A Novel Framework for Assessment of Deep Face Recognition Systems in Realistic Conditions

Yuhang Lu, Luca Barras, Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Email: firstname.lastname@epfl.ch

***Abstract*—Deep convolutional neural networks have shown remarkable results on face recognition (FR). Despite their significant progress, the performance of current face recognition techniques is often assessed in benchmarks under not always realistic conditions. The impact of outdoor environment, post-processing operations, and unexpected human behaviors are not sufficiently studied. This paper proposes a universal methodology that systematically measures the impact of various types of influencing factors on the performance of FR methods. Based on extensive experiments and analysis, the key influencing factors are identified, highlighting the need for suitable precautions on modern FR systems. The robustness of the state-of-the-art deep face recognition techniques is further benchmarked with our assessment framework. The best-performing CNN architecture and discriminative loss function are identified, in order to better guide the deployment of an FR system in real world.**

***Index Terms*—Face Recognition, Influencing Factors, Assessment Framework, Benchmarking**

I. Introduction

Face recognition has become a prominent biometric technology in our society, frequently used in multiple areas, such as access control, video surveillance, and automatic annotation, to mention a few. With the advancement of deep convolutional neural networks (DCNNs), deep learning-based methods [1]–[5] trained on large-scale face datasets have demonstrated huge success in face recognition tasks. In recent years, the powerful network architecture [4], [6]–[8] and discriminative learning approaches [1]–[3], [5] have further boosted the performance of face recognition systems, many of them achieving more than 99% accuracy on some public face recognition benchmarks [9]–[11]. Nevertheless, despite the impressive success of deep learning-based face recognition techniques, most of the current methods are optimized and assessed under conditions that do not always match realistic situations.

Face recognition systems are, in fact, often deployed in the wild, where captured images can suffer from poor illumination conditions, sensor noises, and random occlusions. Aside from the impact of extrinsic environment, the preprocessing operations applied to the face images during transmission, for instance compression, denoising and resizing can bring adversarial artifacts and reduce the discriminative power of a face recognition model. A deep FR system can also exhibit notable biases due to the limited scale and the diversity of the training set, e.g. show lower accuracy in certain gender, ethnicity, age groups [12]–[14], or in specific poses [15]. Therefore, it is crucial to understand in what circumstances and to which extent, the performance of a face recognition system could be impacted by influencing factors.

Analysis on the robustness of CNN-based face recognition models has been reported by [16]–[18], where the authors assessed the impact of face variations caused by standard image processing operation, illumination, occlusion, and misalignment. But their experiments were conducted on underperforming FR methods using less challenging datasets, and more importantly, they only considered limited types of influencing factors. Our work provides a more general assessment framework, which takes into account a wider range of factors and additionally provides a robustness benchmark for the most recent state-of-the-art FR techniques.

To sum up, the contributions made in this paper can be summarized as follow:

- A universal assessment framework is proposed to measure the impact of different types of influencing factors on the performance of deep face recognition techniques. The source code for the framework and evaluation protocols will be released to benefit the community.
- The significant influencing factors are identified based on the assessment framework and extensive experiments, which highlights the need for suitable precaution on current FR systems.
- A benchmark is performed on seven state-of-the-art deep FR techniques to understand the robustness of different CNN architectures and discriminative loss functions.

II. Related Work

Understanding and explaining the behavior of learning-based face recognition systems requires comprehension of the influencing factors that impact the decision. [16], [17] investigated the robustness of a DCNN-based face recognition algorithm under facial appearance changes caused by illumination, pose, occlusion, and common image processing. The authors collected various task-specific databases, e.g. AR face database [19] and FERET [20] dataset, to identify the impact of different influencing factors. However, these datasets only

The authors acknowledge support from CHIST-ERA project XAIface (CHIST-ERA-19-XAI-011) with funding from the Swiss National Science Foundation (SNSF) under grant number 20CH21 195532.

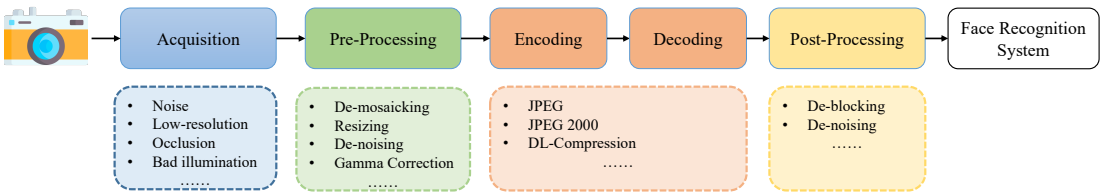

Fig. 1: A realistic data acquisition and transmission pipeline, including various natural image distortions and processing operations.

contain limited number of subjects, i.e. less than 100, which can result in biased conclusions.

Another research in [18] grouped the possible influencing factors into two categories: image quality variants, e.g. blurring, JPEG compression, salt-and-pepper noise, and missing data, and model-related variants, e.g. model architecture. The impact of these factors on four deep face recognition models, e.g. GoogLeNet [21], VGGFace [22], have been analyzed with LFW dataset. Nevertheless, the tested deep models are outdated and the selected dataset is relatively small. In fact, current state-of-the-art face recognition methods [1]–[3] have already achieved rather robust results on classical LFW dataset [9] and its variations [10], [11]. Moreover, the experiments and analyses are only performed on face verification task, while ignoring the face identification task that is different and of equal importance. Our work focuses on the most recent deep face recognition solutions and conduct experiments with larger-scale datasets and multiple performance metrics.

More recently, researchers have focused on more specific factors that could affect a deep face recognition system. Vítor et al. [12] model the gender distribution in the training dataset and study how the under-represented female identities in training data will degrade the female recognition accuracy in testing phase. Similar work has been undertaken to analyze the impact of biased distribution of age [14] and ethnicity [13] in training images. In addition, deep model-specific factors, such as loss function, have also shown a different tendency towards data distortions [15].

While numerous work has been reported in this area, there is a lack of generic and flexible methodology that systematically measures the influence of a diverse range of factors from different dimensions. In this work, a new assessment framework for deep face recognition models is introduced that both identifies possible influencing factors and benchmarks model performance under different realistic conditions.

III. Methodology of Performance Assessment

In this section, the potential influencing factors are categorized into two groups following the suggestion in [23]. Then, a universal assessment approach is introduced, which efficiently measures the impact of different types of factors using the same framework. Finally, a rigorous benchmarking approach is proposed to assess the robustness of deep FR models.

A. Data Quality-related Factors

The most recent face recognition approaches are data-driven and often trained on millions of constrained and good-quality face images. However, face recognition systems are often deployed in real-life applications with very challenging conditions. The quality of the perceived face image depends heavily on the environment and internal processing approaches during delivery. Fig. 1 depicts a typical data acquisition and transmission pipeline in real world. In this paper, the term 'Probe face' is used to refer to the perceived image to be identified, while 'Gallery faces' refer to those stored in the database. The influencing factors will only distort the probe faces. The most prominent factors are listed as follows.

Noise: Noise is a typical distortion especially when images are captured in a low illumination condition. This is often the case when an FR system is deployed in the wild. To simulate the noise, an additive Gaussian white noise is applied to the probe face and the pixel values are clipped to [0, 255]. In this paper, the variance value σ is selected in a range from 5 to 50. In addition, Poissonian-Gaussian noise [24] is also included to better reflect the realistic noise levels, whose parameters are learned from a group of real noisy pictures.

Resolution: When compared to standard-quality images, low-resolution face loses discriminative identity information, significantly reducing the accuracy of a face recognition system. This is often the case in an outdoor environment. In this framework, the low-resolution effect is synthesized by downsampling face images using bicubic interpolation with the scales of 2, 4, 6 and 8.

Enhancement: In realistic situations, the image captured in the wild can suffer from poor illumination. Image enhancement is a very frequently used technique in order to adjust the image for better display or further image analysis. The contrast and brightness of probe images are modified through both linear and nonlinear adjustments. The former simply adds or reduces a constant pixel value in HSV space while the latter adopts gamma correction.

Compression: Lossy compression is widely applied to digital image and video processing to ease transmission or storage. In this framework, the JPEG compression artifacts are used to probe images and the impact of different quality factors, i.e. from 10 to 95, to recognition system is assessed. As AI-based compression techniques are becoming increasingly popular in the community, a deep image compression technique [25] is also applied to the probe set and compares its impact to that of JPEG.

Denoising: Images captured by a digital camera often pick up noise from various sources. A typical way to reduce noise is by smoothing, which is a low-pass filtering applied to the image but tends to blur the image. The blurring effect is simulated by applying Gaussian filters with kernel size σ

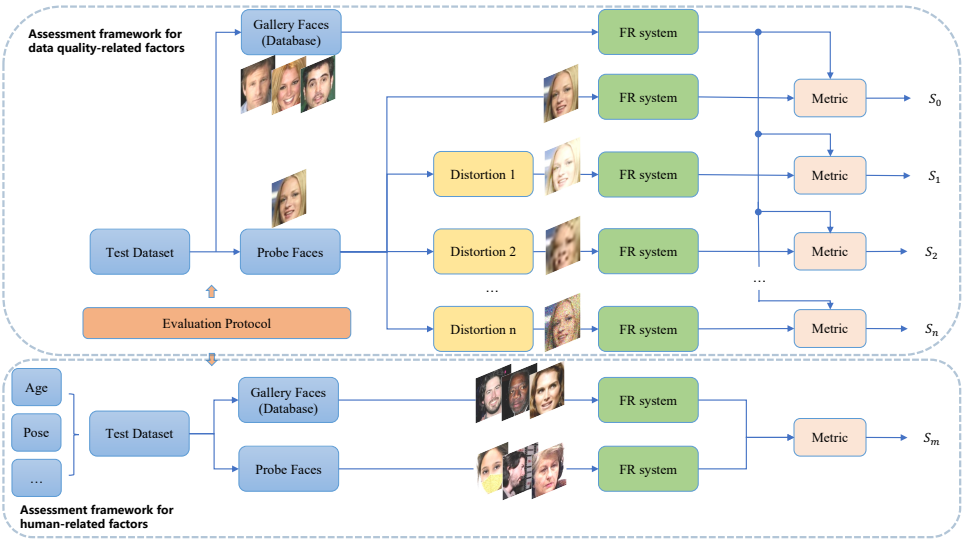

Fig. 2: The proposed assessment framework contains two parts. The first part measures the impact of the data-quality related influencing factors while the second part measures for human-related factors.

ranging from 3 to 11. Meanwhile, learning-based denoising techniques are gradually deployed in practice. They recover a noisy image with higher quality but often bring unpredictable artifacts. The impact of applying DnCNN technique [26] is assessed in our framework.

Combination: As is depicted in Fig. 1, it is even more common that the captured face data suffers from multiple distortion and processing operations. The mixture of the distortions and operations above is considered, making the test data better reflect complex real-world scenarios.

B. Human-related Factors

Besides the possible influence from extrinsic environment and internal processing operations, the mismatching between the probe and gallery face images could also potentially result in poor behavior of a face recognition system. The impact of the following factors is measured with ArcFace [1] method.

Occlusion: In real world, face occlusion is very common with the wearing of glasses and more recently masks due to the COVID-19. To measure the impact of these occlusions on a face recognition model, an evaluation protocol for Webface-OCC dataset [27] is created, which synthesize glasses and mask occlusions on the CASIA-Webface dataset [28].

Age: Faces change through time and it can be challenging to recognize a person after a large gap in time. AgeDB [29] contains images of celebrities of various ages. Evaluation protocols for different age gaps are created to measure the influence of age variation on face recognition. Specifically, the age for identities in gallery set is fixed in a certain range, while probe faces are, for example, ten years older.

Ethnicity: In real life, both humans and deep FR models are often biased in identifying faces from different ethnicity groups, because certain groups appear more frequently in their environment and training datasets. The impact of ethnicity distribution is studied by evaluating model performance on VMER [30], a dataset based on VGGFace2 [31] dataset but filter identities into 4 ethnicity groups: African, East-Asian, Caucasian, Indian. An evaluation protocol is created that performs face recognition on the four groups separately.

Pose variation: In real-world situations, faces are captured in various poses, which causes geometric distortions during alignment and makes the face recognition more challenging. To measure the impact of the pose variation on the face recognition models, LFR [32] dataset is used which contains celebrities under three poses, i.e. left, front, and right. To better reflect realistic conditions, the gallery set only contains front faces, while the probe set will change the face poses.

C. Assessment Framework

The performance of a face recognition system is traditionally assessed through the following steps. First, face images from a test dataset split into probe and gallery groups according to specific recognition task and corresponding evaluation protocols. Then the FR model extracts the deep features of both probe and gallery faces and calculates the similarity between them. Finally, performance metrics are computed and reported based on the similarity score and evaluation protocol.

This work extends the traditional assessment approach based on the two categories of influencing factors defined above. Fig. 2 divides the assessment framework into two parts to make it easier to understand. As illustrated in the figure, to measure the impact of data quality-related factors, the corresponding distortions, e.g. compression, blurring effect, etc., are separately applied only to the probe faces before feeding to the model and comparing with the standard gallery faces. As for the factors that rely on human characteristics and behaviors, it is challenging to synthesize such variants on each subject. This assessment framework adopts multiple specific task-oriented datasets to measure their impact respectively and design evaluation protocols from scratch. It is worth noting that the evaluation protocols always put standard faces to gallery set, for example front pose, mid-age photos without oc-

TABLE I: 1:1 face verification evaluation of four state-of-the-art face recognition methods using our assessment framework. The impact of various types influencing factors at different severity levels are presented. This work reports 1:1 verification accuracy for experiments on LFW and TAR (@FAR=1e-4) on IJB-C. Notations are explained as follows. Low-Res: Low resolution; DL-Comp: Deep learning-based compression; Gau: Gaussian; GammaCorr: Gamma correction.

Backbone	Loss	Dataset	Unaltered	Gau Noise			Low-Res			Gamma Corr				Gau_Blur			DnCNN	JPEG			DL-Comp		
				5	30	50	x4	x6	x8	0.1	0.75	2.5	5	3	7	11		10	30	60	High	Med	Low
ResNet	MV-Softmax	LFW Acc(%)	99.83	99.79	99.45	98.30	99.35	94.90	86.05	99.74	99.83	99.81	99.40	99.80	99.56	98.56	99.80	99.55	99.81	99.81	99.83	99.76	98.33
	CircleLoss		99.76	99.81	99.56	97.80	99.58	95.63	87.88	99.76	99.76	99.75	99.51	99.80	99.58	98.71	99.78	99.56	99.80	99.76	99.78	99.70	98.70
	ArcFace		99.78	99.75	99.56	98.90	99.31	95.03	87.78	99.70	99.81	99.78	99.43	99.73	99.56	98.75	99.77	99.56	99.80	99.85	99.73	99.71	98.63
	MagFace		99.75	99.73	99.51	99.18	99.38	95.30	87.53	99.76	99.76	99.73	99.43	99.80	99.67	99.06	99.75	99.56	99.78	99.73	99.75	99.76	98.66
	MV-Softmax	IJB-C TAR(%)	94.94	94.61	89.54	76.47	90.62	64.49	28.86	92.45	94.87	94.39	86.62	94.92	92.17	81.63	94.19	89.22	94.38	94.71	-	-	-
	CircleLoss		95.77	95.54	90.62	75.68	92.57	71.09	35.74	93.95	95.76	95.25	87.17	95.67	93.44	84.26	95.13	92.32	95.32	95.62	-	-	-
	ArcFace		94.70	94.41	89.76	79.84	89.87	62.92	28.43	91.98	94.62	94.02	86.31	94.62	91.63	80.13	93.90	90.19	94.05	94.43	-	-	-
	MagFace		95.48	95.29	91.98	84.89	91.44	65.72	21.27	93.10	95.38	94.99	88.43	95.31	93.02	82.83	94.77	91.97	94.93	95.32	-	-	-

clusion, in order to reflect the situation in real-world database. Last but not least, by simply replacing the 'FR system' in the above workflow, the proposed assessment framework also identifies the difference when changing the architecture or training loss of a deep face recognition model. It provides a robustness benchmark for deep models and gives valuable instructions of how to choose face recognition techniques in real-world applications.

D. Robustness Benchmark for Face Recognition Models

Modern DCNN-based face recognition techniques often improve their accuracy by adopting more powerful backbone networks as feature extractor or proposing more discriminative loss functions. But the current best-performed architecture and loss function may not necessarily be robust enough towards different influencing factors. Here, robustness of the following four popular and advanced network architectures and loss functions are benchmarked separately in realistic scenarios.

Network architecture: The CNN architectures selected by past state-of-the-art [16], [18] are neither powerful enough nor efficient in computation and not popular anymore in face recognition community. Our assessment framework benchmarks the robustness of two powerful and two light-weight CNN backbone networks: ResNet-152 [6], since released, is one of the most widely used CNN architecture. EfficientNet-B0 [7] is one of the state-of-the-art on ImageNet [33] benchmark. LightCNN [8] and MobileFaceNet [4] are popular light-weight feature extractors and are often deployed in mobile devices due to low computational cost and storage.

Discriminative loss function: Training loss plays an important role in learning discriminative facial representations. Compared to previous work [15], four popular loss functions are selected, namely MV-Softmax [5], ArcFace [1], CircleLoss [3], and MagFace [2], among which ArcFace is perhaps the most widely adopted loss function for face recognition system.

IV. Experiments

In this section, the implementation details of our extensive experiments are introduced. Then the experiment results produced by our assessment framework are presented.

A. Implementation Details

1) Dataset: This work employs the cleaned version of MS-Celeb-1M dataset [34], often referred as MS1M-V2 [1], as training dataset. It contains 3.3M images and 72.7k identities and is one of the largest training sets for face recognition task. The LFW [9] and IJB-C [35] benchmarks are adopted as base test data. In addition, WebFace-OCC [27], LFR [32], AgeDB [29], and VMER [30] are used to identify the influence of occlusion, pose variations, biased age and ethnicity distributions. During training and testing, all the images are cropped to 112x112 pixels and aligned based on given facial landmarks.

2) Training Details: All the models are trained from scratch for 17 epochs by stochastic gradient descent and optimized with the same parameters for a fair comparison. The initial learning rate is set to 0.1 and is divided by 10 at 10, 13, and 16 epochs. The feature embedding size is 512 for all the methods. In the robustness benchmark settings, the ArcFace loss and ResNet are used respectively for the experiment on backbones and loss functions due to their broad popularity in the community.

3) Evaluation Details: Two face recognition tasks are performed during evaluation, namely 1:1 verification and 1:n identification. The former refers to verifying whether a pair of faces belong to the same subject while the latter aims at identifying the probe face from a database. For data quality-related factors, the standard protocols along with the datasets are adopted. For human-related factors, bias in the test dataset has been avoided and both verification and identification protocols are created from scratch.

4) Performance Metrics: Face verification system is often assessed by mean accuracy (ACC), and receiver operating characteristic (ROC). For experiment on IJB-C dataset, true accept rate (TAR) is additionally reported when false accept rate (FAR) is very low, namely 1e-4, which is more popular in nowadays biometric applications. In the closed-set identification scenario, the commonly used metric Rank-N is reported, which measures on what percentage of probe searches return results within the top k rank-ordered portion.

B. Results of Data Quality-related Factors

The assessment framework measures the performance deterioration caused by the previously described image quality related factors and presents a subset of them in Table I.

The verification accuracy for experiments on LFW dataset is reported. Unlike the results from previous work [18], LFW dataset is less challenging to modern FR methods. All the models show relatively robust results on it, except for extremely low-resolution faces. Meanwhile, it is interesting to observe that the learning-based compression brings less

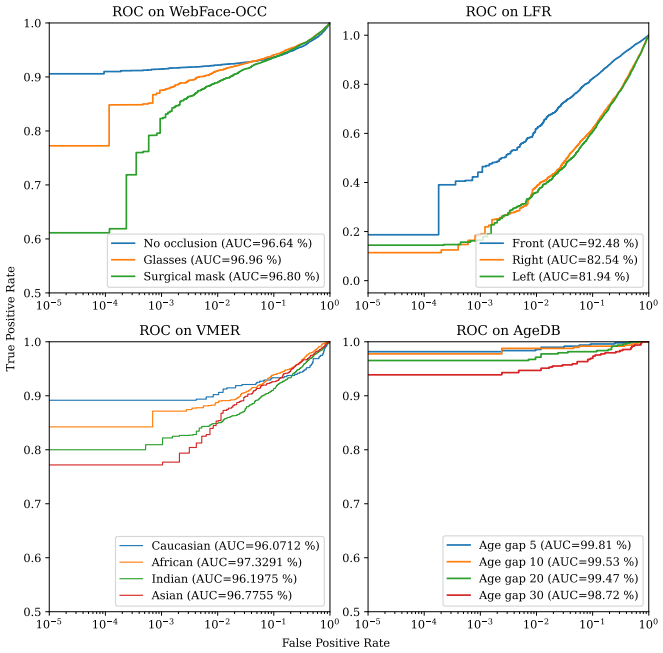

Fig. 3: ROC curves of 1:1 verification protocols of four different types of human-related variants. The figures use logarithmic scale to better show the difference.

TABLE II: 1:n face identification evaluation of ArcFace model on WebFace-Occ dataset. Rank-1, 5, 10 identification accuracy are reported.

Method	Occlusion	Rank-N		
		Rank-1	Rank-5	Rank-10
ResNet-ArcFace	w/o Occ	89.05	90.71	91.00
	Glasses	84.10	88.61	89.80
	Surgical Mask	74.85	82.83	84.96

negative impact to FR system than conventional JPEG compression when they achieve comparable bit-rates, for example 'JPEG 10' and 'DL-Comp Med'. Similarly, the learning-based denoising algorithm maintains more identity information while better reduces noise than Gaussian blur operation.

In Table I, the TAR (@FAR=1e-4) scores on the IJBC dataset are presented. As a result, low-resolution effect brings the most significant influence to all face recognition techniques. Noise corruption or blurry effect will also notably degrade the performance. On the contrary, compression artifacts and contrast changes show relatively less influence.

C. Results of Human-related Factors

In this section, the impact of occlusion, pose variation, age and ethnicity distribution is summarized.

By observing the results in Table II and Fig. 3, the model accuracy is degraded by both types of occlusion, although it is more robust to glasses occlusion. It is reasonable because people wearing masks appears less common in the training set. But with the spread of Covid-19, it is well-needed for the current FR systems to better identify people who wear masks.

TABLE III: 1:n face identification evaluation of ArcFace model on LFR dataset. Rank-1, 5, 10 identification accuracy are reported.

Method	Probe-set	Rank-N		
		Rank-1	Rank-5	Rank-10
ResNet-ArcFace	Front	46.73	57.64	62.21
	Left	20.25	30.50	35.38
	Right	19.42	29.50	34.01

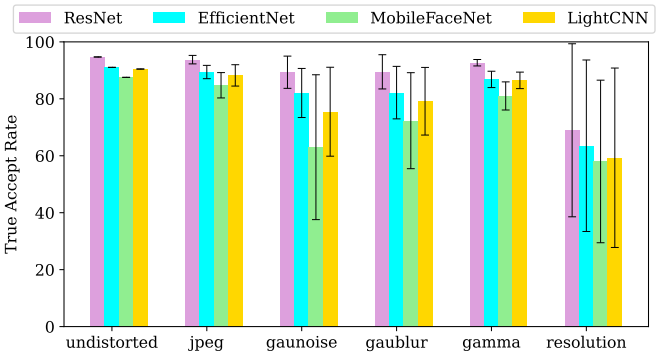

Fig. 4: 1:1 face verification results of four backbone architectures. The reported result for each factor is averaged for different severities.

The ROC curve in Fig. 3 and Table III clearly show the model is not invariant to the pose. There is a big performance gap between front and side faces. On the other hand, there is small difference between the impact from right and left poses.

The experiment results on VMER dataset show that the model recognizes better the Caucasian group as expected while showing relatively lower accuracy on India and East-Asian, indicating the imbalance of ethnicity in training data.

According to the results on AgeDB in Fig. 3, the bigger the age gap, the more the discriminative features of the probe face will change and the lower the performance. But their impact on modern FR techniques is still limited.

D. Results of Robustness Benchmark

The proposed assessment framework measures the robustness of the backbone CNN architectures and loss functions towards realistic influencing factors.

Table I compares the results for different loss functions when fixing ResNet as backbone. In general, the recognition performance on distorted data of the four methods is proportional to the performance on clean data but with a few exceptions. For example, Magface is clearly more robust to noise distortion than others when increasing the severity level. But it has lower scores on low-resolution data.

Similarly, the same ArcFace loss is used while changing the backbone networks in order to identify the best-performed architecture. Fig. 4 illustrates the TAR scores averaged for each image distortion category for four deep face recognition models. The error bar indicates how the model performance varies when changing the severity of one type of distortion. The ResNet outperforms all other architectures in terms of TAR score and meanwhile, shows a much smaller deviation

toward different noise levels and compression ratios. Due to the light-weight nature and low network capacity, the MobileFaceNet and LightCNN network is less robust than the other two. But LightCNN is clearly a better option out of the two for real-time application in terms of robustness.

Compared to loss function, backbone networks are more sensitive to realistic distortions. Therefore, a powerful backbone network should be first considered in real-world deployment instead of the loss function.

V. Conclusion

To better understand the behavior of learning-based face recognition systems, this paper analyzes a large number of influencing factors. A generic performance assessment methodology is proposed that systematically measures the influence of these factors on FR systems. Extensive experiments show that modern face recognition techniques are robust to many disturbances, but still prone to performance deterioration under low resolution and noisy data and ill-posed faces. Another interesting finding is that ResNet backbone is generally more robust than other architectures when facing distortions.

References

[1] 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*, 2019, pp. 4690–4699.

[2] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 225–14 234.

[3] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6398–6407.

[4] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[5] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 241–12 248.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[8] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[9] 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.

[10] 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.

[11] 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*, vol. 5, p. 7, 2018.

[12] V. Albiero, K. Zhang, and K. W. Bowyer, "How does gender balance in training data affect face recognition accuracy?" in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020, pp. 1–10.

[13] J. Coe and M. Atay, "Evaluating impact of race in facial recognition across machine learning and deep learning algorithms," *Computers*, vol. 10, no. 9, p. 113, 2021.

[14] V. Albiero, K. Bowyer, K. Vangara, and M. King, "Does face recognition accuracy get better with age? deep face matchers say no," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 261–269.

[15] G.-S. J. Hsu, H.-Y. Wu, and M. H. Yap, "A comprehensive study on loss functions for cross-factor face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 826–827.

[16] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How image degradations affect deep cnn-based face recognition?" in *2016 international conference of the biometrics special interest group (BIOSIG)*. IEEE, 2016, pp. 1–5.

[17] M. Mehdipour Ghazi and H. Kemal Ekenel, "A comprehensive analysis of deep learning based representation for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 34–41.

[18] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations," *Iet Biometrics*, vol. 7, no. 1, pp. 81–89, 2018.

[19] A. Martinez and R. Benavente, "The ar face database: Cvc technical report, 24," 1998.

[20] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[21] 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*, 2016, pp. 2818–2826.

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[23] F. Pierre. (2021) Xaiface deliverable. [Online]. Available: https://xaiface.eurecom.fr/

[24] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian, "Practical poissonian-gaussian noise modeling and fitting for single-image raw-data," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1737–1754, 2008.

[25] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[26] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[27] B. Huang, Z. Wang, G. Wang, K. Jiang, K. Zeng, Z. Han, X. Tian, and Y. Yang, "When face recognition meets occlusion: A new benchmark," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4240–4244.

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

[29] 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*, 2017, pp. 51–59.

[30] A. Greco, G. Percannella, M. Vento, and V. Vigilante, "Benchmarking deep network architectures for ethnicity recognition using a new large face dataset," *Machine Vision and Applications*, 2020.

[31] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[32] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "Lfr face dataset: Left-front-right dataset for pose-invariant face recognition in the wild," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. IEEE, 2020, pp. 124–130.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[34] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[35] 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)*. IEEE, 2018, pp. 158–165.