DifFIQA: Face Image Quality Assessment Using Denoising Probabilistic Diffusion Models – Supplementary Material

Žiga Babnik, Peter Peer, Vitomir Štruc
University of Ljubljana, Kongresni trg 12, SI-1000 Ljubljana, Slovenia
{ziga.babnik,vitomir.struc}@fe.uni-lj.si, peter.peer@fri.uni-lj.si

Abstract

In the main part of the paper, we evaluated the proposed DifFIQA technique in comprehensive experiments across 7 diverse datasets, in comparison to 10 state-of-the-art (SOTA) competitors, and with 4 different face recognition models. In this supplementary material, we now show additional results using the same setup as in the main part of the paper that: (1) illustrate the performance of the model at another discard rate, (2) show the average performance of the proposed approach across all datasets and FR models and in comparison to all considered SOTA techniques for two different discard rates, and (3) provide details on the runtime complexity of the DifFIQA model. Additionally, we also discuss the limitation of the proposed FIQA models and provide information on the reproducibility of the experiments described in the main part of the paper.

1. Additional Results

Comparison to SOTA techniques. In Table 1, we present additional comparisons to the ten state-of-the-art techniques already considered in the main part of the paper. However, here the results are reported for a lower drop rate of 0.2. We note again that the performance of FIQA techniques is most relevant at lower drop rates, since this facilitates real-world applications, as also emphasized in [8].

From the presented results, we observe that the distilled model, DifFIQA(R) yields the lowest average pAUC scores (computed over the seven test datasets), when used with the AdaFace, ArcFace and CosFace models. With the CurricularFace model, DifFIQA(R) is the runner-up with performance close to the best performing CR-FIQA technique. It is worth noting that among the tested methods, four FIQA techniques performed significantly better than the rest across the four different FR models, i.e., CR-FIQA [2], FaceQAN [1] and the two diffusion-based models proposed in this paper, DifFIQA and DifFIQA(R). However, the distilled DifFIQA(R) technique is overall the top performer and fares particularly well on the most challenging datasets considered in the experiments, i.e., IJB-C and XQLFW.

Table 1. **Comparison to the state-of-the-art.** The table reports pAUC scores at a discard rate of 0.2 and a FMR of 10^{-3}. Average results across all datasets are marked $\overline{\text{pAUC}}$. The best result for each dataset is shown in bold, the overall best result is colored green, the second-best blue and the third-best red.

FIQA model	Adience	CALFW	CFP-FP	CPLFW	IJB-C	LFW	XQLFW	$\overline{\text{pAUC}}$
Cross-Model: AdaFace - pAUC@FMR=10^{-3}(↓)								
FaceQnet [5]	0.969	0.960	0.772	0.935	1.133	0.934	0.969	0.953
SDD-FIQA [7]	0.884	0.911	0.632	0.789	0.854	0.857	0.907	0.833
PFE [9]	0.873	0.917	0.659	0.772	0.918	0.854	0.885	0.840
PCNet [11]	1.003	0.985	0.893	0.926	0.843	0.730	0.999	0.911
MagFace [6]	0.890	0.900	0.632	0.747	0.915	0.735	0.958	0.825
LightQNet [3]	0.890	0.925	0.711	0.784	0.846	0.837	0.836	0.833
SER-FIQ [10]	**0.871**	0.930	0.563	0.715	0.812	0.982	n/a	0.812
FaceQAN [1]	0.905	0.942	**0.474**	0.700	0.800	**0.721**	0.764	0.758
CR-FIQA [2]	0.890	**0.887**	0.504	**0.684**	0.796	0.755	0.830	0.764
FaceQgen [4]	0.889	0.967	0.774	0.778	0.877	0.887	0.814	0.855
DifFIQA	0.897	0.932	0.500	0.698	0.813	0.770	0.769	0.768
DifFIQA(R)	0.893	0.913	0.505	0.696	**0.796**	0.752	**0.754**	0.758
Cross-Model: ArcFace - pAUC@FMR=10^{-3}(↓)								
FaceQnet [5]	0.957	0.970	0.761	0.918	1.123	0.934	0.933	0.942
SDD-FIQA [7]	0.841	0.931	0.637	0.829	0.806	0.857	0.874	0.825
PFE [9]	**0.823**	0.943	0.624	0.833	0.844	0.854	0.746	0.810
PCNet [11]	1.013	0.998	0.910	0.809	0.770	**0.697**	1.003	0.886
MagFace [6]	0.852	0.925	0.683	0.809	0.867	0.712	0.961	0.830
LightQNet [3]	0.840	0.930	0.706	0.857	0.788	0.814	0.772	0.816
SER-FIQ [10]	0.840	0.934	0.508	0.797	0.732	0.982	n/a	0.798
FaceQAN [1]	0.850	0.957	**0.470**	0.771	0.731	0.699	0.710	0.741
CR-FIQA [2]	0.861	**0.912**	0.475	0.791	**0.724**	0.732	0.764	0.751
FaceQgen [4]	0.857	0.980	0.823	0.834	0.823	0.865	0.786	0.853
DifFIQA	0.848	0.931	0.493	**0.771**	0.743	0.759	0.696	0.749
DifFIQA(R)	0.840	0.920	0.484	0.772	0.732	0.752	**0.688**	0.741
Cross-Model: CurricularFace - pAUC@FMR=10^{-3}(↓)								
FaceQnet [5]	0.941	0.964	0.692	0.914	1.139	0.960	0.990	0.943
SDD-FIQA [7]	0.838	0.932	0.556	0.802	0.806	0.865	0.867	0.810
PFE [9]	**0.815**	0.937	0.539	0.793	0.848	0.863	0.900	0.814
PCNet [11]	1.000	0.993	0.931	0.938	0.776	0.732	0.971	0.906
MagFace [6]	0.841	0.921	0.624	0.779	0.875	0.736	0.901	0.811
LightQNet [3]	0.827	0.938	0.574	0.815	0.787	0.834	0.857	0.805
SER-FIQ [10]	0.832	0.926	0.493	0.747	0.725	0.986	n/a	0.784
FaceQAN [1]	0.843	0.948	0.453	0.736	0.730	**0.713**	0.908	0.762
CR-FIQA [2]	0.859	**0.908**	**0.428**	**0.729**	0.734	0.746	0.902	0.758
FaceQgen [4]	0.858	0.972	0.754	0.806	0.824	0.894	0.836	0.849
DifFIQA	0.851	0.919	0.499	0.738	0.738	0.771	0.863	0.768
DifFIQA(R)	0.832	0.922	0.467	0.740	**0.723**	0.764	0.883	0.762
Same-Model: CosFace - pAUC@FMR=10^{-3}(↓)								
FaceQnet [5]	0.962	0.970	0.761	0.917	1.139	0.934	0.933	0.945
SDD-FIQA [7]	0.873	0.931	0.637	0.832	0.806	0.857	0.874	0.830
PFE [9]	**0.856**	0.943	0.624	0.837	0.848	0.854	0.746	0.816
PCNet [11]	1.005	0.998	0.910	0.861	0.776	**0.697**	1.003	0.893
MagFace [6]	0.882	0.925	0.683	0.808	0.875	0.712	0.961	0.835
LightQNet [3]	0.880	0.930	0.706	0.855	0.787	0.814	0.772	0.821
SER-FIQ [10]	0.863	0.934	0.508	0.790	0.725	0.982	n/a	0.800
FaceQAN [1]	0.890	0.957	**0.470**	0.759	0.741	0.699	0.710	0.747
CR-FIQA [2]	0.884	**0.912**	0.475	0.778	0.734	0.732	0.764	0.754
FaceQgen [4]	0.880	0.980	0.823	0.821	0.824	0.865	0.786	0.854
DifFIQA	0.881	0.931	0.493	**0.758**	0.738	0.759	**0.696**	0.751
DifFIQA(R)	0.870	0.931	0.484	0.758	**0.723**	0.759	**0.696**	0.746

[†]SER-FIQ was used to create XQLFW, so the results here are not reported for a fair comparison.

Table 2. **Average performance over all seven test datasets and four FR models at a drop rate of 0.2.** The results are reported in terms of average pAUC score at the FMR of 10^{-3}. The proposed DifFIQA(R) approach is overall the best performer. The best result is colored green, the second-best blue and the third-best red.

FaceQnet [5]	SDD-FIQA [7]	PFE [9]	PCNet [11]	MagFace [6]	LightQNet [3]	SER-FIQ [10]	FaceQAN [1]	CR-FIQA [2]	FaceQgen [4]	DifFIQA	DifFIQA(R)
0.9458	0.8244	0.8197	0.8989	0.8253	0.8183	0.7985	**0.7519**	**0.7567**	0.8527	0.7591	**0.7518**

Table 3. **Average performance over all seven test datasets and four FR models at a drop rate of 0.3.** The results are reported in terms of average pAUC score at the FMR of 10^{-3}. The proposed DifFIQA(R) approach is overall the best performer. The best result is colored green, the second-best blue and the third-best red.

FaceQnet [5]	SDD-FIQA [7]	PFE [9]	PCNet [11]	MagFace [6]	LightQNet [3]	SER-FIQ [10]	FaceQAN [1]	CR-FIQA [2]	FaceQgen [4]	DifFIQA	DifFIQA(R)
0.9315	0.7483	0.7497	0.8691	0.7635	0.7412	0.7292	0.6847	**0.6800**	0.7954	**0.6822**	**0.6768**

Table 4. **Detailed analysis of the runtime performance of DifFIQA in ms.** The reported results were computed over the entire XQLFW dataset and for each component of the model separately. For DifFIQA the times are presented separately for the initialization t_i, the forward process t_f, the backward process t_b, embedding of the images t_{fr}, and the quality calculation t_q steps. The symbol Σ denotes the overall runtime.

Model component runtime	t_i	t_f	t_b	t_{fr}	t_q	Σ
Runtime in ms ($\mu \pm \sigma$)	0.166 ± 0.006	0.192 ± 0.010	842.041 ± 9.068	66.224 ± 0.689	166.335 ± 1.750	1074.627 ± 11.458

Overall performance. To further illustrate the performance of the proposed DifFIQA and DifFIQA(R) techniques, we present in Tables 2 and 3 the average pAUC scores for two discard rates (0.2 and 0.3), computed over the seven test datasets and all four considered FR models. The reported results again support the findings already made above, i.e., FaceQAN, CR-FIQA, and our proposed techniques significantly outperform all other FIQA techniques, while DifFIQA(R) performs overall the best.

Runtime complexity. In the main part of the paper, we analyzed and tested all considered techniques from a runtime-performance perspective. Here, we explore the runtime complexity of DifFIQA in more detail to get better insight into the computationally most demanding steps of the approach. The whole method includes five steps: the initialization step (i), which creates all the necessary image copies and converts them into tensors, the forward diffusion step (f), the backward diffusion step (b), the image embedding step (fr), and the quality score calculation step (q). As can be seen from the reported results in Table 4, DifFIQA takes 1074ms on average to estimate the quality of a single face image. Recall, that the distilled approach requires only around 1ms for the same task. By far the most demanding part of the quality estimation procedure is the backward diffusion process, which iteratively denoises the given images, with an average time of a little more than 840ms. Even though we use only 5 iterations, we create for a single image 10 noisy copies of the original and the flipped version. All of these images are then passed through the denoising network, which accounts for the high time complexity of the backward process. The generation of image embeddings also requires some time, i.e., 66ms, as the step encapsulates the collection of all starting, noisy and reconstructed images into a single tensor as well as the forward pass through the FR model. In total, the image embedding steps need to produce embeddings for 60 images, all constructed from the given input sample. The score computation also takes close to 170ms, because it includes the calculation of five separate cosine similarities for all image copies, calculation of the average value over all copies and the data transfer from VRAM to RAM.

2. Limitations

The proposed DDPM-based DifFIQA technique ensure highly competitive FIQA performance, but also has some **limitations**. One obvious limitation is the computational complexity that affects the model's runtime performance, as emphasized throughout the paper. While this can be addressed through a distillation procedure, the distillation process removes the relation between the (noising and denoising) tasks and image quality, and consequently impacts the interpretability of the results. From a conceptual point of view, the nosing and denoising steps probe the quality of the facial images by (in a sense) first obscuring important facial features and then measuring the ability to restore the obscured features through denoising. Such restoration-based solutions may depend, to a significant degree, on the restoration model utilized, which in our case is a CNN-based UNet that implements the denoising diffusion. While such models are known to be able to capture local image characteristics very well, they may be less capable in capturing key global image properties, and we plan to explore transformer-based models in our future work to further improve on this limitation.

3. Reproduciblity

We would like to note that all of our experiments are fully reproducible. Most of the models used for the implementation and testing of DifFIQA and DifFIQA(R) are publicly available from the official repositories, while all others can be obtained by request from the authors, i.e.:

- AdaFace:
  https://github.com/mk-minchul/AdaFace
- ArcFace:
  https://github.com/deepinsight/insightface
- CosFace:
  https://github.com/deepinsight/insightface
- CurricularFace:
  https://github.com/HuangYG123/CurricularFace
- FaceQnet:
  https://github.com/javier-hernandezo/FaceQnet
- SDD-FIQA:
  https://github.com/Tencent/TFace/tree/quality
- PFE:
  https://github.com/seasonSH/Probabilistic-Face-Embeddings
- PCNet:
  Requested from authors
- MagFace:
  https://github.com/IrvingMeng/MagFace
- LightQNet:
  https://github.com/KaenChan/lightqnet
- SER-FIQ:
  https://github.com/pterhoer/FaceImageQuality
- FaceQAN:
  https://github.com/LSIbabnikz/FaceQAN
- FaceQgen:
  https://github.com/javier-hernandezo/FaceQgen
- CR-FIQA:
  https://github.com/fdbtrs/CR-FIQA
- Diffusion models:
  https://github.com/lucidrains/denoising-diffusion-pytorch

Additionally the source code for DifFIQA, including all training and testing scripts, model design and learned weights, is available from: https://github.com/LSIbabnikz/DifFIQA.

References

[1] Ž. Babnik, P. Peer, and V. Štruc. FaceQAN: Face Image Quality Assessment through Adversarial Noise Exploration. In *Proceedings of the IAPR International Conference on Pattern Recognition (ICPR)*, pages 748–754, 2022.

[2] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer. CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] K. Chen, T. Yi, and Q. Lv. LightQNet: Lightweight Deep Face Quality Assessment for Risk-Controlled Face Recognition. *Signal Processing Letters*, 28:1878–1882, 2021.

[4] J. Hernandez-Ortega, J. Fierrez, I. Serna, and A. Morales. FaceQgen: Semi-Supervised Deep Learning for Face Image Quality Assessment. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2021.

[5] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, and L. Beslay. Biometric Quality: Review and Application to Face Recognition with FaceQnet. *arXiv preprint arXiv:2006.03298*, 2020.

[6] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. MagFace: A Universal Representation for Face Recognition and Quality Assessment. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, 2021.

[7] F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y.-G. Wang. SDD-FIQA: Unsupervised Face Image Quality Assessment with Similarity Distribution Distance. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, 2021.

[8] T. Schlett, C. Rathgeb, J. Tapia, and C. Busch. Considerations on the Evaluation of Biometric Quality Assessment Algorithms. *arXiv preprint arXiv:2303.13294*, 2023.

[9] Y. Shi and A. K. Jain. Probabilistic Face Embeddings. In *Proceedings of the CVF/IEEE International Conference on Computer Vision (CVPR)*, pages 6902–6911, 2019.

[10] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5651–5660, 2020.

[11] W. Xie, J. Byrne, and A. Zisserman. Inducing Predictive Uncertainty Estimation for Face Verification. In *British Machine Vision Conference (BMVC)*, 2020.