Evaluating Image Quality Estimators for Face Matching

Praneet Singh Haoyu Chen Edward J. Delp Amy R. Reibman

Elmore Family School of Electrical and Computer Engineering, Purdue University

Abstract

Understanding the quality of a face image can be useful for improving the performance of automated face matching systems. With the increasing number of face quality estimators (QEs) being proposed recently, it is important to have systematic methods to evaluate and compare the performance of these QEs. In this paper, we describe two existing strategies for evaluating face QEs, and propose several new approaches. Our new approaches focus on targeted QE evaluation using carefully constructed image datasets. We show that these strategies lead to important insights about the effectiveness of existing face QEs.

1. Introduction

Quality estimation is an integral component of systems for processing images and videos [11]. While conventional quality focuses on perceptual quality, for specific tasks, many more things can affect quality (e.g., the capture, and post-processing, the environment, background, and lighting). Input images that have bad quality will likely result in poor performance or an unverifiable result.

In this paper, we focus on the task of face recognition, or face matching, and consider how to rigorously evaluate the effectiveness of a QE for its use in a real system. To actually improve system performance, a face QE needs to be accurate and create actionable and interpretable output scores. If a QE computes a particular score, how should that score be interpreted and what action should be taken?

Perceptual QEs are primarily evaluated by computing the correlation between the computed quality score and subjective ground truth from human viewers. Evaluation methods to prove the effectiveness of QEs for other non-perceptual tasks, however, are still being defined. QEs for fingerprint images have been evaluated using Error vs. Reject (EvR) curves [8]. Face QEs have subsequently also been evaluated using EvR [19, 18] as well as performance curves for Best, Medium, and Worst (BMW) subsets[12]. A QE for person re-identification (re-id) [2] has also been evaluated using both EvR and BMW.

None of the current evaluation methods for face QEs inform about when a QE is not effective so that its design can be improved. They all modify a collection of data according to the QE scores and compute overall performance on the remaining data. Therefore, we propose new evaluation strategies for face-matching QEs that move beyond evaluation strategies of "how well does my QE work on an existing dataset". Our proposed evaluation strategies are designed to be explanatory and to explore QE effectiveness at evaluating the impact of a specific aspect of face quality.

Note that methods to evaluate QE performance are distinct from methods to evaluate the task performance itself. The latter is typically evaluated using accuracy on a fixed dataset like LFW [13] or IJB-C [15], with accuracy measured using true positive rate (TPR) for a chosen false positive rate (FPR). In addition, considering interpretability of the output scores of a QE allows appropriate action to be taken based on a particular output score. This is not the same as the current research thrust of interpretable machine learning algorithms.

In this paper, we describe both existing methods and new methods to evaluate QE performance for the task of face matching for identification, verification, or recognition. In particular, we seek where a QE fails rather than where it succeeds, to obtain insights of when a QE is effective or not. Section 2 describes the six quality estimators we consider in this paper, all of which take a single input image and output a quality score. Section 3 explores the existing evaluation protocols and some insights that can be obtained by applying them to the six QEs. We also provide an upper bound for these protocols to understand how much further improvement may be possible for a face QE. Morover, this section offers a caution about how implementation mistakes could affect the correctness of these evaluations.

Next, in Section 4 we propose new evaluation protocols that are designed to target two potential weaknesses in existing face QEs. The first test explores consistent and interpretable scores given distorted or perturbed images, while the second test explores robustness to two face alignment strategies. These two evaluations are only examplars, since a wide range of additional protocols can be envisioned using these principles.

2. Quality Estimators

2.1. QEs Designed For Face Matching

Three recent no-reference learning based approaches for face image quality estimation considered here are FaceQNet [12], SER-FIQ [19] and SDD-FIQA [18].

FaceQNet [12] is a supervised approach proposed to correlate the quality of a face image to its expected accuracy for face recognition. It employs a BioLab-ICAO framework to create image quality ground-truths such that the quality score is related to the ICAO compliance level. State-of-the-art deep learning frameworks are then trained to predict image quality scores.

In contrast, SER-FIQ [19] is an unsupervised approach that uses feature vector robustness to assign a quality score to face images. Here, face images are passed several times through a recognition network like ArcFace with dropout [6] enabled. Dropout introduces randomness into the feature vectors generated for a given image. The SER-FIQ quality score of a face image is the Euclidean distance between the different feature vectors. A lower variation is assigned a higher SER-FIQ score. Because SER-FIQ requires an underlying network, the training details regarding this network are necessary for interpreting the results.

SDD-FIQA [18] is another unsupervised approach for face image quality estimation. It relies on the same basic principles used to design recent learning-based face recognition systems, namely that a high-quality face image should be similar to its intra-class samples and dissimilar to its inter-class samples. For each face image, SDD-FIQA first computes a similarity distribution distance, using the Wasserstein Distance, between its intra-class and inter-class distributions. This distance is used as a ground-truth to train a regression network. Similar to SER-FIQ, SDD-FIQA relies on an underlying recognition network to create intra-class and inter-class distributions that help generate quality scores. In addition, SDD-FIQA also depends on a fixed database to define the intra-class and inter-class members used to create its ground truth.

Both SER-FIQ and SDD-FIQA rely on an underlying face recognition network to generate their quality scores, specifically ArcFace [6] with a ResNet [10] backbone trained on MS1-MV2 [9] dataset. To ensure fair evaluation and create some separation between the QEs and the matching system, for our face matcher, we use ArcFace [6] with MobileFaceNet [3] as the backbone trained on the MS1-MV1 dataset [5]. This change in network backbones and training dataset reduces dependencies between the face matching system and the QEs.

2.2. QEs Designed For Perceptual Quality

We also consider three conventional no-reference image quality estimators: BRISQUE [16], NIQE [17], and PIQUE [20]. These QEs are designed to assess perceptual, not task-related quality. For example, they consider the question "do people think this image has high quality?" Including these in our experiments provides a useful contrast to illustrate how effective face QEs are in the domain of face matching.

Both BRISQUE and NIQE use statistical features to quantify the naturalness of an image. A distinction between the two is that BRISQUE is trained with collection of natural and distorted images, whereas NIQE is solely trained with a collection of natural images. PIQUE[1] does not require training, but instead extracts block-based spatial features to decide whether distortion is present.

3. Evaluation Protocols

In this section, we present two protocols that have been previously used to evaluate face matching QEs. For the first protocol we also describe a warning about its implementation details.

We evaluate all experiments in this section using the IJB-C dataset [15], specifically using only the image subset of IJB-C that contains 17,474 images of 3,464 subjects. We detect faces in these images using the MTCNN detector [21] and then align them using similarity transformations before applying the ArcFace [6] face matcher. All quality scores are estimated on the aligned IJB-C face images.

3.1. Error Versus Reject (EvR) Protocol

This experiment characterizes whether a quality measure can effectively rank images by their usefulness and potential reliability to a system. In this experiment, we rank-order the query images according to each QE. We reject a fraction of query images based on each QE, and evaluate the performance on the dataset with those query images removed. Note that this is an evaluation protocol, and may not be something that is implemented in a real system. However, as a protocol, it allows us to observe both how well the QE orders low-quality images (by reading from the left of the plot), as well as how well it orders high-quality images (by reading from the right of the plot). This protocol was introduced by [8] and has been extensively used to evaluate biometric and face quality measures [8, 14, 7, 1, 19].

Figure 1 shows the results from applying this protocol to the IJB-C database. This uses over 33 million image pairs and 6 QEs. We see that all face-based QEs (solid lines) perform relatively well. Performance, in terms of FNMR, improves dramatically when only 5% of the images in the

[1] PIQUE is the name created by the authors, but the method has recently been referred to as PIQE.

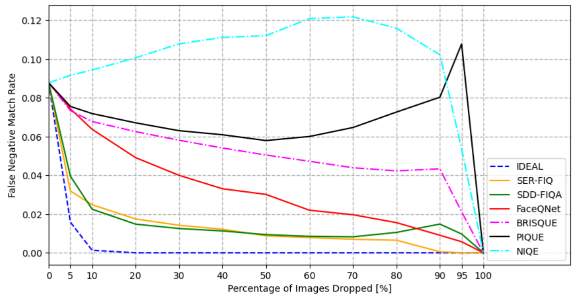
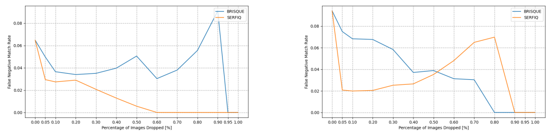


Figure 1: EvR on the full set of pairs from IJB-C image dataset.



(a) Run 1

(b) Run 2

Figure 2: EvR using random subset of pairs (out of 330 million pairs) leads to incorrect estimation of performance

dataset are rejected. SER-FIQ performs best of the three face QEs for the high-quality images (right portion of plot), and FaceQNet also begins to outperform SDD-FIQA in this region. However, the perceptual QEs fare less well in this protocol. BRISQUE performs best, and the performance with NIQE actually degrades, except for the high-quality region where more than 90% of low-quality images are dropped. PIQUE performs reasonably well to begin with, but its performance begins to suffer when between 60-90% of query images are discarded.

In addition, Figure 1 shows the Ideal case, where images are dropped from the comparison based on a plausible best-case scenario. For every subject/identity, we compute the cosine similarity between all their images. The average of these similarities describes how closely each image of this subject is in the feature space, relative to other images of the same subject. Images are then dropped according to these average similarity scores.

It is notable that the performance with the Ideal case is extremely good when only 10% of the worst images are dropped from the dataset. This implies that there are actually quite few low-quality images in the IJB-C image dataset, and that overall dataset performance is governed by a small fraction of low-quality images.

This protocol requires using the exhaustive set of all possible pairs of images to obtain an accurate evaluation. We demonstrate this with the following experiment. We sample 2500 images at random from the 17,474 images in IJB-C and create all possible pairs among these 2500 images. We evaluate the EvR protocol for both SER-FIQ and BRISQUE using this subset. Results are shown in Figure 2(a). We repeat the experiment with a different sample of 2500 images to obtain the results in Figure 2(b). Both results lead to quite different conclusions, and neither are similar to the performance demonstrated in Figure 1.

Therefore, it is critical when evaluating the EvR protocol to use the exhaustive set of pairs in a dataset to obtain an accurate assessment. Unfortunately, this can be prohibitive. The next commonly implemented protocol reduces the required number of pairs significantly.

3.2. Best-Middle-Worst (BMW) Protocol

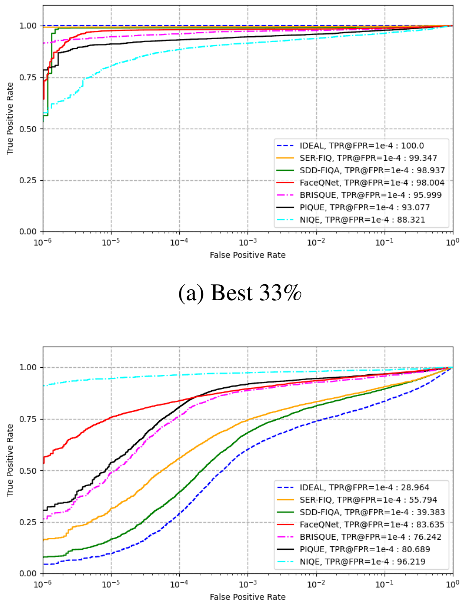
A second commonly used protocol to evaluate face-based QE performance we term the “Best-Middle-Worst” (BMW) performance protocol. In this protocol, we partition a dataset into three sets based on the quality score, and then demonstrate that the subsets (ideally) create ordered performance. This protocol has been used in Face Q-NET, NIST fingerprint quality. Relative to the EvR protocol, this protocol requires the use of fewer pairs, because it only considers pairs within each partition.

There are two ways to gain insights from the results of this protocol for performance analysis. The first is to compare how well all QEs do for their best, middle, and worst partitions. The second is to compare, for a given QE, its best, middle, and worst partitions.

Figure 3 shows the performance of the best and worst subsets, when each subset is chosen according to a particular QE. Performance for the Ideal case (defined above) is also shown. For the subset of images chosen as Best, for each QE, which is shown in Figure 3(a), we see that both SER-FIQ and SDD-FIQA perform well across all FPR values, although SER-FIQ outperforms SDD-FIQA at the lowest FPRs. FaceQNet performs quite well for FPRs greater than 10^{-5}. Among the three perceptual QEs, BRISQUE performs reasonably well at choosing the Best subset, and actually outperforms SDD-FIQA and FaceQNet for the lowest FPRs. The worst performer is NIQE.

Considering Figure 3(b), which shows the performance on the subsets that each QE selects as Worst, we see that SDD-FIQA is most effective, since its chosen Worst subset has the worst performance across the range. Both BRISQUE and PIQUE perform reasonably well, outperforming FaceQNet for the lowest FPRs. The worst performer is NIQE.

It is more difficult to meaningfully compare the performance across QEs for the middle portion of this protocol, because the expected performance of this subset should neither be good nor bad. These results are omitted. Instead, to view the performance of the Middle subset in context, it is useful to compare performance across Best, Middle, and Worst for each QE separately. This is shown in Figure 4 for the three face QEs, and allows us to observe how well each



(a) Best 33%

(b) Worst 33%

Figure 3: BMW: matching performance on (a) Best, (b) Worst subsets, across QEs

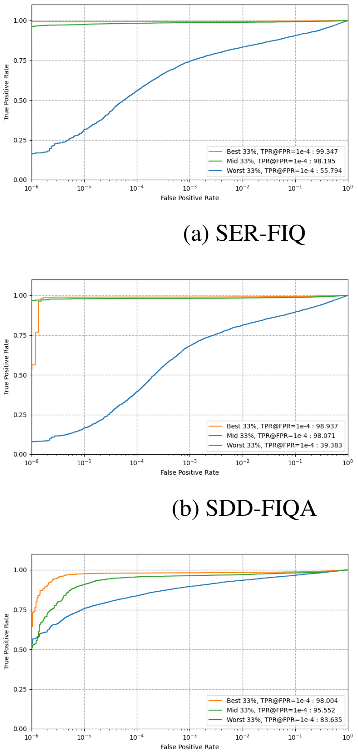
QE does in rank ordering images of best and worst quality.

In Figure 4(a), we see again that SER-FIQ is highly effective at separating the best and worst images in the dataset, with the Middle subset having slightly worse performance than the Best subset. Figure 4(b) demonstrates that SDD-FIQA is very effective at separating the Best from the Worst, but at the lowest FPRs its Middle subset actually has better performance than its Best subset. Figure 4(c) shows that FaceQNet does partition the sets relatively well, within itself, but the performance of all subsets is significantly less distinct than for the other two face QEs.

Chen et al. [2] proposed a modification on the basic BMW protocol, in the context of the task of person re-identification (re-id). Instead of explicitly considering matching performance on each subset, they evaluate the robustness of each subset to further degradations. In particular, they show that the images in the best partition are more robust to JPEG compression than those in the Middle or Worst partitions, which could lead to application in real system by adjusting compression without compromising task performance. However, as the image of interest (face image in our case) gets smaller, the implication of this evaluation protocol diminishes.

4. Targeted Testing Strategies

In this section, we describe two new evaluation protocols, that serve as examples for a more targeted testing strategy for face QEs. These new evaluation protocols are essential because the current ones do not provide insights or information about where a QE is not providing a meaningful score. Our new methods are designed to determine if a QE is robust to explainable changes to the input image. A similar strategy has been applied to evaluate perceptual QEs in [4]. Here, we consider the impact of perturbations like blur, noise, and JPEG compression, as well as robustness to two different face alignment strategies. We do this by evaluating each QE over carefully constructed subsets of face images.



(a) SER-FIQ

(b) SDD-FIQA

(c) FaceQNet

Figure 4: BMW: matching performance on Best, Middle, and Worst subsets for three face QEs

4.1. Adding Targeted Perturbations

We begin by investigating how robust a QE is when comparing face images under different perturbation types, across a range of different subjects. Specifically, we consider the binary classification question – can a QE consistently predict when performance degradation happens across multiple perturbations and across multiple subjects?

We consider 9 types of synthetic perturbations, as detailed and illustrated in Figure 5. We randomly select 100 subjects from the IJB-C image dataset and apply multiple levels of each perturbation. Each of the perturbed images,

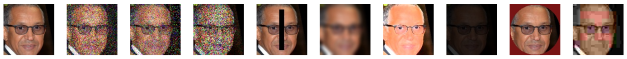


Figure 5: Perturbations used: original, Gaussian noise, salt & pepper noise (snp), speckle noise, occlusion, blur, brightening, darkening, colored border, JPEG compression (from left to right)

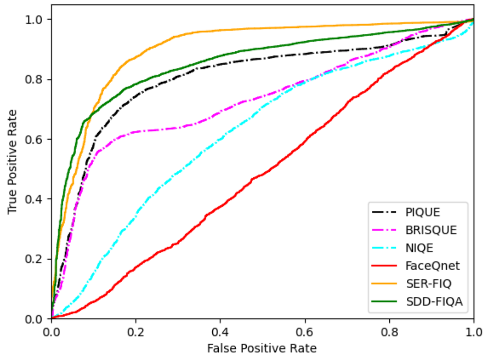
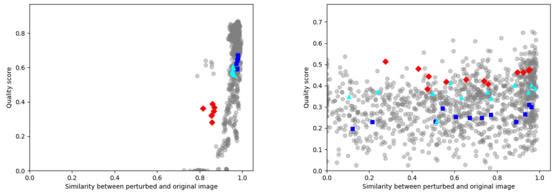


Figure 6: ROC curve predicting performance degradation

as well as the original image, is then associated with two performance-related metrics. The first is TPR @ FPR=1e-3 when comparing to every other face image in IJB-C dataset. The second is the cosine similarity between the perturbed image and the original image in feature space, which describes how much the perturbed image deviates from the original in feature space. The targeted test explores whether a QE reflects this difference.

To characterize the effectiveness of each quality metric regarding whether face-matching performance will degrade (i.e., meaningful degradation), we use an ROC curve. The results are shown in Figure 6 for each QE considered. We can see that SER-FIQ outperforms all other QEs, although SDD-FIQA performs slightly better for the lower FPR. Interestingly, FaceQNet performs worst on this evaluation. PIQUE performs strongly, due to its robustness at accurately predicting perceptual quality given perturbations.

To further explore the performance of the QEs across perturbation types and subjects, we explore the cosine similarity and quality scores for all perturbed images. Figure 7 shows two sample scatterplots, one for SER-FIQ with randomly imposed color borders, and one for FaceQNet with JPEG perturbation. Each point corresponds to one perturbed image, and we highlight the images of three subjects with unique color markers. Figure 7-(a) shows that SER-FIQ is highly responsive to color changes in the borders of the images. It predicts a strong quality difference even though the face matcher creates little deviation in the cosine similarity. This is known as False Differentiation (FD) [4], where a QE predicts different quality even when the actual performance is unchanged. Further, for FaceQNet the quality score with JPEG varies little despite significant changes in the cosine similarity. This is known as a False Tie (FT), where a QE predicts similar quality even when the actual quality varies significantly. These two situations can be problematic when an application uses a QE score to take action, but the score does not reflect actual recognition performance. Note that these targeted tests identify a weakness even in a QE that performed well using the EvR evaluation protocol.



(a) SER-FIQ with color border (b) FaceQNet with JPEG

Figure 7: Quality score vs. cosine similarity

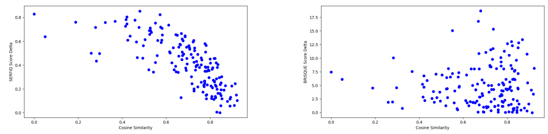
4.2. Robustness To Face Alignment Methods

To achieve face matching in a compete system, images are first processed a face detector and detected faces aligned before being processed by a matching step. Here, we consider the question: can a QE characterize the impact of different face alignment methods on face matching performance? We align 175 randomly selected images from the LFW [13] dataset using two different alignment strategies. One uses the cosine rule and aligns the face so the eye landmarks are horizontal, and the other uses a similarity transformation. We ensure that only the alignment strategy varies; in both cases, we use MTCNN face detector to detect faces and landmarks and use ArcFace with a MobileFaceNet backbone as our face matcher. The goal is to examine if the QE scores reflect face matcher performance.

Figure 8 shows results for the two alignment strategies on the randomly selected images from the LFW dataset [13], for both SER-FIQ and BRISQUE. The y-axis is the difference between the QE scores of the two alignments, and the x-axis is the cosine similarity between the two. BRISQUE does a poor job predicting the impact of alignment, which is not surprising as it is not designed for face matching. In contrast, SER-FIQ does reasonably well, with scores that increase as the similarity decreases. However, if one were to fit a nonlinear monotonic curve to this data, there would still be significant fitting error, indicating that this QE could be more effective in this application.

5. Concluding Discussion

In this paper, we reviewed existing evaluation protocols, such as EvR and BMW, for determining the effectiveness



(a) SER-FIQ

(b) BRISQUE

Figure 8: Comparison between two face alignment strategies: change in QE score vs. cosine similarity

of face QEs. We demonstrated an upper bound (ideal case) for each evaluation protocol, and delineated requirements for achieving a rigorous comparison. When proven effective, a QE can be used for real world applications such as prioritizing available resources to more reliable images, or increasing human involvement on less reliable images.

However, showing good results in traditional evaluation protocols (e.g., EvR) does not guarantee effectiveness across meaningful application scenarios. Our proposed targeted tests expose weaknesses even in reasonably effective QEs, and thus are an important component of evaluating QEs. Further, this paper has only scratched the surface of the range of possible targeted testing. There are many experiments can be done to evaluate the robustness of a QE in different scenarios.

A good quality estimator should handle multiple tasks; in particular, face matching requires detection, alignment, and recognition. Most QEs are only designed for and evaluated on one of these tasks, and the existing face QEs assume an already detected and aligned face. Further work is required to develop a QE that is effective across all these tasks.

References

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

[2] H. Chen, E. J. Delp, and A. R. Reibman. Estimating image quality for person re-identification. In *IEEE International Workshop on Multimedia Signal Processing*, 2021.

[3] 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*, pages 428–438. Springer, 2018.

[4] F. M. Ciaramello and A. R. Reibman. Supplemental subjective testing to evaluate the performance of image and video quality estimators. In *Human Vision and Electronic Imaging*, Jan. 2011.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Deepglint: Face feature test/trillion pairs. http:// trillionpairs.deepglint.com/overview. Accessed: 2022-05-01.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[7] A. Dutta, R. Veldhuis, and L. Spreeuwers. A Bayesian model for predicting face recognition performance using image quality. *IEEE International Joint Conference on Biometrics*, 2014.

[8] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):531–543, April 2007.

[9] 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*, pages 87–102. Springer, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] S. S. Hemami and A. R. Reibman. No-reference image and video quality estimation: Applications and human-motivated design. *Signal Processing: Image Communication*, Aug. 2010.

[12] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. FaceQnet: Quality assessment for face recognition based on deep learning. *Proceedings of the International Conference on Biometrics*, 2019.

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[14] H. Kim, S. H. Lee, and M. R. Yong. Face image assessment learned with objective and relative face image qualities for improved face recognition. *IEEE International Conference on Image Processing*, pages 4027–4031, 2015.

[15] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. IARPA Janus benchmark - C: Face dataset and protocol. In *International Conference on Biometrics (ICB)*, pages 158–165, 2018.

[16] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[17] A. Mittal, R. Soundararajan, and A. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 03 2013.

[18] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[19] P. Terhörst, J. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[20] Venkatanath N, Praneeth D, M. C. Bh, S. S. Channappayya, and S. S. Medasani. Blind image quality evaluation using perception based features. *Twenty First National Conference on Communications (NCC)*, 02 2015.

[21] J. Xiang and G. Zhu. Joint face detection and facial expression recognition with MTCNN. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427, 2017.