Explainability of the Implications of Supervised and Unsupervised Face Image Quality Estimations Through Activation Map Variation Analyses in Face Recognition Models

Biying Fu[1], Naser Damer[1,2]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]Department of Computer Science, TU Darmstadt, Darmstadt, Germany
Email: biying.fu@igd.fraunhofer.de

Abstract

It is challenging to derive explainability for unsupervised or statistical-based face image quality assessment (FIQA) methods. In this work, we propose a novel set of explainability tools to derive reasoning for different FIQA decisions and their face recognition (FR) performance implications. We avoid limiting the deployment of our tools to certain FIQA methods by basing our analyses on the behavior of FR models when processing samples with different FIQA decisions. This leads to explainability tools that can be applied for any FIQA method with any CNN-based FR solution using activation mapping to exhibit the network's activation derived from the face embedding. To avoid the low discrimination between the general spatial activation mapping of low and high-quality images in FR models, we build our explainability tools in a higher derivative space by analyzing the variation of the FR activation maps of image sets with different quality decisions. We demonstrate our tools and analyze the findings on four FIQA methods, by presenting inter and intra-FIQA method analyses. Our proposed tools and the analyses based on them point out, among other conclusions, that high-quality images typically cause consistent low activation on the areas outside of the central face region, while low-quality images, despite general low activation, have high variations of activation in such areas. Our explainability tools also extend to analyzing single images where we show that low-quality images tend to have an FR model spatial activation that strongly differs from what is expected from a high-quality image where this difference also tends to appear more in areas outside of the central face region and does correspond to issues like extreme poses and facial occlusions. The implementation of the proposed tools is accessible here [1].

[1]https://github.com/fbiying87/Explainable_FIQA_WITH_AMVA

1. Introduction

Face recognition (FR) systems are becoming more widely used in our daily life, be it for security-relevant areas such as border control or for unlocking your personal devices such as smartphones. This spread relates highly to the performance improvements due to advances made in deep-learning methods for FR [6, 2, 3].

Low utility [18] face samples largely effects the performance of FR algorithms [1]. Therefore, to improve the performance of FR and overcome the aforementioned challenges, advances have been made to enhance the performance and dependability by choosing to use high-quality face images. Face image quality is also used to weight face embeddings and comparison scores when performing multi-frame (video) [21, 5, 20] or multi-spectrum [19] face verification. Facial image quality assessment (FIQA) methods have been developed to evaluate this metric for facial images. Recent advances in the development of FIQA methods with deep-learning-based approaches already show good results in improving the FR performance. Few works focus also on building FIQA methods based on interpretable reasoning in terms of uncertainty [26], embedding robustness [27], or embedding magnitude that correlate to the sample location with respect to its class [21]. However, little effort was put into the explainability of these methods.

To address this unscouted field, we develop a set of novel explainability tools. To start with, and to enable our tools to be deployed to both supervised and unsupervised FIQA approaches, we do not look into the responses of the FIQA approach itself, but rather the FR model behavior when processing images with various FIQ decisions. Given that the general spatial activation of FR models to low or high-quality images is rather similar, we take this analysis to the higher derivative level by looking at the activation map variation analyses. The newly proposed tools successfully target answering three questions related to the explainability

of FIQA decisions. These questions address the following issues: 1) What makes the face images of high quality (in comparison to low) in the view of different FIQA based on the FR model behavior, 2) what makes the decision of high or low quality different between different FIQAs based on the FR model behavior, and 3) how do the activation mappings caused by face images in the FR model deviate from what is expected from high-quality images and how does that reflect in their quality score.

The structure of our work is as follows: we first introduce the related work in terms of FIQA methods in Section 2. In section 3, we introduce our proposed explainability tools. The experimental setup is described in Section 4, while the analyses based on our proposed tools are followed in Section 5. We conclude our paper in Section 6 where we shortly summarized our main findings.

2. Related Works

The estimation of face images utility with FIQA methods helps FR systems to improve their performance by avoiding the processing of low-quality, and thus low-utility captures.

Most recent works of FIQA methods mainly focused on enhancing the performance of the proposed metric in terms of the FR accuracy when neglecting low-quality samples. An example of that is the probabilistic Face Embeddings (PFEs) proposed by Shi et al. [26] where they represented each face image as a Gaussian distribution in the latent space, where the variance of the Gaussian indicates the uncertainty in the feature space. This uncertainty is used as a measure of quality. Another example is the SDD-FIQA [23] where Ou et al. proposed a supervised method using generated quality pseudo-labels by calculating the Wasserstein Distance between the intra-class similarity distribution and inter-class similarity distributions of sample identities. With these peseudo-labels, the network trains a regression network for the quality prediction. Other FIQA metrics using the trustworthiness as in SER-FIQ [27] or MagFace [21] by learning a universal feature embedding which magnitude is a direct measure of quality, are further introduced in Section 4.

Recent works have tried to have a detailed look into the contribution of facial parts to the estimated face image quality (FIQ). Fu et al. [10] investigated the different face subregions (including eyes, mouth, and nose) and showed their relative importance towards face utility by comparing general image quality assessment (IQA) metrics on these areas. However, this work only looked at IQA metrics and thus does not provide many insights on the explainability of FIQA. Further works have looked into the spatial activation maps of supervised FIQA and IQA methods and analyzed how they are affected when facing different sample categories such as low and high-quality [9], or even masked faces [11]. However, these efforts were limited to the supervised FIQA methods and did not address the better performing unsupervised FIQA methods, as the nature of the unsupervised approaches does not allow for rational activation map analyses. Moreover, none of the previous FIQA methods tried to provide an explaination of a quality decision by looking into the spatial interpretation of the response of FR models to what is deemed as low or high-quality samples, rather than just analyzing the FIQA behavior. In this work, we propose a generalized methodology to enhance the explainability of the behavior of FR models on low and high-quality samples, as well as face image quality estimation decisions, by analyzing the FR activation mappings, rather than these of FIQA.

3. Methodology

In this section, we propose a novel set of tools to explain the face image quality and its effect on FR model behavior, independent of the underlying working principles of the FIQA methods itself, by focusing on the response of the FR models to samples labeled with different qualities. We leveraged the process of activation mapping of a visualization network to display the scaled activation weighted by the face embedding of the FR network. This mapping links the content of face embeddings with the pixels in the input face image. Based on the activation mapping, we draw statistic characteristics for images with different face qualities and thus enable analyses of the response of FR models to low and high face image qualities and the quality interpretation of single samples based on FR model responses.

To illustrate our proposed method with a concrete example, let us assume that we chose MagFace [21] as the underlying unsupervised FIQA method and used the ResNet-100 [13] model trained with ArcFace loss [6] as the face recognition model to extract the face embeddings. We further used ScoreCAM [28] as the approach for the activation mapping process. The activation mapping visualized the deepest convolution layer of the Res-Net100 and upsampled it to overlay to the input layer. The scaled version of the activation measures how the output changes to the face embedding. For each image, the activation mapping (AM) provided an output activation map with each pixel value noted as $a_{i,j}, i = 1 : 112, j = 1 : 112$ of the size 112x112. However, this concept can be extended to any FIQA method and CNN-based FR model. More details about the exact models used in this work and the reason for this selection are provided later in Section 4.

Using the selected FIQA metric, e.g., MagFace, we calculate the face image scores for a given face images database. The calculated face image scores are used as ground truth to determine the group of low and high-quality images. In our experiment, we chose the 10% of face images with the lowest and 10% of face images with the highest FIQ scores. For simplicity, these two groups are named

H and L individually.

For each of these two individual groups of H and L images, we introduce and define the mean activation mappings (MAM) and denote them as MAM_H and MAM_L. These maps have the same dimension as the input with 112x112 pixels. Each element in the MAM is noted as $\overline{a_{i,j}}$ and it is derived from equation (1) using the activation value of each single sample $a_{i,j}$ with the running index $i = 1 : 112$ and $j = 1 : 112$:

$$\overline{a_{i,j}} = \frac{1}{N}\sum_{k=1}^{N} a_{i,j}^{k}, \tag{1}$$

where N is the number of images within the H or L groups respectively. As we aim to measure the variability in the activation mapping, we introduce and define the activation mapping variation map (AM-V). We further denote them as AM-V_H and AM-V_L respectively, for high-quality or low-quality samples. These maps also have the dimension of the input with a shape of 112x112 and each element in the AM-V is the $\text{s}_{i,j}$ and is derived according to equation (2):

$$\text{s}_{i,j} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(a_{i,j}^{k} - \overline{a_{i,j}})^2}, \tag{2}$$

where N has the same meaning as in Equation (1) and $a_{i,j}$ are extracted from elements of the activation mapping.

In order to reduce the influence of outliers in the MAM, we further looked at the median activation mapping (MDAM). The notation of MDAM_H and MDAM_L are also noting the MDAM for low and high quality sample sets. The element of the MDAM is denoted as $\widetilde{a_{i,j}}$ and is derived as:

$$\widetilde{a_{i,j}} = Median(a_{i,j}^{k}), k = 1..N. \tag{3}$$

We further introduce the activation mapping Median variation (AM-MV), called AM-MV_H and AM-MV_L. The associated equation for each element of these maps $\widetilde{s_{i,j}}$ is given as:

$$\widetilde{s_{i,j}} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(a_{i,j}^{k} - \widetilde{a_{i,j}})^2}, \tag{4}$$

where both AM-MV maps have the same dimension of 112x112 pixels.

Both the defined AM-V and AM-MV maps present a visualization tool to look into the spatial areas where a relatively large variation in the activation of the FR occurs, with respect to a set of high or low-quality images. This will help identify the spatial regions responsible for the certain quality decision, despite the low consistency of these areas' location across different images.

As will be shown later, the differences between the MAM (or MDAM) of image sets of different qualities do not uncover a lot of explainability information. Therefore, to uncover the spatial related differences between these groups, we rather analyze the differences between the variations in the activation mapping (AM-V or AM-MV). We introduce these differences as the Differential activation mapping variation (D-AM-V) as in (5) and the Differential activation mapping Median variation (D-AM-MV), as in (6):

$$\text{D-AM-V} = |\text{AM-V}_H - \text{AM-V}_L|, \tag{5}$$

and

$$\text{D-AM-MV} = |\text{AM-MV}_H - \text{AM-MV}_L|, \tag{6}$$

where both equations (5) and (6) can be extended to look at the differences of variations of any sets of images, not only L and H, but also to sets of images determined to be H or L by different FIQA approaches.

The proposed visualization maps provide a useful tool to analyze the differences in FR model responses to sets of facial images belonging to different sets, here sets with different FIQ determined by any FIQA approaches, or sets of a certain FIQ label determined by different FIQA approaches.

So far we introduced methods to visualize and analyze the behavior of FR models between sets of images to enable a better understanding of FR response to images of different quality levels and thus understanding the used FIQA. To further analyze the quality decision of a single face image using its FR model response, we introduce the activation deviation from the MAM (AD-MAM). The elements of the AD-MAM is noted as d_i, j and is calculated of an image x with its activation mapping element $a_{i,j}$ and its absolute deviation from the mean activation mapping of the high quality sets $\overline{a_{i,j}^{H}}$,

$$d_{i,j} = |a_{i,j} - \overline{a_{i,j}^{H}}|, \tag{7}$$

this can be calculated for different sets of images that build the MAM, here we focus our analyses on the MAM_H, and thus we note our AD-MAM as AD-MAM_H.

4. Experimental Setup

In this section, we first introduce the face image database used to evaluate the methodology proposed in Section 3. This was followed by the description of the three recent deep-learning-based FIQA metrics and one general IQA metric used as examples in our work. Finally, we used ResNet100 trained with ArcFace loss as the basic FR model used as the backbone of our explainability efforts. A short experiment overview is provided before introducing the final results and more detailed analysis.

4.1. Database

VGGFace2 [4] dataset is a large-scale database containing face images with a large variety in quality distribution which makes it a challenging FR database. The images have

high diversity in poses and complex acquisition conditions. For the main analysis in this paper, we only used the official test dataset containing 500 subjects. To reduce heavy computation and a more balanced database, we randomly selected 30 face images of each subject representing the full database. This made a total of 15000 images, the list of randomly selected images from each identity will be made publicly available to enable reproducibility.

Data preprocessing includes face detection, cropping, and alignment. The Multi-task Cascaded Convolutional Networks (MTCNN) [30] framework is used to detect faces from the VGGFace2. The detected faces are further cropped and aligned using similarity transform to 112x112 pixels, such that all the face images are standardized for comparison.

4.2. FIQA methods

In our experiment, we selected four different quality metrics, one is an supervised general image quality called BRISQUE [22] as a baseline and three specifically designed methods for face image quality assessment, FaceQnetV1 [15], MagFace [21], and SER-FIQ [27], the first is a supervised FIQA and the later two are unsupervised. the following introduces these methods shortly.

BRISQUE [22] proposed by Anish et al. in 2012 is an opinion-aware image quality assessment method trained with human opinion scores. The assessment of the image is purely based on natural scene statistics learned from distortion-generic images. The method is built on the finding by Rudermann [24] that natural scene images have a luminance distribution similar to a normal Gaussian distribution. Handcrafted features were derived to quantify the deviation from the Gaussian due to image distortions. The quality estimations by BRISQUE were previously found to have a strong correlation to face image utility [9].

FaceQnet [15] proposed by Hernandez-Ortega et al. in 2019 is a supervised FIQA method. The BioLab-ICAO framework was used to label the ground-truth score for the training image according to the ICAO compliance level [17]. This score is used to train the regression layer on top of the feature extraction layers. FaceQnet is based on fine-tuning a pre-trained FR network (RseNet-50) and the successive regression layer to associate an input image to a utility score that determines the appropriateness of the input image to an FR model. In this work, we used the latest version published in [14], i.e. FaceQnetV1[2].

SER-FIQ [27] is an unsupervised deep-learning-based FIQA approach that applies stochastic variations on face representations learned from a deep-learning-based FR model by using dropout. This method mitigates the need for any automated or human labeling. The face image was passed to several sub-networks of a modified FR network by using different dropout patterns. Images with high utility are expected to possess similar face representations resulting in low variance. Thus, this proposed metric linked the robustness of face embeddings directly with FIQ.

[2]https://github.com/uam-biometrics/FaceQnet

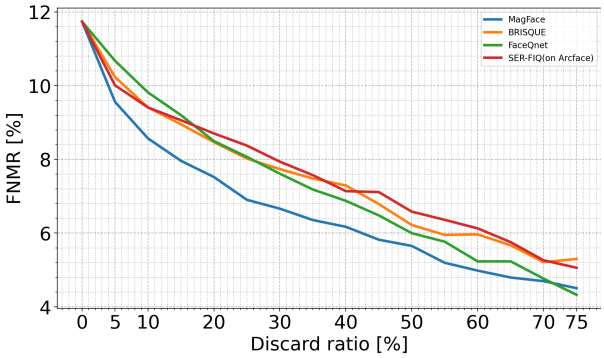

Figure 1. Error versus reject curve shows the FNMR at fixed FMR=0.1% decays with increasing number of worse quality images are discarded. All four FIQA methods seem to perform well as a strong decrease in error is observed when the predicted low-quality images were removed.

MagFace [21] by Meng et al. is another recently proposed unsupervised FIQA method based on using the adaptive loss incorporating the face image quality to the magnitude of the face embedding. This method can derive both the face representation and the face image quality from calculating the magnitude of the face embedding. The loss optimizes the inter-class variability and intra-class similarity. The MagFace version used in this work is trained on MS-Celeb-1M database and used the ResNet-100 as the backbone.

All selected FIQA methods perform well on the VGGFace2 database. This can be seen in the error versus reject characteristic (ERC) presented in Figure 1. The ERC shows the relative performance of the FR system when rejecting different ratios of the evaluation data with the lowest quality according to each FIQA metric. Figure 1 shows the ERC with the false non-match rate (FNMR) at different ratios of rejected (low quality) images using a fixed false-match rate (FMR) at 0.1%. The error clearly decreased as the number of worst quality samples are discarded.

In the overlapping ratio matrix shown in Figure 2, we see further that the set of the lowest and highest 10% sample images are not fully identical for different FIQA methods, indicating that the proposed metrics in Section 3 are derived from different base samples. The largest overlap is found for both unsupervised FIQA methods, i.e. for MagFace and SER-FIQ.

4.3. Face Recognition Solution

We use the ResNet-100 trained with ArcFace loss as the main FR solution to visualize the activation mapping of the input face images. This **ArcFace** [6] model is trained using

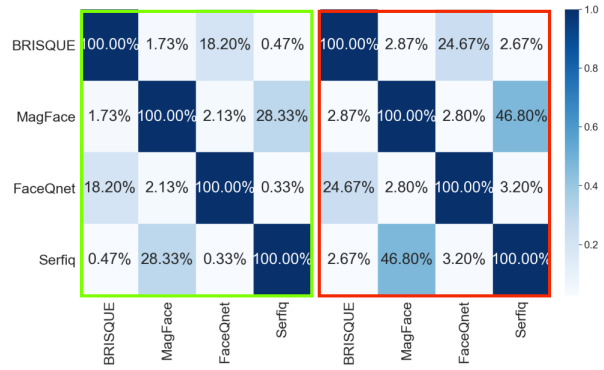

Figure 2. We displayed the samples' overlap ratio between the samples of the highest quality (10% of the data) on the right between every pair of quality estimation methods, and the same for the 10% of the lowest quality on the left.

the MS1M dataset [12]. The loss function applied additive angular margin to improve the discriminative power of the FR model. We chose this model because of its improved accuracy on LFW [16] 99.83% and YTF DB [29] 99.02%.

4.4. Activation mapping method

ScoreCAM [28] is used as the activation mapping method to display the activations of the deepest convolution layer of the Res-Net100 and upsampled to overlay to the input layer. This choice is motivated by the extensive successful use of ScoreCam as an activation mapping tool in various biometric domains [8, 7]. ScoreCAM provides a scaled version of the activation for the face embeddings. The weights of the scale factors are derived from this embedding. We chose ScoreCAM as it seems to provide more realistic explainability analyses to other methods such as CAM [31] and Grad-CAM [25].

4.5. Experiment overview

Given the methodology proposed in Section 3 and the experimental setup in Section 4, we intend to show: (1) difference between high and low-quality decision of FIQA based on AM-V/AM-MV and D-AM-MV, (2) differences within the low and high-quality decisions across FIQA methods, and (3) individual sample quality explainability with its AD-MAM of the individual FIQA method.

All four FIQA metrics were used to determine the face image quality of the used database. Out of which we determine the 10% of the face images with the highest and lowest face qualities to the group H and L, which makes 1500 face images in each of these two groups. ScoreCAM builds the activation mapping from the FR using ArcFace model for these input images of both groups H and L. For the FR model trained with ArcFace loss, we used the official Pytorch version from the official Github [3].

[3]https://github.com/deepinsight/insightface

5. Results and Analyses

This section is structured in three main parts related to explaining FIQ estimation within and across the decisions of different FIQA methods. The methodologies used as explainability tools are introduced in Section 3.

1. What makes the face images of high quality (in comparison to low) in the view of different FIQA based on the FR model behaviour? Figure 3 and Figure 4 depict the results for the MAM and MDAM for each FIQA method individually, we do not notice major differences between MAM and MDAM. Taking a look at the distributions for the mean activation mapping for H and L, no visible significant differences are noticed. This acts as the major motivation behind our proposed analyses based on activation map variation, rather than the activation maps themselves. Even though the difference for the MAM/MDAM between the H and L sets is not strongly visible, there are strong variations in the AM-V/AM-MV. Comparing the value distributions for AM-V_H and AM-V_L, we clearly observe that the variations for L are significantly larger compared to H, the same can be observed for AM-MV_H and AM-MV_L. This might indicate that the variability in activation mapping is stronger for low-quality face images. Looking at the AM-V/AM-MV for L and H, we noticed that the values are typically higher in L on the borders of the images, while it is higher in the middle face region for H. This indicates that the lower utility of L is based on the FR model focusing more often on the border areas rather than the center, in comparison to the H set. This result is further confirmed by looking at the D-AM-V mapping, where we can see that stronger deviations are observed on the left and right borders of the face image.

Generally, all considered FIQA methods lead to similar MA-V/AM-MV and D-AM-V/D-AM-MV observations, indicating that the effect of quality differences (as per different FIQAs) on the FR model is of the same nature. Based on these observations and to answer the question driving this subsection, we can notice that despite the similarity of the general activation maps of low and high-quality images (similar MAM/MDAM for L and H), what makes an image high quality is the consistent low activation on the areas outside of the face center, while low-quality images, despite general low activation in these areas, have high variations of activation there. In simple words, low-quality images do attract the attention of FR model in areas outside of the center face area, however, in different locations and less consistently, which can be caused by the different reasons for the degradation of quality.

2. What makes the decision of high or low quality different between different FIQAs based on the FR model

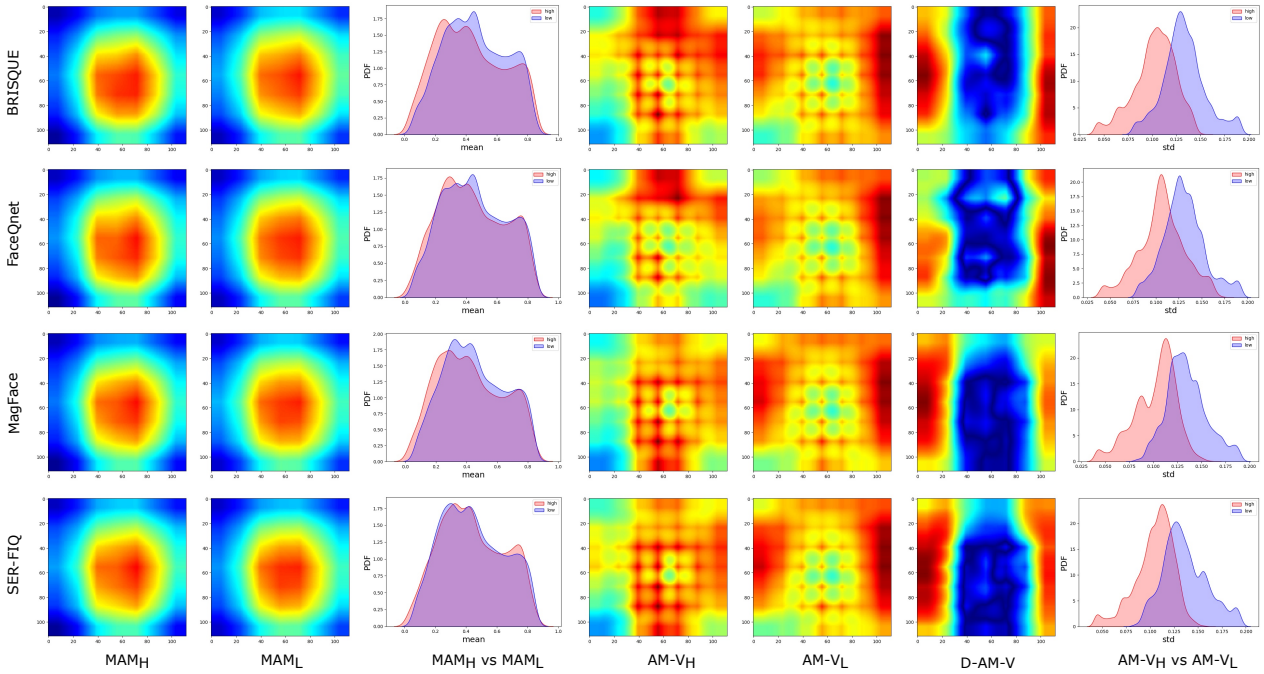

Figure 3. The results are derived from the MAM for each FIQA metric individually. Even though the MAM$_H$ and MAM$_L$ are similar for both H and L groups, a strong deviation can be observed in terms of AM-V, indicating low-quality images have a stronger deviation from the mean compared to high-quality images, especially on image borders. It is to be noted that for the visualization purpose, these deviation maps AM-V$_H$, AM-V$_L$, and D-AM-V are scaled.

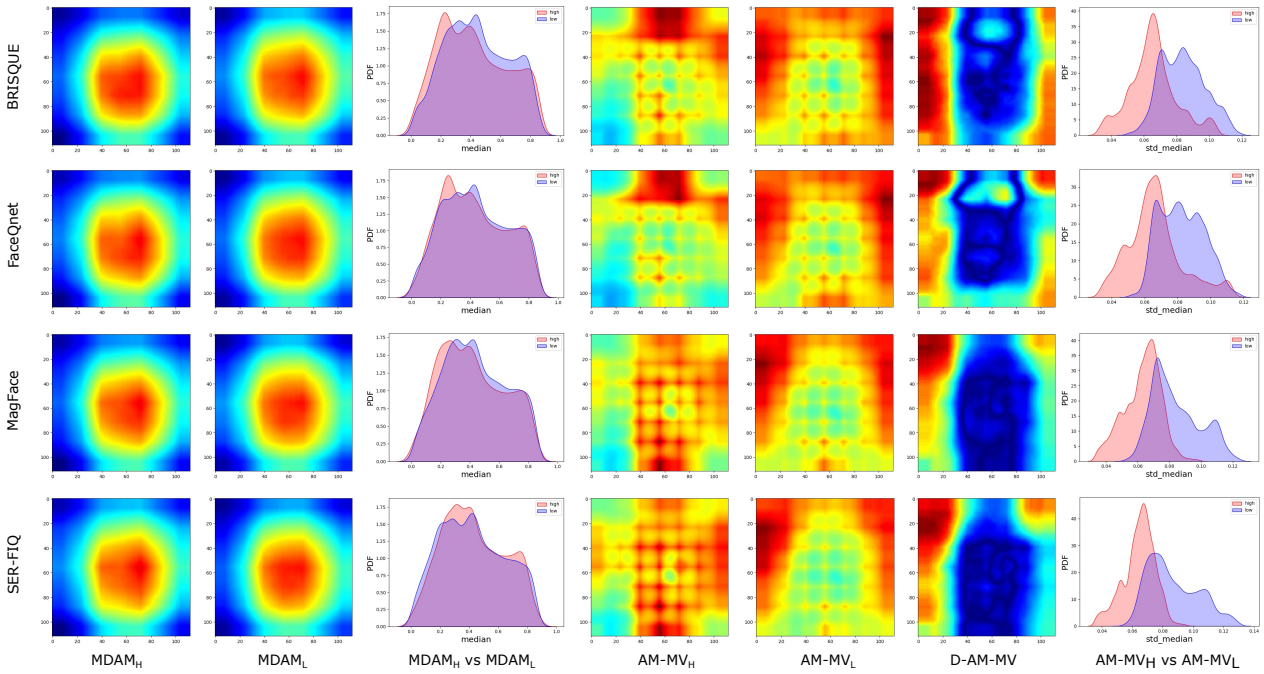

Figure 4. The results are derived from the MDAM for each FIQA metric individually. The same result is obtained as for MAM in Figure3, where even stronger variations are observed from the median for low-quality images. Low-quality images tend to attract the attention of FR model in areas outside of the center face area, which can be caused by different reasons like e.g., postures or occlusions. The variation maps are re-scaled for better visualization.

behaviour? In Figure 5 we see visualization of our D-AM-V mapping across FIQA methods. The D-AM-V here differs than the one presented in Equation (5) by looking across FIQA methods rather than quality sets, here the D-AM-V_H = AM-$V_{H,FIQA-1}$ − AM-$V_{H,FIQA-2}$ and D-AM-V_L = AM-$V_{L,FIQA-1}$ − AM-$V_{L,FIQA-2}$. D-AM-V_H and D-AM-V_L show the differential activation mapping variations for H and L sets individually between pairs of FIQA metrics. D-AM-V_H and D-AM-V_L between the unsupervised MagFace and the supervised methods FaceQnet and BRISQUE show similar tendencies. Samples selected to be H by the supervised methods tend to cause higher activation variations in the top and bottom of the image when compared to MagFace. Samples selected to be L by the supervised methods tend to cause higher activation variations on the left and right edges of the face image when compared to MagFace. This consistent activation variation on areas out of the face center can rationalize the performance differences between FaceQnet and BRISQUE on one side, and MagFace on the other side, see Figure 1. There are fewer differences in the activation variations caused by both the H and L set between the supervised (and with poorer performance) FIQA methods (BRISQUE and FaceQnet) both in terms of magnitude (distribution shifts) and clear spatial distribution. The same can be seen between the two high-performing unsupervised methods (SER-FIQ and MagFace). This leads to answering the question behind this sub-section by stating that compared to high performing FIQA methods, lower performing FIQA methods have larger FR activation variation on the edges of the face image.

How do the activation mappings caused by face images in the FR model deviate from what is expected from high quality images and how does that reflect in their quality score? In Figure 6 we depicted sample images of high and low face qualities. Each sample subject, we provided one original image with the FIQ score from all four FIQA metrics below, one image overlayed with the activation mapping from the ArcFace FR solution, and four overlayed with the AD-MAM$_H$ map of each FIQA method. These differential activation mappings are for BRISQUE, FaceQnet, MagFace, and SER-FIQA in the correct ordering starting from upper left to bottom right displayed for each sample subject. These AD-MAM$_H$ maps show areas which could cause the degradation in qualities as they deviate from the mean template and is emphasized in the visualization.

From Figure 6, it can be generally concluded that the AD-MAM$_H$ for all FIQA methods contain larger and higher values for low quality images in comparison to high quality images. Also in this comparison, the high value areas in the AD-MAM$_H$ for low quality images tend to appear more in the areas around the face rather than the center of the face. This is less apparent in the high quality images. This larger and higher values in AD-MAM$_H$ corresponds to the low quality estimated across the four FIQA methods and in many cases it is related to less than optimal poses, face occlusions, and overall low image sharpness. To answer the question motivating this sub-section, our analyses and proposed explainability tools reveal that low quality images tend to have a FR model activation map that strongly differs than that of what is expected from a high quality image. This difference also tends to appear more in the areas outside of the central face region.

6. Conclusion

Making the face image quality estimation explainable is a challenging task that goes beyond analyzing the FIQA network itself. Most recent works put more focus on designing FIQA methods that perform well without looking into what is the response of an FR model to high or low-quality face image. In this work, we presented a novel set of explainability tools to enhance the visual explainability of FIQ estimation decisions based on the variation analyses in FR models. The proposed tools can be applied for any FIQA method with any CNN-based FR solution using activation mapping to exhibit the network's activation derived from the face embedding. By showing the intra-groups and cross-method inter-groups statistics of the network's activation, we try to relate explainability to groups of H and L face quality image sets for the individual FIQA method. We demonstrate that even though the MAM between H and L is small, the variations in activation mapping for L are larger compared to H. This points out that the low-quality images tend to cause the FR network to focus on areas outside of the central face area, however, in an inconsistent manner, as the reason causing the low quality can vary largely. We additionally link this observation to the relative performance of different supervised and unsupervised FIQA approaches. Finally, we look at the explainability of the quality decision of individual face images by analyzing the differences between their activation maps in FR models and the maps expected from high-quality images, pointing out consistent differences in FR behavior between high quality and low-quality face images.

Acknowledgements: This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Lacey Best-Rowden and Anil K. Jain. Learning face image quality from human assessments. *IEEE Trans. Inf. Forensics Secur.*, 13(12):3064–3077, 2018.

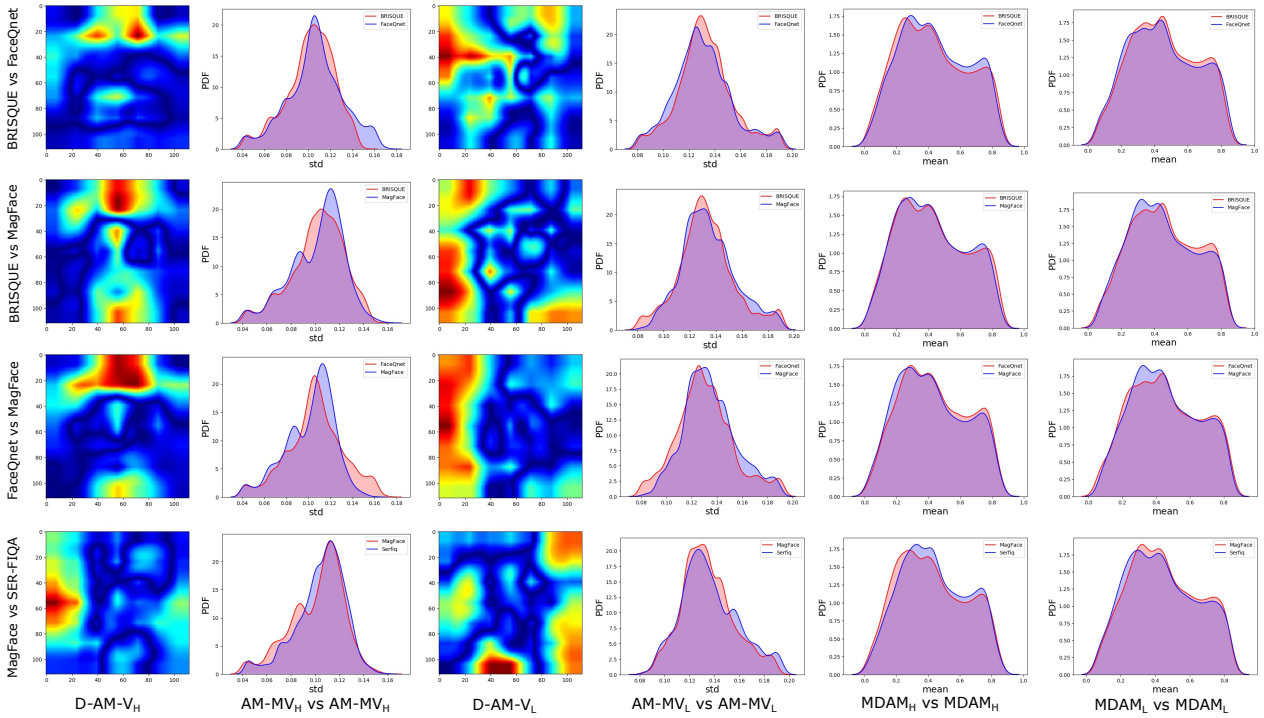

Figure 5. Results are derived when comparing D-AM-V for H and L sets across different FIQA methods. D-AM-V$_H$ and D-AM-V$_L$ between the unsupervised MagFace and the supervised methods FaceQnet and BRISQUE show similar tendencies. D-AM-V$_H$ and D-AM-V$_L$ tend to show stronger and more consistent differences between supervised and unsupervised methods, hinting a link between the lower performance of FaceQnet and BRISQUE and the focus of the FR model on areas outside of the face center in an unexpected manner
.

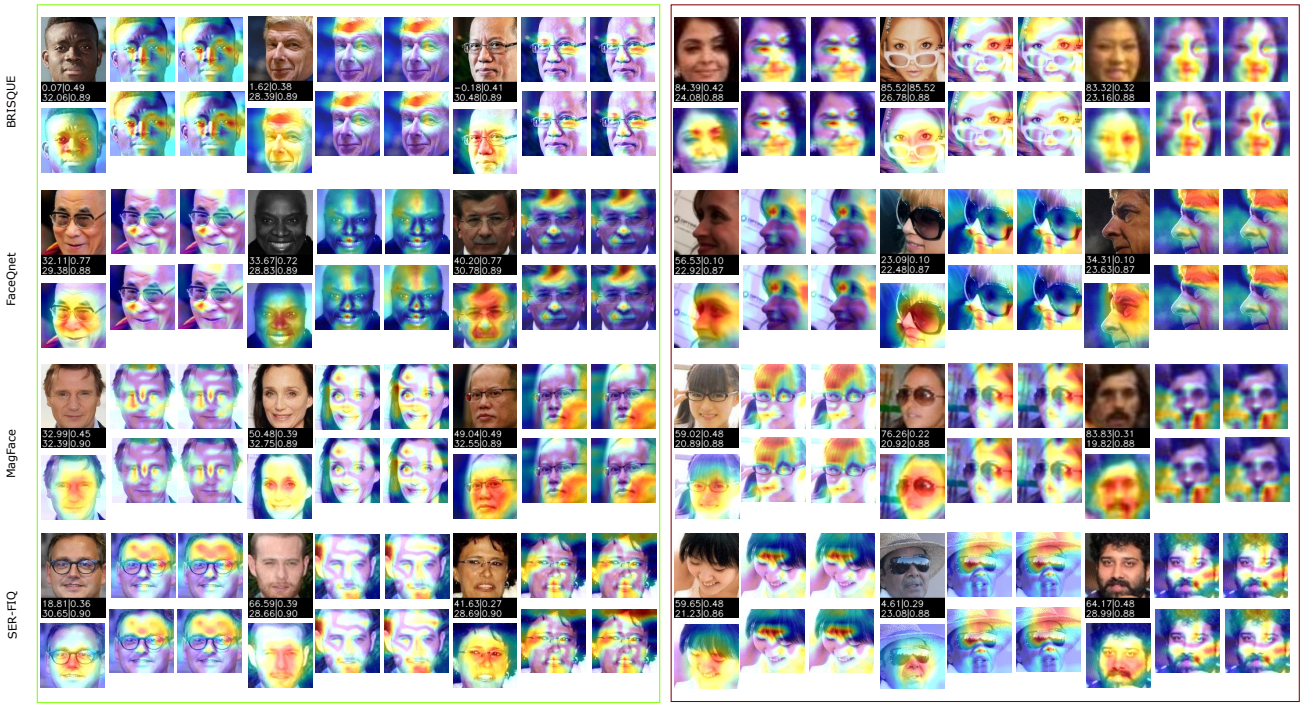

Figure 6. Selected face images with high and low face qualities are displayed. Each sample subject contains one original image with quality scores illustrated below, one image overlapped with the activation mapping derived from ArcFace FR, and four overlayed AD-MAM$_H$ maps. These differential maps indicate the deviations from the MAM$_H$ and show areas which could cause the degradation in qualities, such as sunglasses, hat, and mustaches.

[2] Fadi Boutros, Naser Damer, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Mixfacenets: Extremely efficient face recognition networks. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–8. IEEE, 2021.

[3] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. *CoRR*, abs/2109.09416, 2021.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 67–74. IEEE Computer Society, 2018.

[5] Naser Damer, Timotheos Samartzidis, and Alexander Nouak. Personalized face reference from video: Key-face selection and feature-level fusion. In *Face and Facial Expression Recognition from Real World Videos - International Workshop, FFER@ICPR 2014, Stockholm, Sweden, August 24, 2014, Revised Selected Papers*, volume 8912 of *Lecture Notes in Computer Science*, pages 85–98. Springer, 2014.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF CVPR*, pages 4690–4699, 2019.

[7] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–8. IEEE, 2021.

[8] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Real masks and spoof faces: On the masked face presentation attack detection. *Pattern Recognition*, 123:108398, 2022.

[9] Biying Fu, Cong Chen, Olaf Henniger, and Naser Damer. A deep insight into measuring face image utility with general and face-specific image quality metrics. *CoRR*, abs/2110.11111, 2021.

[10] Biying Fu, Cong Chen, Olaf Henniger, and Naser Damer. The relative contributions of facial parts qualities to the face image utility. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2021.

[11] Biying Fu, Florian Kirchbuchner, and Naser Damer. The effect of wearing a face mask on face image quality. *CoRR*, abs/2110.11283, 2021.

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conf. on Computer Vision*, pages 87–102. Springer, 2016.

[13] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6307–6315. IEEE Computer Society, 2017.

[14] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *CoRR*, abs/2006.03298, 2020.

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

[16] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, Oct. 2007.

[17] ISO/IEC JTC1 SC17 WG3. Portrait Quality - Reference Facial Images for MRTD. International Civil Aviation Organization, 2018.

[18] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC 29794-1:2016 Information technology - Biometric sample quality - Part 1: Framework. International Organization for Standardization, 2016.

[19] Khawla Mallat, Naser Damer, Fadi Boutros, and Jean-Luc Dugelay. Robust face authentication based on dynamic quality-weighted comparison of visible and thermal-to-visible images to visible enrollments. In *22th International Conference on Information Fusion, FUSION 2019, Ottawa, ON, Canada, July 2-5, 2019*, pages 1–8. IEEE, 2019.

[20] Brianna Maze, Jocelyn C. Adams, James A. Duncan, Nathan D. Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA janus benchmark - C: face dataset and protocol. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pages 158–165. IEEE, 2018.

[21] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2021.

[22] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012.

[23] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. SDD-FIQA: unsupervised face image quality assessment with similarity distribution distance. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7670–7679. Computer Vision Foundation / IEEE, 2021.

[24] Daniel L Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.

[25] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2):336–359, 2020.

[26] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6901–6910. IEEE, 2019.

[27] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 5651–5660, 2020.

[28] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

[29] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 529–534. IEEE Computer Society, 2011.

[30] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016.

[31] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society, 2016.