electronics

MDPI

Article

Face Recognition Bias Assessment through Quality Estimation Models

Luis Lopez Paya [1], Pedro Cordoba [2], Angela Sanchez Perez [2], Javier Barrachina [2], Manuel Benavent-Lledo [1], David Mulero-Pérez [1] and Jose Garcia-Rodriguez [1,*]

[1] Department of Computer Technology, University of Alicante, 03080 Alicante, Spain; llopez@dtic.ua.es (L.L.P.); mbenavent@dtic.ua.es (M.B.-L.); dmulero@dtic.ua.es (D.M.-P.)
[2] Facephi Research Lab, 03001 Alicante, Spain; pcordoba@facephi.com (P.C.); asanchezperez@facephi.com (A.S.P.); jbarrachina@facephi.com (J.B.)
* Correspondence: jgarcia@dtic.ua.es

Abstract: Recent advances in facial recognition technology have achieved outstanding performance, but unconstrained face recognition remains an ongoing issue. Facial-image-quality-evaluation algorithms evaluate the quality of the input samples, providing crucial information about the accuracy of recognition decisions. By doing so, this can lead to improved results in challenging scenarios. In recent years, significant progress has been made in assessing the quality of facial images. The computation of quality scores has become highly precise and closely correlated with the model results. In this paper, we reviewed and analyzed the existing biases of cutting-edge quality-estimation techniques for face recognition. Our experimentation focused on the quality estimators developed by MagFace, FaceQNet, and SER-FIQ and were evaluated on the CelebA reference dataset. A study of bias in the face-recognition model was conducted by analyzing the quality scores presented in each article. This allowed for an examination of existing biases within both the quality estimators and the face-recognition models.

Keywords: quality; bias; face recognition; deep learning

check for updates

Citation: Lopez Paya, L.; Cordoba, P.; Sanchez Perez, A.; Barrachina, J.; Benavent-Lledo, M.; Mulero-Perez, D.; Garcia-Rodriguez, J. Face Recognition Bias Assessment through Quality Estimation Models. *Electronics* **2023**, *12*, 4649. https://doi.org/10.3390/electronics12224649

Academic Editor: Jenhui Chen

Received: 19 October 2023
Revised: 10 November 2023
Accepted: 12 November 2023
Published: 15 November 2023

1. Introduction

Estimating the quality of a biometric image for facial recognition is a critical aspect in facial-recognition systems, as image quality can have a significant impact on system performance. The main challenges in estimating the quality of a biometric image for face recognition include:

- Variability of the capture conditions: Biometric facial images can be captured under a variety of conditions, such as different lighting, viewing angles, resolutions, and backgrounds. Image quality can be adversely affected by these variations.
- Noise and distortion: Biometric images can contain noise, distortion, and artifacts such as blur, background noise, reflections, and occlusions. These factors may impede the identification of accurate facial features.
- Capture equipment quality: The quality of the camera or sensor used to capture the image can vary widely. Low-quality images captured with inferior equipment may encounter issues regarding focus and resolution, affecting image quality adversely.
- Pose and facial expression: Image quality can vary based on pose (the orientation of the face) and facial expression. Extreme poses or expressions can hinder the identification and comparison of facial features.
- Facial coherence: In some cases, the image may not contain all the facial features necessary for accurate identification, such as missing parts of the face due to occlusion (e.g., glasses or a thick beard).
- Inadequate image quality: Low-quality images may lack sufficient detail for accurate identification, resulting in false negatives or false positives.

To address these challenges, specific techniques and algorithms have been developed to assess the quality of facial biometric images. These techniques typically involve extracting image features and evaluating the presence of problems such as noise, blur, and other artifacts. The estimated quality is then used to make decisions, such as selecting the best images for comparison or requesting a new capture if the quality is insufficient. Improving the quality of biometric images is critical for improving the accuracy and reliability of facial-recognition systems. The correlation is such that the quality scores can be used to analyze certain behaviors of the FR model, as was already proposed in [1].

The gap against which no biometric quality estimator is completely secure refers to the fact that, even when advanced techniques and algorithms are used for quality assessment, they are prone to errors. Multiple reasons contribute to the existence of this gap:

- Complexity of images: Biometric images, such as facial images, are inherently complex and can contain a wide variety of features and factors that affect their quality. Some of these factors may be subtle or difficult to detect, making it challenging to create perfectly accurate quality estimators.
- Variability in the capture conditions: As previously discussed, the variations in the capture conditions can lead to differing interpretations of image quality. Quality estimators may not be able to account for the numerous possible situations in which an image is taken, leading to varying judgments of image quality.
- Unpredictable subjects and scenarios: Subjects may have unique facial characteristics and changing expressions, which can affect image quality. In addition, unpredictable events in the environment, such as sudden changes in lighting or the presence of obstacles, can affect image quality.
- Technological limitations: Quality estimators are dependent on the technology and algorithms used. If the algorithms are not sufficiently robust or if the technology is outdated, they may be unable to detect all conditions that affect image quality.
- Subjectivity in perceived quality: Image quality is subject to personal perception and can vary from one individual to another. What one person considers as high-quality may not necessarily hold true for someone else.

In summary, the inability to develop a completely secure biometric quality estimator is due to the inherent complexity of biometric images and the variability of the capture conditions and subjects. Although technology and algorithms have advanced, the chances of overlooking quality issues or making mistakes in biometric image quality assessment always exist. Therefore, it is crucial for biometric systems to consider this limitation and incorporate supplemental security and verification measures to mitigate the possibility of the identification of errors.

For this reason, this paper conducted a study on the false rejection rates of the current state-of-the-art biometric quality estimators and the biases present in FIQA models, as well as their correlation with the respective facial recognition models. We analyzed and compared the distributions of the quality scores for different populations, taking into account factors such as hair color, accessories, or facial features that may obscure parts of the face, as well as age, among different FIQA estimators.

Finally, we evaluated the potential impact of these biases on specific demographics and demonstrate that incorporating a biased quality estimator or facial recognition model diminishes the accuracy of facial recognition and compromises its effectiveness and security, particularly in the digital age, where impersonation is becoming more prevalent. Although this is an ever-evolving field, there is significant room for improvement, as a biased FIQA model can have serious consequences for the system, given that these models do not perform uniformly across diverse populations.

The study of biases in the SER-FIQ, MagFace, and FaceQNet quality estimation models is of paramount importance in the field of facial recognition, as these models play a crucial role in the accuracy and reliability of biometric applications. Biases can introduce significant inaccuracies and inequalities in the assessment of biometric image quality, which, in turn, can affect the fairness and objectivity of facial recognition. Identifying and quantifying biases, whether related to gender, age, hair color, or other factors, is essential for understanding and addressing potential sources of error in these models. By identifying and mitigating biases, we can make progress toward more-equitable and -accurate facial-recognition systems that are fair and useful to a wide range of users and demographic groups, thereby contributing to a more-ethical and -responsible use of this technology.

Addressing the gaps in biometric image quality estimation is an ongoing challenge in the field of biometrics and facial recognition. While it may be difficult to completely eliminate these gaps, several strategies and approaches can be implemented to improve the accuracy and robustness of the quality estimators.

The remainder of this paper is organized as follows. Section 2 explains the most-relevant FIQA approaches. In Section 3, the experimental setup is described. Section 4 reports the results. Finally, in Section 5, the conclusions of this study are drawn.

2. Related Works

In this section, we aim to contextualize the most-recent advancements in Facial Image Quality Assessment (FIQA), focusing on three categories within existing FIQA approaches [2]. We categorized the most-recent techniques into articles that introduce innovative methods for calculating quality scores derived from Facial Recognition (FR) model predictions, followed by the use of a regression model to infer these quality scores from raw images. We also discuss articles that propose adjustments to the training of FR models to incorporate information regarding image usefulness in the FR prediction, as well as articles introducing modified inference methods.

2.1. Pseudo-Quality Label

The primary objective of this first group is to generate automatic pseudo-quality labels for training a regression model, enabling the prediction of the usefulness of an image for a specific facial-recognition model. Variations among these methods lie in how they extract pseudo-quality labels from the FR model.

FaceQNet [3] and FaceQNet2 [4] aim to develop FR-agnostic FIQA models that score based on how close the image is to being ICAO compliant for the FR models. In order to achieve this, they follow the scheme shown in Figure 1. Firstly, they select the best image for every user based on the ICAO standards. Once they have a reference image for each user, several FR models are used to extract the embeddings for every image. The pseudo-quality labels are calculated as the Euclidean distance between the embedding of the reference image to the respective images of the same user. These models also leverage transfer learning, wherein the FR model weights are frozen and a trainable dense layer is appended to estimate the quality score [5].

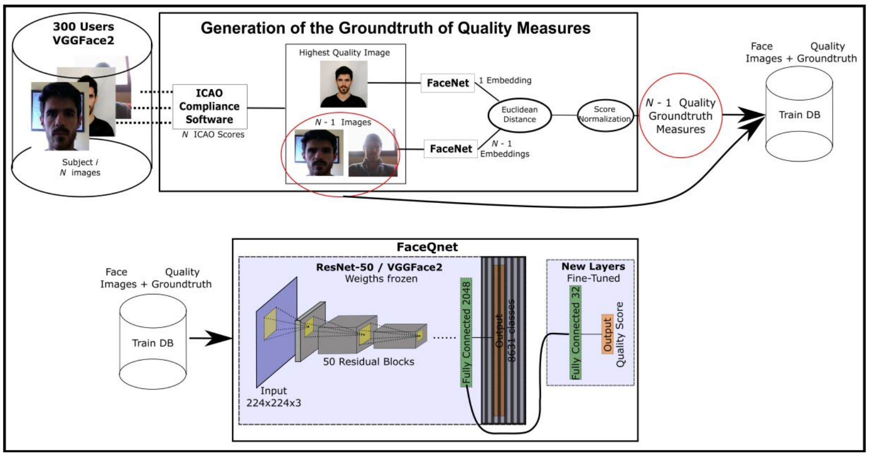

Figure 1. Score calculation techniques for the use of a regression model to replicate these scores. Pseudo-quality label generation proposed by FaceQNet [3].

PCNET [6] introduces a quality system that estimates image quality by measuring its similarity to another image of the same individual, based on pairwise image comparisons. When presented with a pair of images of sufficient quality for face recognition, where the FR model demonstrates reliability, the image pair is assigned a high-quality score. Works related to PCNET include [7–9].

SDD-FIQA [10] proposes an alternative approach, suggesting that image usefulness is determined not only by its proximity to embeddings of images from the same user, but also by its distance from embeddings of images from different users. The scoring mechanism involves computing scores using the Wasserstein distance between the distributions of similarities among images from the same user and different users. A larger distance between these two distributions results in a higher quality score [11]. Finally, a face-recognition model is trained using the obtained labels, similar to the approach in FaceQNet.

FaceQAN [12] leverages adversarial noise exploration for FIQA. Unlike most methods that require quality label generation, this unsupervised approach is based on adversarial examples and relies on the analysis of adversarial noise, which can be computed with any FR model. By comparing the embeddings of adversarial examples and the original input sample, FaceQAN is able to calculate a quality score that is an excellent predictor of the sample's utility for face recognition. FaceQAN can also operate with any face-recognition model, although the proposed model is preferred for improved quality estimation compared to other methods in the literature [13–17].

FaceQgen [18] is a no-reference quality assessment approach for face images based on a Generative Adversarial Network (GAN). No pre-labeled quality measures are needed for training. Instead, the quality assessment is based on measuring the similarity between the original and restored images, as lower-quality images undergo more substantial changes during the restoration process. Despite the lower accuracy compared to other methods, it proves the potential of semi-supervised learning approaches for FIQA.

CR-FIQA [19] presents a novel approach for assessing the quality of face images by predicting their relative classifiability. This classifiability is determined by analyzing the distribution of feature representations in angular space in relation to the class center and the nearest negative class center. The research demonstrates a robust correlation between the quality of face images and their relative classifiability through empirical analysis, given that this characteristic is exclusively observable in the training dataset; the authors advocate acquiring this knowledge by scrutinizing internal network observations throughout the training process, thus enabling one to assess the quality of unseen samples more effectively.

2.2. Face Recognition Training Modification

This approach uses face-recognition models to generate a set of embeddings that provide an assessment of the image quality. MagFace's method [20] adapts the training of a face-recognition model to extract embeddings that measure image utility. Quality is determined by the module of the vector of the extracted embedding, as shown in Figure 2. The bigger the modulus, the higher the quality of the image is. This technique presents a loss function established on the ArcFace loss [21–24]. The loss function permits a greater embedding separation for higher magnitudes, ultimately positioning high-quality images closer to the center of the class. This approach learns more from superior-quality images. The method proposed by MagFace improves the one proposed by ArcFace in terms of comparative results on the same datasets.

Another method used is the one proposed by Probabilistic Face Embeddings (PFEs) [25], which defines the image quality as the harmonic mean over the variance vector. This method learns to predict the uncertainty of a given image by estimating the mean and variance of the calculated embeddings. The mean vector represents the embedding of the selected image, while the variance vector signifies the image's uncertainty in the feature space.

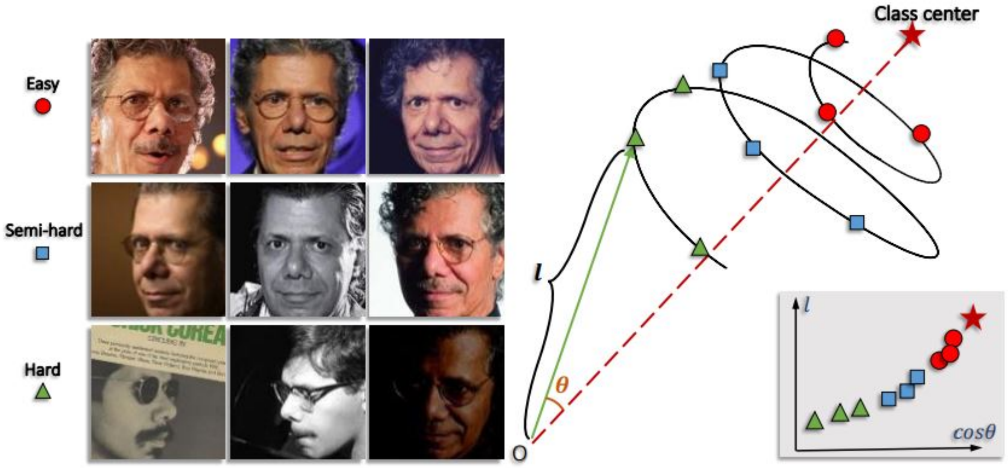

Figure 2. Techniques that modify the training of a model to obtain image utility information through embeddings. Definition and attainment of quality proposed by MagFace [20].

More recently, the authors in [26] provided a deep insight into the IQA and FIQA algorithms and how they affect FR. While face-specific metrics rely on features of aligned face images, general image quality metrics can be used on the global image and relate to human perceptions. This work analyzed 25 different quality metrics on various datasets, and the findings demonstrated a distinct connection between the acquired image metrics and their utility for face images, even in cases where they were not explicitly designed for face utility assessment. Conversely, individually crafted features exhibit a lack of overall stability and demonstrate markedly inferior performance compared to general face-specific quality metrics.

2.3. *Face Recognition Prediction Modification*

Finally, the methods in this category distinguish themselves by obviating the need to train a face-recognition model. Instead, these approaches focus on altering inference methods to obtain image quality. One notable approach within this category is SER-FIQ [27] represented in Figure 3, which harnesses dropout properties to randomly deactivate and activate neurons in the network layers to prevent overfitting. A series of embeddings, 100 proposed by the authors, are produced for a single input image using in each embedding a different dropout. Thanks to the sigmoid function of the normalized negative sum of the distance between all pairs of the produced embeddings, the image quality value is obtained. Similar works in this line of research can be found in [28–30].

In summary, techniques for biometric image quality estimation, such as SER-FIQ, MagFace, and FaceQNet, represent valuable resources in the field of facial recognition. However, none of these methods can fully address the challenges of biometric image quality estimation comprehensively. SER-FIQ focuses on evaluating image quality by predicting the efficacy of facial representation; MagFace employs deep learning to assess both image and facial quality; FaceQNet utilizes a convolutional neural network to estimate image quality in terms of authenticity and manipulation. Despite their distinct approaches and capabilities, no single technique can account for the complexities and variations inherent in biometric images, including diverse capture conditions, the presence of artifacts, and subtle quality issues. The persistent gap underscores the intricate nature of biometric images and the continuous evolution of image manipulation techniques. It highlights the significance of multifaceted approaches and ongoing improvements in biometric image quality to achieve more-accurate and -secure facial recognition.

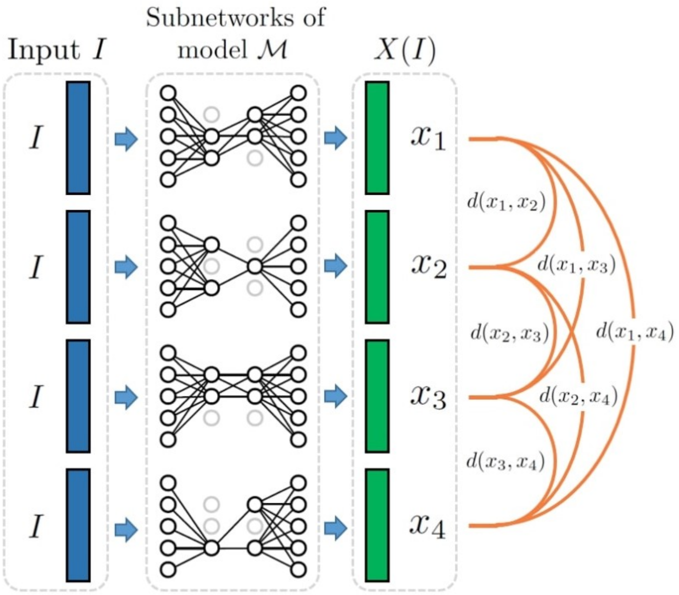

Figure 3. Techniques that propose a modified inference method. Quality attainment proposed by SER-FIQ [27].

3. Experimental Setup

This section describes the details of the experimental setup for the study that was carried out.

3.1. Dataset

To assess the quality of our estimation models, we utilized the CelebA dataset [31]. Comprising 200 k images from 10 k individuals, all of which were captured in unconstrained conditions, we selected this dataset due to its extensive and diverse attribute labels. In our study of bias, we specifically focused on information pertaining to hair type, face occlusion, age, and sex. We ensured that all technical terminology was explained and adopted a logical, objective tone throughout. We also adhered to conventional academic structure and formatting standards, including consistent citation usage and footnoting style. From this large dataset, we extracted a more-manageable and -controlled subset. We selected 10 images from 1000 different identities, resulting in a total of 10,000 images. Consequently, we compiled a diverse and extensive dataset suitable for conducting our research.

3.2. Face-Recognition Models

To obtain embeddings for all images in the dataset, we utilized two models, ArcFace https://github.com/deepinsight/insightface (accessed on 11 November 2023) [21] and MagFace https://github.com/IrvingMeng/MagFace (accessed on 11 November 2023) [20]. Both models were trained with the MS1M dataset [32]. Using the error versus reject curves [33], we assessed the effectiveness of the FIQA models with these two approaches. Identity verification was carried out by measuring the similarity between two embeddings using the cosine similarity.

3.3. Face Image Quality Assessment Models

For this work, we selected three state-of-the-art models of face image quality assessment: FaceQNet, MagFace, and SER-FIQ. We selected one model for each of the three categories we introduced in Section 2: FaceQNet for the pseudo-quality label category, MagFace for the face recognition training modification category, and SER-FIQ for the face recognition inference modification category.

In the case of FaceQNet https://github.com/uam-biometrics/FaceQNet (accessed on 11 November 2023) and MagFace https://github.com/IrvingMeng/MagFace (accessed on

11 November 2023), we used publicly available trained models from their respective official GitHub pages. In the case of SER-FIQ, we chose to use the ArcFace model to implement the inference modification to calculate the quality scores for this approach.

3.4. Experimentation

The main goal of this work was to analyze the bias in the FIQA models and, therefore, demonstrate if the face-recognition models they were trained for are biased. As was shown in [1], the quality scores given by the FIQA models are highly correlated with the reliability of the face-recognition model they were trained to assess. Due to this fact, the bias study of the FIQA models will be highly correlated with the bias of the face-recognition models. Since the quality assessment deeply influences the results of the bias study, we first analyzed the performance of the quality-assessment models. The second step was to carry out a bias study that consisted of the analysis of the different distributions of the quality scores given by these models for different population groups. The criteria used to form these population groups were hair color, type of face occlusion, age, and sex. With this study, it will be possible to observe whether these models assign lower quality scores to any of these groups, which could possibly cause issues for the users belonging to that group.

4. Results

4.1. Face Quality Assessment Performance

The study of the FNMR was carried out on a part of the CelebA dataset. To assess the quality of a face recognition system, it is necessary to evaluate the performance of the model according to the FNMR (Figure 4). The evolution of the FNMR will show the sensitivity and the degree of security that a face-recognition model possesses. The representation of the FNMR curves is affected by the FR model and by the quality score that each model estimates for each image. In this section, two face-recognition models, Insightface and MagFace, and three quality estimation models, FaceQNet, SER-FIQ, and MagFace, were evaluated. The quality models were evaluated against the Insightface model to see how the models perform in an unbiased FR model, and the MagFace model was evaluated against the MagFace quality scores to see how the use of the FR model affects its own quality-estimation model. Biometric systems perform best when tested on datasets on which they have been trained. This is not the case in real-life applications. For this reason, we evaluated the performance of the quality estimation and FR models on a new dataset for all models.

To show the behavior of the face-recognition models based on the quality estimation techniques, the evolution of the FNMR curves for the values of 0.5%, 1%, 2%, 5%, 15%, and 20% is shown.

For an FNMR initially set at 0.5%, as shown in the figure, the system had a very low similarity threshold. This makes the system more vulnerable and less secure, as it is more permissive to noise and poor-quality images. This factor worsens the FNMR calculations as the system contains a high false rejection rate and, with it, the utility of the system. As can be seen in the figure, the difference between the quality models was small. All models had the same behavior up to the 50% rejection point of the images. At this point, the images that formed the FNMR curve were the ones with the best quality and had a significant impact on the metric. In the case of FaceQNet, the metric improved, but it was still the worst curve compared to the other models. In the case of SER-FIQ, with a rejection of around 90% of the images, it showed a significant improvement in its metric. Therefore, SER-FIQ performed better when the image quality was high. As for MagFace, we obtained the curve of the scores extracted with respect to the Insightface model (over the model) and the curve with respect to MagFace's own FR model. It can be seen how both models performed almost on a par up to the point where the images increased in quality according to the MagFace model. With the increase in quality, the MagFace face-recognition model performed better, obtaining a metric with up to 0.1% over the curve of the FNMR on top model.

With the increase of the defined FNMR threshold set at 1.0%, 2.0%, and 5%, the evolution of the FNMR curves was affected in the same way as with a similarity threshold set for an FNMR of 0.5%. As can be seen in the figure, as the FNMR threshold increased, the system became more-secure and less-tolerant. From these thresholds, we will see how the quality estimation models actually work.

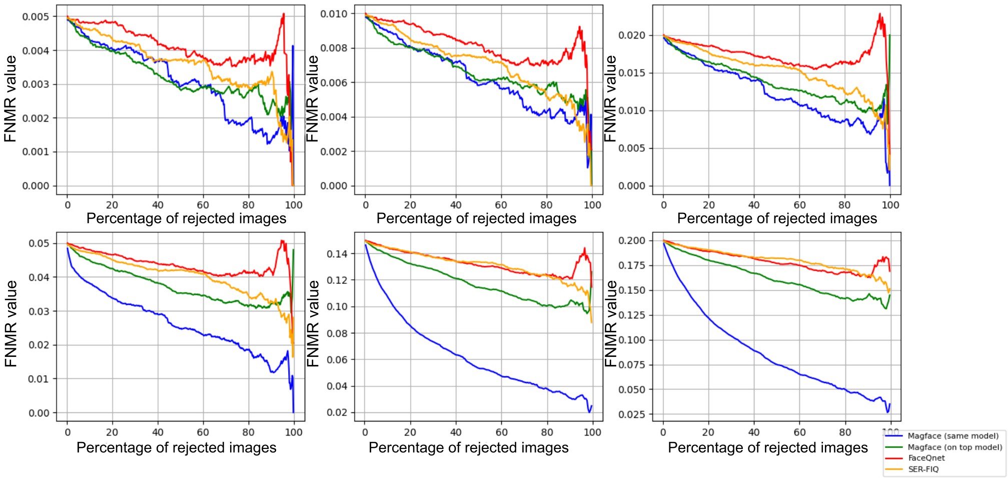

Figure 4. FNMR curves.

For a similarity threshold corresponding to an FNMR of 15%, it is shown how the FNMR evolved for high similarity thresholds, a system that was less permissive with the inputs. As can be seen, the Insightface FR model performed better with the scores extracted by the MagFace model.

On the other hand, the metric of the MagFace model (same model) was obtained, which improved with increasing system security. It can be seen how the curve decreased in the FNMR value considerably with respect to the percentage of rejected images in the set. The MagFace model obtained a better metric with respect to the rest of the models.

In addition, the analysis was performed for an FNMR value set at 20%, as shown in the figure. For this higher similarity threshold, the system behaved in the same way as for the 15% threshold. The Insightface FR model showed an improvement in the FNMR curves. This was because the model was not optimal in the similarity extraction performed to obtain the similarity matrix. Even so, the MagFace quality estimation model was still the best-performing model for this Insightface model. On the other hand, we obtained the curve of the MagFace model (same model), which still presented the best metric of all models due to the optimal similarity extraction by the FR model and the use of its own quality scores provided by the model.

4.2. Face Quality Assessment Bias

Biometric image quality estimation plays a crucial role in facial-recognition systems, directly influencing their accuracy and reliability. However, as we advance in the application of techniques like SER-FIQ, MagFace, and FaceQNet, it is essential to recognize the growing concern about biases in these models. Biases in quality estimation can arise from various sources, such as imbalances in the datasets used to train these models, thereby reflecting biases present in society and the available images. This study aimed to analyze and understand the presence of biases in SER-FIQ, MagFace, and FaceQNet, examining how these biases can affect the assessment of biometric image quality and, ultimately, equity and accuracy in facial-recognition applications.

4.2.1. Bias Assessment According to the Person's Hair Color

The study of biases in the hair color of individuals within the SER-FIQ, MagFace, and FaceQNet quality estimation models revealed an important aspect of equity and accuracy in biometric image assessment. It was observed that these models can exhibit certain levels of bias with respect to hair color, raising concerns about their ability to provide unbiased quality estimations. Bias in hair color can arise due to imbalances in the training data, where certain demographic groups may be underrepresented or misrepresented. This study aimed to quantify and assess the extent of such bias in SER-FIQ, MagFace, and FaceQNet, with the goal of raising awareness about these issues and working towards improving equity in biometric image quality assessment in the field of facial recognition.

In order to study the biases through hair color, we obtained graphs with the quality scores for each of the models. In the FaceQNet model, the bias due to hair color was studied as shown in Figure 5. As can be seen, the FaceQNet model is biased. When the model received an image of people with blond or brown hair color, it estimated a lower image quality than people with black and gray hair. As for the estimation between people with gray hair and black hair, the model estimated similar qualities, but performed better for people with gray hair.

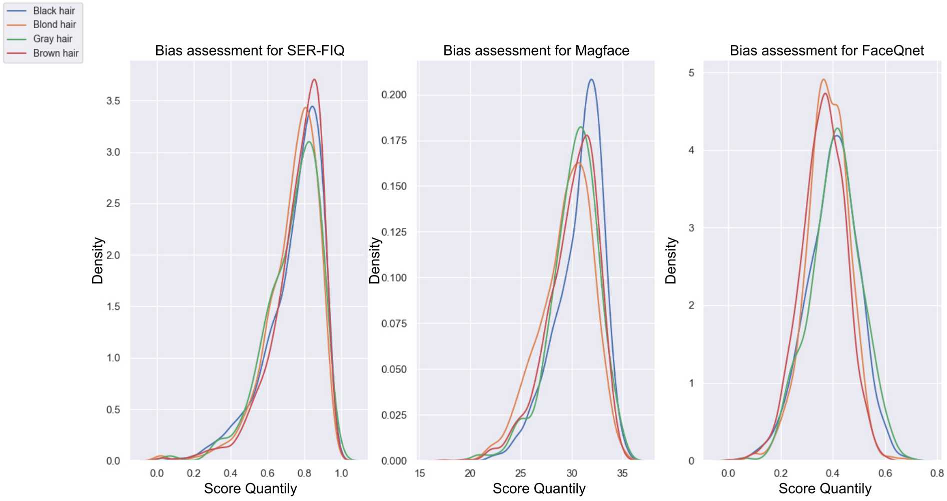

Figure 5. Quality scores by hair color.

To evaluate the bias of the MagFace model, the same criterion as in the previous models was used. Figure 5 shows the distribution of the quality scores of the MagFace model on the user's hair color. This model obtained a higher quality score on images of people with black hair. In the case of gray and brown hair, the model behaved similarly, obtaining worse quality scores than for users with black hair. In this case, the MagFace model performed worse for people with blond hair. For this type of person, the model estimated lower quality scores than for other hair colors. The model made a clear distinction between the different hair colors, and there was an important bias in this aspect. As for the SER-FIQ model, Figure 5 shows how the model presented a similar behavior to all the users' hair types, but made a very clear distinction with brown-haired users. For this hair type, the SER-FIQ model estimated higher quality scores than for the other hair colors. On the other hand, there were gray-haired users for whom black-haired users will have higher quality and blond-haired users will have lower quality.

4.2.2. Bias Assessment According to Facial Occlusions

Figure 6 shows the different quality estimation curves according to the facial occlusions contained in the image. In this case, it was clear that the FaceQNet model estimated low qualities for people with glasses. This can be a problem since, by estimating a low quality for this type of occlusion, when a system is quite safe, it will run the risk of rejecting all images of people with glasses, something that would be wrong since wearing glasses should not be estimated as a worse quality. This type of problem usually occurs due to reflections on the glasses when taking the picture. On the other hand, the model worked worse for people without beards than for people with beards, which indicates that the model has probably been trained with more images of users with beards than without beards. Finally, the model performed similarly for users with beards and with users with mustaches, but estimated a better quality for people with beards. The model contained an important bias with which people with glasses will be greatly affected when the system validates their image as good in terms of quality for subsequent facial recognition.

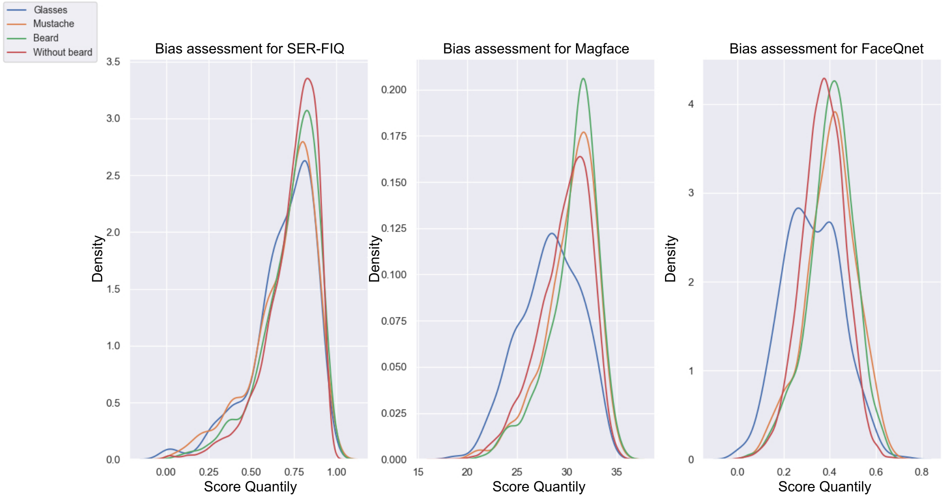

Figure 6. Quality scores by facial occlusions.

As in the other models, MagFace made a distinction when it came to quality on people with glasses, giving a lower quality to such images. In this case, the difference is very noticeable when viewing Figure 6. As for people with or without beards and people with mustaches, the model was also biased, although not to as great of an extent as it was with people wearing glasses. This difference in quality can cause a major problem when it is used for facial recognition.

Figure 6 depicts the behavior of the SER-FIQ model with respect to a series of occlusions. For the occlusions, the SER-FIQ model showed a significant bias between the different occlusions. The occlusion that estimated the worst quality score was for glasses wearers. It was not as significant a bias as in FaceQNet, but it did affect this factor of glasses in terms of quality. The SER-FIQ model estimated higher quality for users without beards; the fact that there was no occlusion in the facial region made this model perform much better. Finally, the model performed similarly for users with mustaches and users with beards, estimating higher quality for users with beards.

4.2.3. Bias Assessment According to the Age of the Person

Research into biases related to age in the SER-FIQ, MagFace, and FaceQNet quality estimation models highlights a fundamental concern in the field of facial recognition. It has been identified that these models can exhibit certain levels of bias with respect to people's age, raising questions about their ability to provide objective and equitable assessments of biometric image quality. Age bias can arise due to the unequal representation of different age groups in the datasets used to train the models, which can lead to inaccuracies in quality estimation and ultimately result in less-reliable performance in facial-recognition applications. This study aimed to quantify and analyze the nature and extent of such bias in SER-FIQ, MagFace, and FaceQNet, with the goal of raising awareness about these issues and promoting improvements in equity and accuracy in quality-estimation models in the field of facial recognition. Another sign of bias in the FaceQNet model can also be seen in Figure 7. The model estimated lower image qualities for young people than for adults.

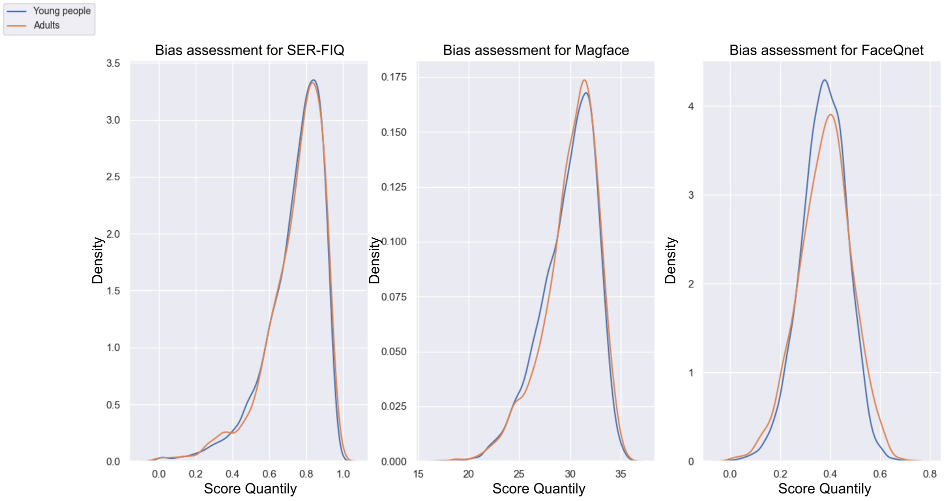

Figure 7. Depending on the age of the person.

As for the age of the users, according to Figure 7, there was a small bias in the model. For young people and adults, the model behaved similarly. However, for young people, the model estimated a higher quality.

In Figure 7, it can be seen how the SER-FIQ model was not biased in terms of the age of the person in the photo. As can be seen in the graph, the model estimated almost the same quality value for young people as for adults. It is not biased in this respect.

4.2.4. Bias Assessment According to the Sex of the Person

In Figure 8, we can see how the sex of the person affected the subsequent quality estimation in the FaceQNet model. As can be seen, there was a large bias between sexes. The model estimated higher image qualities for the male sex than for the female sex. In view of the results, this quality-estimation model will have low quality scores for female glasses wearers, an important bias for subsequent face recognition. We evaluated how the sex of the person affected the quality of the image defined by the model. Figure 8 shows the behavior of MagFace. This model had a significant bias with respect to the sex of the person. It can be seen how the model estimated high quality values when the sex was male. In contrast, if the sex was female, the model's quality score was lower.

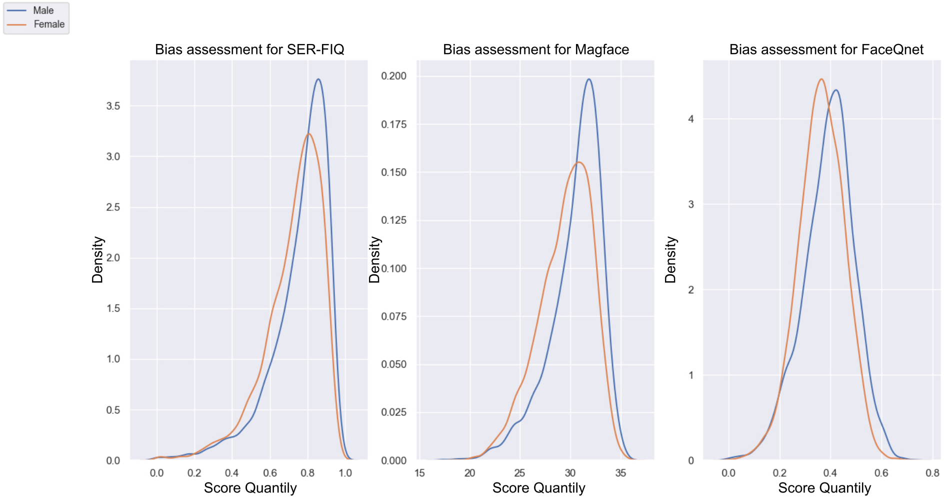

Figure 8. According to the sex of the person.

MagFace had important biases when estimating the quality of an image. One of them was the fact of being male or female. If one was of the female gender, the image was more likely to be rejected before performing face recognition. On the other hand, the model made a clear distinction between people with glasses and those without. The quality value of an image was greatly affected by the fact of wearing glasses. This bias hinders good face recognition.

Because MagFace's quality scores were extracted from the feature vectors that the model extracted from each image, looking at the biases present in MagFace's model, it can be seen that MagFace's face-recognition model was also biased for the same aspects. This occurred because MagFace's FR model is in charge of determining the final quality score for each image.

Like the FaceQNet model, SER-FIQ obtained better quality scores for images featuring men than for images featuring women. Significant bias can be seen in the bias evaluation curve.

The SER-FIQ model worked well on people of any age. On the other hand, it obtained a higher bias among people of different sexes and accentuated the loss of quality on people wearing glasses.

5. Conclusions

Studying biases in a biometric image quality estimation models is of paramount importance in today's context. This imperative arises from the escalating centrality of facial and biometric recognition systems across diverse domains, encompassing security, online authentication, and automated decision-making. Detecting and understanding biases in these models not only ensures a fair and equitable application of the technology, but also helps prevent unintended consequences, such as discrimination and privacy breaches. In-depth research on these biases not only reveals existing challenges, but also provides a solid foundation for improving the models and promoting more ethical and responsible practices in the field of biometrics and facial recognition. Ultimately, this kind of research is essential to ensure that biometric technology serves as a precise and fair tool for the collective benefit of society.

In conclusion, this investigation underscored the critical importance of addressing and understanding biases in biometric-image-quality-estimation models, such as SER-FIQ, MagFace, and FaceQNet. These biases have the potential to undermine equity and objectivity in facial-recognition applications, compromising the accuracy and reliability of the systems. Identifying gaps in these models, whether associated with gender, age, hair color, or other factors, is an essential step toward improving fairness and ethics in the field. While it may not be possible to completely eliminate these gaps, multifaceted approaches can be adopted, including diversifying training datasets, ongoing evaluation, education, and awareness, as well as technological improvements. By actively recognizing and addressing these issues, we can progress towards impartial and precise image-quality-estimation systems in the realm of facial recognition, thereby meeting higher ethical standards and benefiting a more-inclusive and -equitable society.

Tackling the disparities in biometric image quality estimation persists as a challenge in the domains of biometrics and facial recognition. Although the absolute eradication of these disparities may prove elusive, a repertoire of strategies and methodologies can be implemented to enhance the accuracy and robustness of quality estimators:

- Advanced algorithms: Continuously improving and developing algorithms that can detect a wide range of image quality issues, including noise, blur, and distortion, constitute a pivotal endeavor.
- Machine learning: Leveraging machine learning techniques to train models on diverse datasets can help improve the ability of quality estimators to adapt to various conditions and populations.

- Diverse training data: Ensuring that training data used for quality estimation include diverse populations, distinct lighting conditions, and a wide range of image variations can reduce biases.
- Bias detection and mitigation: Implementing techniques to detect and mitigate biases in quality-estimation models, especially those related to demographics, can enhance fairness and accuracy.
- Feedback loops: Establishing feedback loops within recognition systems to reevaluate and potentially reacquire low-quality images can help improve the overall quality of the biometric database.
- User-centric approaches: Integration of user input and preferences into quality-estimation processes can augment user experience and engender trust in biometric systems.
- Continuous research: Maintaining an active research effort in the field to stay updated with the latest advancements in image quality estimation is crucial.
- Ethical considerations: Ensuring that ethical considerations are part of the development process, including addressing potential biases and privacy concerns, is essential.
- Standards and regulations: Adhering to established standards and regulations for biometric data collection and quality assessment can help ensure consistency and fairness.
- Transparency and accountability: Promoting transparency in the development and deployment of biometric systems and holding developers accountable for their systems' performance and biases is important.

While the complete elimination of disparities in biometric image quality estimation may remain an elusive goal, amalgamating these strategies can serve to diminish errors, amplify system efficacy, and guarantee that biometric technology attains heightened reliability and equity for all end-users.

Author Contributions: Conceptualization, L.L.P. and P.C.; methodology, P.C.; software, L.L.P.; validation, L.L.P., P.C. and A.S.P.; formal analysis, L.L.P.; investigation, L.L.P.; resources, P.C.; data curation, M.B.-L. and D.M.-P.; writing—original draft preparation, L.L.P., P.C., M.B.-L. and D.M.-P.; writing—review and editing, L.L.P., P.C., M.B.-L., D.M.-P., J.G.-R., A.S.P. and J.B.; visualization, M.B.-L. and D.M.-P.; supervision, J.G.-R. and J.B.; project administration, P.C., J.G.-R. and J.B.; funding acquisition, J.G.-R. and J.B. All authors have read and agreed to the published version of the manuscript.

Funding: We would like to thank the "A way of making Europe" European Regional Development Fund (ERDF) and MCIN/AEI/10.13039/501100011033 for supporting this work under the TED2021-130890B (CHAN-TWIN) research Project funded by MCIN/AEI/10.13039554/501100011033 and the European Union NextGenerationEU/PRTR; HORIZON-MSCA-2021-SE-0 action number: 101086387, REMARKABLE, Rural Environmental Monitoring via ultra wide-ARea networKs And distriButed federated Learning; CIAICO/2022/132 Consolidated group project "AI4Health" funded by the Valencian government and International Center for Aging Research ICAR funded project "IASISTEM". This work was also supported by two Spanish national and regional grants for Ph.D. studies, FPU21/00414 and CIACIF/2021/430.

Data Availability Statement: The dataset analyzed during the current study is available at: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 11 November 2023), and the face-recognition models are available at: https://github.com/deepinsight/insightface (accessed on 11 November 2023), https://github.com/IrvingMeng/MagFace (accessed on 11 November 2023) and https://github.com/uam-biometrics/FaceQNet (accessed on 11 November 2023).

Acknowledgments: We would like to thanks Facephi for its support. The experiments were made possible by a generous hardware donation from NVIDIA.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Terhörst, P.; Kolf, J.N.; Damer, N.; Kirchbuchner, F.; Kuijper, A. Face Quality Estimation and Its Correlation to Demographic and Non-Demographic Bias in Face Recognition. In Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 28 September–1 October 2020; pp. 1–11. [CrossRef]
2. Méndez-Vázquez, H.; Chang, L.; Rizo-Rodríguez, D.; Morales-González, A. Evaluación de la calidad de las imágenes de rostros utilizadas para la identificación de las personas. *Comput. Sist.* **2012**, *16*, 147–165.
3. Hernandez-Ortega, J.; Galbally, J.; Fierrez, J.; Haraksim, R.; Beslay, L. FaceQNet: Quality Assessment for Face Recognition Based on Deep Learning. In Proceedings of the 2019 International Conference on Biometrics (ICB), Crete, Greece, 4–7 June 2019.
4. Hernandez-Ortega, J.; Galbally, J.; Fierrez, J.; Beslay, L. Biometric Quality: Review and Application to Face Recognition with FaceQNet. *arXiv* **2020**, arXiv:2006.03298.
5. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823. [CrossRef]
6. Liang, F.; Shen, C.; Yu, W.; Wu, F. Towards Optimal Power Control via Ensembling Deep Neural Networks. *arXiv* **2018**, arXiv:1807.10025. [CrossRef].
7. Wang, F.; Chen, L.; Li, C.; Huang, S.; Chen, Y.; Qian, C.; Loy, C.C. The Devil of Face Recognition is in the Noise. *arXiv* **2018**, arXiv:1807.11649. [CrossRef].
8. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. SphereFace: Deep Hypersphere Embedding for Face Recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
9. Best-Rowden, L.; Jain, A.K. Learning Face Image Quality From Human Assessments. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 3064–3077. [CrossRef]
10. Ou, F.Z.; Chen, X.; Zhang, R.; Huang, Y.; Li, S.; Li, J.; Li, Y.; Cao, L.; Wang, Y.G. SDD-FIQA: Unsupervised Face Image Quality Assessment with Similarity Distribution Distance. *arXiv* **2021**, arXiv:2103.05977v1.
11. Kharchevnikova, A.; Savchenko, A. Efficient video face recognition based on frame selection and quality assessment. *Peerj Comput. Sci.* **2021**, *7*, e391. [CrossRef] [PubMed]
12. Babnik, Ž.; Peer, P.; Štruc, V. FaceQAN: Face Image Quality Assessment Through Adversarial Noise Exploration. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 748–754. [CrossRef]
13. Grm, K.; Dobrišek, S.; Scheirer, W.J.; Štruc, V. Face hallucination using cascaded super-resolution and identity priors. *arXiv* **2018**, arXiv:1805.10938v2. [CrossRef].
14. Schlett, T.; Rathgeb, C.; Henniger, O.; Galbally, J.; Fierrez, J.; Busch, C. Face Image Quality Assessment: A Literature Survey. *ACM Comput. Surv.* **2022**, *54*. [CrossRef]
15. Boutros, F.; Damer, N.; Kirchbuchner, F.; Kuijper, A. ElasticFace: Elastic Margin Loss for Deep Face Recognition. *arXiv* **2021**, arXiv:2109.09416v4. [CrossRef].
16. Wang, F.; Cheng, J.; Liu, W.; Liu, H. Additive Margin Softmax for Face Verification. *IEEE Signal Process. Lett.* **2018**, *25*, 926–930. [CrossRef]
17. Khan, S.; Hayat, M.; Zamir, W.; Shen, J.; Shao, L. Striking the Right Balance with Uncertainty. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
18. Hernandez-Ortega, J.; Fierrez, J.; Serna, I.; Morales, A. FaceQgen: Semi-Supervised Deep Learning for Face Image Quality Assessment. In Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021; pp. 1–8. [CrossRef]
19. Boutros, F.; Fang, M.; Klemt, M.; Fu, B.; Damer, N. CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 5836–5845.
20. Meng, Q.; Zhao, S.; Huang, Z.; Zhou, F. MagFace: A Universal Representation for Face Recognition and Quality Assessment. *arXiv* **2021**, arXiv:2103.06627v4.
21. Deng, J.; Guo, J.; Yang, J.; Xue, N.; Cotsia, I.; Zafeiriou, S.P. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
22. Zhang, X.; Zhao, R.; Qiao, Y.; Wang, X.; Li, H. AdaCos: Adaptively Scaling Cosine Logits for Effectively Learning Deep Face Representations. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
23. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]
24. Wang, F.; Xiang, X.; Cheng, J.; Yuille, A.L. NormFace. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017. [CrossRef]
25. Chen, K.; Lv, Q.; Yi, T. Fast and Reliable Probabilistic Face Embeddings in the Wild. *arXiv* **2021**, arXiv:2102.04075. [CrossRef].

26. Fu, B.; Chen, C.; Henniger, O.; Damer, N. A Deep Insight Into Measuring Face Image Utility with General and Face-Specific Image Quality Metrics. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 905–914.
27. Terhörst, P.; Kolf, J.N.; Damer, N.; Kirchbuchner, F.; Kuijper, A. SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
28. Sellahewa, H.; Jassim, S. Image-Quality-Based Adaptive Face Recognition. *IEEE Trans. Instrum. Meas.* **2010**, *59*, 805–813. [CrossRef]
29. Chen, J.; Deng, Y.; Bai, G.; Su, G. Face Image Quality Assessment Based on Learning to Rank. *IEEE Signal Process. Lett.* **2015**, *22*, 90–94. [CrossRef]
30. Goel, T.; Murugan, R. Classifier for Face Recognition Based on Deep Convolutional-Optimized Kernel Extreme Learning Machine. *Comput. Electr. Eng.* **2020**, *85*, 106640. [CrossRef]
31. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
32. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 87–102.
33. Grother, P.; Tabassi, E. Performance of biometric quality measures. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 531–543. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.