Supplemental Material for Harnessing Unrecognizable Faces for Improving Face Recognition

Siqi Deng Yuanjun Xiong‡ Meng Wang Wei Xia‡ Stefano Soatto
AWS AI Labs

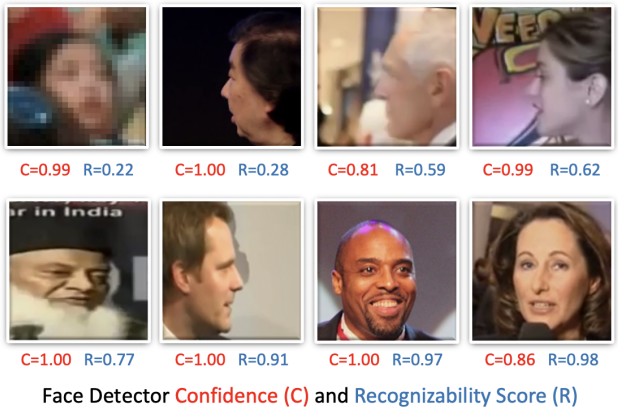

Figure 1. A comparison between face detection confidence and embedding-based recognizability score (ERS). The former is the output of a face detector and measures the likelihood that the image contains a face while the ERS measures if the face can be recognized. The face detection confidence scores are from Retinaface [2] detector.

Figure 2. Examples of the artificially corrupted images from DeepGlint 1K described in main paper Sec2.1 and Sec3.3. These images were used in the direct approach to generate the UI centroid.

Appendix A. Additional Experimental Results

Qualitative Comparison: Detection Confidence v.s. ERS

In Fig. 1 we show a comparison between face detection confidence and embedding-based recognizability score (ERS). The face detection confidence scores are from Retinaface [2] detector ‡. It can be seen that face detection confidences do not capture face recognizability because it is not tasked to do so. In light of this, we introduce ERS.

More Results on UI Centroid Generation

As discussed in the main paper Sec3.3, we explored two approaches for obtaining the UI images, the direct approach and the clustering-base approach. Here we provide more experiment results.

To artificially corrupt the images from the DeepGlint dataset, we applied low-level image processing techniques including Gaussian blur, down-sampling, motion blur, occlusion, affine transformation and rotation. We found it empirically easy to implement the direct approach as one does not need to make sure all the faces are completely unrecognizable, the approach is robust to the existence of some recognizable faces in the set. In Fig. 2, we show qualitative examples of the artificially corrupted images from DeepGlint 1K. It can be seen that many of the images have identity unrecognizable faces, while some of them are still somewhat recognizable to the human eyes.

It is worth mentioning that after running clustering on the DeepGlint 10K images where 1K are corrupted, the resultant UI cluster centroid is also very close (cosine distance is 0.0135) to that from directly averaging the 1K. This finding further supports our hypothesis that UI images tend to distribute closely around one centroid.

Benchmark Results on YoutubeFaces We show results on verification benchmark from YoutubeFaces in Table 2. Our method slightly improves the baseline: most of the faces are predicted as high recognizability (99% of the frames have ERSs above 0.8), thus ERS aggregation is similar to average pooling.

Exploration on Re-Id and Fashion Retrieval

‡Work done when at Amazon.

‡https://github.com/deepinsight/insightface/tree/master/detection/RetinaFace

		Ethnicity							Total
		African	Asian	European	Hispanic	Indian	Other	Unknown	
Gender	Female	24898	536	109132	1880	66	82	10	136604
	Male	155783	1150	99093	8908	322	93	102	265451
Total		180681	1686	208225	10788	388	175	112	402055

Table 1. Morph [5] dataset image statistics breakdown by Gender / Ancestry .

Method	Baseline (AvePool)	ERS
Accuracy (%)	96.62	96.64

Table 2. Templated-based face verification test on YoutubeFaces benchmark.

We perform person re-identification clustering embedding clustering on Market1501 [8] dataset which contains low recognizability examples which are labeled "junk" and "distractors" in this gallery set. Likewise for partially perturbed Deepfashion [4] In-Shop dataset (most image retrieval datasets do not contain natural quality corruption, so we manually perturb the recognizability). We observe low recognizability miscellaneous samples can gather in one cluster similar to those of faces (Fig. 3). After devising the associated ERS, it can be observed, from Fig.7 and Fig.8 in the main paper, that it also correlates with the input image recognizability, consistent with our findings on the face.

As a quick experiment to test if our method has the potential to generalize to tasks beyond face recognition, we apply our method on Market1501 [8] and show results in Table 4. It can be seen that our method can improve both mAP and rank-K accuracy.

Appendix B. Other Implementation Details

Face Clustering Parameters

We provide code for face clustering along with the submission, see "code.zip". We use HAC algorithm [1] to cluster normalized embedding features. For reference, when running face clustering on WIDERFace validation split with features extracted with Arcface model "MS1MV2: MS1M-ArcFace" from Insightface model zoo [‡]. We find distance threshold 1.0 and single linkage suitable using Scipy function "fcluster" [‡].

The optimal clustering parameters may vary from one embedding model to another, but we empirically find that the algorithm is not very sensitive to the distance threshold in order to obtain a set of unrecognizable images. Grid searching with distance step 0.1 usually gives satisfying results, and one set of parameters can generalize to multiple models.

[‡]https://github.com/deepinsight/insightface/tree/master/model_zoo

[‡]https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html

Figure 3. Similar to our findings in face clustering, miscellaneous low recognizability examples can gather in one cluster in person re-id (top, using Market1501 [8] dataset) and fashion retrieval datasets (bottom, using Deepfashion [4] dataset).

	FRR@FAR=		
	1e-2	1e-3	1e-4
Baseline	0.6820	0.9788	0.9960
ERS ($\gamma = 0.50$)	**0.4718**	0.5853	0.7133
ERS ($\gamma = 0.60$)	**0.4763**	**0.5772**	**0.7019**
ERS ($\gamma = 0.70$)	0.4896	**0.5768**	**0.6948**
ERS ($\gamma = 0.80$)	0.5178	0.5871	0.6820

Table 3. Tinyface face verification benchmark for ERS threshold selection, best two results each column are in bold. $\gamma = 0.60$ yields the best overall performances at multiple FARs.

Selection of ERS threshold

We use images from the TinyFace [7] dataset to generate a 1v1 face verification protocol, test on it using ERS as the recognizability measure to select the best threshold (Table. 3). We can see $\gamma = 0.60$ yields best overall performances at multiple FARs.

Setting	mAP	Rank-N Acc		
		1	5	10
Baseline [9]	0.7204	0.8786	0.9525	0.9697
ERS ($\gamma = 0.70$)	0.7314	0.8884	0.9569	0.9734

Table 4. Application on Market1501 for person re-id.

Implementation details of FaceQnet and SER-FIQ

We followed the best practice possible to ensure the comparison with FaceQnet [3] and SER-FIQ [6] to be fair. We used the original implementation [‡] from SER-FIQ to get prediction scores, during which we also adapt their preferred face detector and face embedding model from Insightface [‡]. To keep the prediction scores from SER-FIQ and face embedding model consistent, we used the same face embedding model from Insightface throughout the IJB-C Covariate Test benchmark to obtain results in the paper Table 1. This is also a demonstration of our method working effectively on an arbitrary face embedding model not trained by us. For FaceQnet, we also used the original implementation [‡] to get prediction scores, during which the preferred face detector was adapted.

Appendix C. Bias Analysis and Discussion

In paper Sec. 1.3, we had a discussion about potential implications of biases, here we conduct preliminary bias analysis and show results.

Test dataset and benchmark We use Morph [5] dataset as our test data for the analysis, basic statistics of the dataset are shown in Table 1.

Within each gender and ethnicity subgroup, we sample genuine and imposter 1v1 pairs to establish face verification benchmark protocol with gender and ethnicity breakdown.

Embedding models We test on two face embedding models previously used in benchmark experiments, one being our reference *CosFace* embedding model (ResNet101 + CosFace Loss + DeepGlintFace), the other being the *SC-Arc* model (ResNet101 + Sub-center Arcface Loss + DeepGlintFace) which has the best performance on IJB-C.

Recognizability prediction distribution We plot the recognizability prediction of $e = 1 - \langle \mathbf{f}_{UI}, \mathbf{f}_i \rangle$. Comparing to ERS defined in Eq.1, the cap at 1 operation is removed for the convenience of observing the raw distribution. We show e density distribution on Morph dataset for CosFace model in Fig. 4, and for SC-Arc model in Fig. 5.

Verification benchmark We also test and show face verification benchmark results with gender and ethnicity breakdown on Morph dataset for CosFace model in Table. 5, and for SC-Arc model in Table. 6.

Conclusions from the preliminary results:

[‡]https://github.com/pterhoer/FaceImageQuality
[‡]https://github.com/deepinsight/insightface
[‡]https://github.com/uam-biometrics/FaceQnet

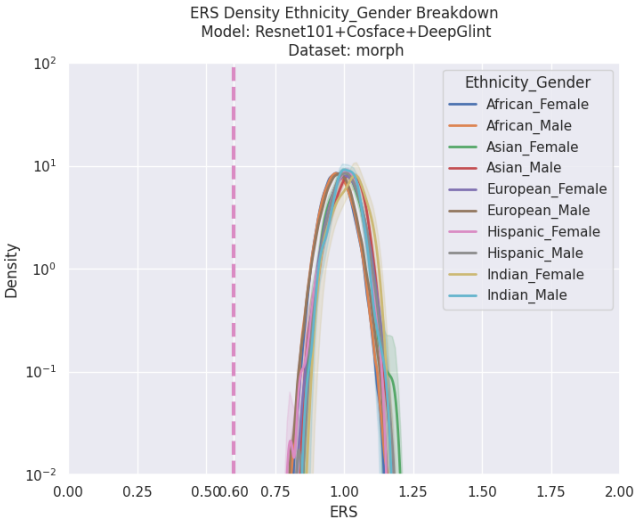

Figure 4. Reference CosFace model (ResNet101 + CosFace Loss + DeepGlintFace) breakdown of ERS on Morph dataset. Vertical dash indicates our threshold at 0.6.

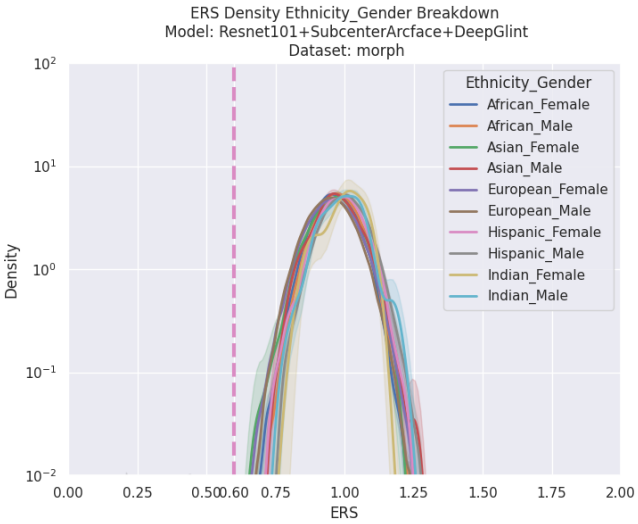

Figure 5. Reference SC-Arc model (ResNet101 + Sub-center Arcface Loss + DeepGlintFace) breakdown of ERS on Morph dataset. Vertical dash indicates our threshold at 0.6.

(1) The different distributions may be attributed to the particularity of the model evaluated and random fluctuations in the data, rather than systematic characteristics of the design. This is evidenced by that across two models, we do not observe one group (defined by Morph dataset ethnicity or gender labels) to consistently have higher or lower ERS in general.

(2) Albeit there could be biases in these models according to verification results, the differences among groups only matter little in the application of the ERS we proposed. The proposed ERS yields high prediction (>0.8) for most images, regardless of the labeled gender and ethnicity. The chosen threshold 0.6 almost never generates false positive

FAR	FRR@FAR=1e-2	FRR@FAR=1e-3
Overall	0.00033	0.00044
African_Female	0.00021	0.00091
African_Male	0.00015	0.00022
Asian_Female	0.00218	0.00437
Asian_Male	0.00000	0.00000
European_Female	0.00046	0.00058
European_Male	0.00028	0.00030
Hispanic_Female	0.00061	0.00061
Hispanic_Male	0.00020	0.00035

Table 5. Reference CosFace model (ResNet101 + CosFace Loss + DeepGlintFace) breakdown 1v1 face verification benchmark on Morph dataset. Average is indicated by "overall". ("indian" and "other" not tested due to insufficient number of images.)

FAR	FRR@FAR=1e-2	FRR@FAR=1e-3
Overall	0.00037	0.00042
African_Female	0.00019	0.00023
African_Male	0.00018	0.00019
Asian_Female	0.00218	0.00218
Asian_Male	0.00000	0.00000
European_Female	0.00048	0.00052
European_Male	0.00030	0.00030
Hispanic_Female	0.00061	0.00061
Hispanic_Male	0.00025	0.00025

Table 6. Reference SC-Arc model (ResNet101 + Sub-center Arcface Loss + DeepGlintFace) breakdown 1v1 face verification benchmark on Morph dataset. Average is indicated by "overall". ("indian" and "other" not tested due to insufficient number of images.)

prediction for recognizability across all these groups.

(3) The raw ERS distributions do not reveal a strong correlation with the model's face verification performance in each subgroup. This is evidenced by, for example, "Asian_Female" group has the highest error rate in both models, but its raw ERS distributions appear to be average among the curves.

It is worth noting that although the breakdown face verification results imply some biases of the face embedding models used, our investigation is far from thorough, given fairness is not the main focus of this work. Comprehensive analysis is needed to draw more solid conclusions, for which we refer readers to the dedicated literature on face representation learning fairness analysis.

References

[1] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[2] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[3] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[4] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006.

[6] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5651–5660, 2020.

[7] Zhifei Wang, Zhenjiang Miao, QM Jonathan Wu, Yanli Wan, and Zhen Tang. Low-resolution face recognition: a review. *The Visual Computer*, 30(4):359–386, 2014.

[8] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[9] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.