/Publications/Research/Best_Practice_Technical_Guidelines_ABC.pdf, 2017. 5

[15] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. FaceQnet: Quality assessment for face recognition based on deep learning. In *ICB*, pages 1–8, 2019. 5

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

[17] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *CVPR*, pages 12715–12725, 2023. 2, 5, 6

[18] Jan Niklas Kolf, Naser Damer, and Fadi Boutros. Grafiqs: Face image quality assessment using gradient magnitudes. In *CVPRW*, pages 1490–1499, 2024. 5

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 3

[20] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. GANDiffFace: Controllable generation of synthetic datasets for face recognition with realistic variations. In *ICCVW*, pages 3086–3095, 2023. 1, 4, 5, 6

[21] NIST. IJB-C Dataset Request Form. https://www.nist.gov/itl/iad/ig/ijb-c-dataset-request-form, 2023. 5

[22] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. SDD-FIQA: Unsupervised face image quality assessment with similarity distribution distance. In *CVPR*, pages 7670–7679, 2021. 4, 5

[23] Fu-Zhao Ou, Baoliang Chen, Chongyi Li, Shiqi Wang, and Sam Kwong. Troubleshooting ethnic quality bias with curriculum domain adaptation for face image quality assessment. In *ICCV*, pages 20718–20729, 2023. 5

[24] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong. CLIB-FIQA: Face image quality assessment with confidence calibration. In *CVPR*, pages 1694–1704, 2024. 4, 5

[25] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. Face image quality assessment: A literature survey. *ACM Computing Surveys*, 54(10):1–49, 2022. 4, 5

[26] Torsten Schlett, Christian Rathgeb, Juan Tapia, and Christoph Busch. Considerations on the evaluation of biometric quality assessment algorithms. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(1):54–67, 2024. 5

[27] Sefik Serengil and Alper Özpınar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilişim Teknolojileri Dergisi*, 17(2):95–107, 2024. 1

[28] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *CVPR*, pages 5651–5660, 2020. 3, 4, 5

[29] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 6

[30] Haiyu Wu, Jaskirat Singh, Sicong Tian, Liang Zheng, and Kevin W Bowyer. Vec2Face: Scaling face dataset generation with loosely constrained vectors. *arXiv preprint arXiv:2409.02979*, 2024. 5, 6

[31] Weidi Xie, Jeffrey Byrne, and Andrew Zisserman. Inducing predictive uncertainty estimation for face recognition. In *BMVC*, pages 1–13, 2020. 4