Multi-Task Explainable Quality Networks for Large-Scale Forensic Facial Recognition

Andrea Macarulla Rodriguez, Luis Unzueta, Zeno Geradts, Senior Member, IEEE, Marcel Worring, Senior Member, IEEE, and Unai Elordi

Abstract—Identifying suspects from surveillance footage is a crucial task in forensic investigations, but it is often hindered by the variable conditions of observation and the large amounts of data. Face image quality (FIQ) is a metric that measures the usefulness of a face sample for facial recognition. Existing methods for automated FIQ assessment only provide a scalar value for quality, and do not indicate which factors are causing low quality. Additionally, these methods are computationally expensive, which makes current FIQ assessment methods unsuitable for large numbers of images. To address these issues, we introduce multi-task explainable quality networks (XQNets). XQNets provide both the quality value and the associated facial and environmental attributes, and automatically learn how each attribute contributes to the quality value during the training process. We also propose a dataset-agnostic quality pairing protocol to ensure that sample pairs are balanced across datasets and evaluations are fair. Our experiments on the LFW and SCFace benchmarks show that our approach generalizes well across different datasets and outperforms state-of-the-art methods. Our method offers a fast, explainable approach to FIQ assessment, making it suitable for large-scale forensic applications.

Index Terms—Face image quality, explainable AI, multi-task learning, forensics.

I. INTRODUCTION

Face recognition (FR) has improved significantly due to advancements in facial recognition algorithms and methodologies [2], [3]. Despite the improvements, recognition error rates remain high in real-world forensic applications, particularly in challenging conditions such as CCTV footage or ATM cameras [4], [5], [6]. As FR systems play an increasingly larger role in crucial decision-making processes, there is a growing need to explain the FR process to humans [7], [8]. As a solution, Face Image Quality (FIQ) assessment methods have been developed to output a quality score that can be represented as a single scalar value or a vector of values to explain how suitable a face image is for face recognition [9], [10], [11]. However, there have been few studies to explain these scores to users and provide an interpretable cause for a face image’s low or high quality [9], [12]. An example of an explainable FIQ analysis pipeline is presented, where images with low quality scores are rejected and accompanied by attributes that help humans understand the system’s decision [12]. This makes FR not only accurate but also explainable.

General image quality assessment algorithms such as BRISQUE [13], NIQE [14], and PIQE [15] do not achieve satisfactory performance when applied to face images because they aim to assess images in terms of subjective (human) perceptual quality [5]. On the other hand, FIQ assessment algorithms are concerned with the assessment of the biometric utility of facial images, which can be objectively defined in the context of specific FR systems. Hence, FIQ assessment methods obtain more accurate results for FR applications [7], [16], [17]. This occurs because FIQ algorithms for the purpose of biometric utility prediction can perform better than a general image quality assessment that has not been developed with facial biometrics in mind.

Predicting recognition utility in FR implies that the quality score has to indicate the “accuracy” or “certainty” of comparison scores generated for a sample pair that includes the assessed sample [7]. Thus, quality should be indicative of the face comparison performance. Note that this entails that the output of a specific FIQ assessment algorithm may be more accurate for a specific FR system. So the FIQ assessment utility prediction effectiveness ultimately depends on the combination of both the FIQ assessment algorithm and the FR system. According to [17], it is desirable to facilitate interoperability such that the FIQ assessment algorithm is predictive of recognition performance in general for a range of relevant systems instead of being dependent on only one.

FIQ measures enable various forensic applications. For example, in real-time recording sessions, photos can be accepted or rejected based on their scalar image quality values. If the image
quality is too low, the system will reject it and collect a new image, which is particularly valuable during first enrolment when a reference photo is not available [5]. Scalar image quality values can also be used as a management indicator by summarizing the effectiveness of the collection process across different sites and conditions [18]. Additionally, they can be used to select the best image from a set of photos [9]. It would be useful to have a FIQ that can explain why an image cannot be used and which facial or environmental attributes the subject needs to improve to increase the quality.

Regarding FIQ assessment, literature often focuses on optimizing quality scores on benchmarks such as LFW or Adience [7], [8], [19]. However, there is a risk of saturation on datasets such as LFW [4], which led to the proposal of XQLFW, a benchmark derived from LFW with pairs of maximum quality difference [4]. The selection of images for XQLFW is based on BRISQUE and SER-FIQ quality scores, but this selection may introduce a bias towards SER-FIQ [4]. Additionally, LFW and XQLFW come with a predefined set of 6000 pairs for evaluation, requiring the generation of a new set of pairs for new datasets like SCFace or ForenFace. To overcome this issue, this study proposes a dataset-agnostic quality pairing (DAQP) protocol to ensure a balanced representation of the whole spectrum of qualities in pair generation. The study evaluates the widely used datasets LFW [20], XQLFW [4], and DAQP on forensic datasets such as SCFace [1] and ForenFace [21].

The overall FIQ value is related to descriptive facial attributes such as deviation from the frontal pose or hot spots; and environmental attributes such as sharpness or deviation from uniform illumination. One way to consider all these variables would be to process each separately and combine the scores afterwards but this would not allow to learn the common aspects and it would be inefficient. Learning paradigms such as multi-task learning (MTL) [22] could help leveraging the domain-specific information in the training signals of related tasks [23], [24], e.g., the FIQ value, and its related attributes to generalize better across different FR models. Moreover, MTL provides several outputs with a single forward inference, which allows accelerating the computation significantly. Thus, in this work, we study how MTL can be exploited to build efficient explainable FIQ assessment systems for large-scale forensic FR applications. More specifically, the contributions we make in this paper are:

1) Multi-task explainable quality networks (XQNets) to efficiently assess FIQ value along with a set of facial and environmental image attributes that explain the calculated FIQ.

2) A dataset-agnostic pairing (DAQP) protocol to evaluate explainable FIQ assessment systems using the whole range of FIQ values in the test dataset ensuring that sample pairs are balanced.

3) Experimental results with the LFW and SCface FR benchmarks, following the DAQP protocol, demonstrating that XQNet has an accurate FIQ across different state-of-the-art FR assessment methods in complex surveillance scenarios and with competitive inference times.

The rest of the paper is organized as follows: Section II describes prior related work; Section III explains our proposed XQNet and DAQP protocol; Section IV presents experimental results with LFW and SCface FR benchmarks. Finally, Section V presents the conclusions and future lines of work.

II. RELATED WORK

FIQ assessment algorithms can be classified as factor-specific and monolithic approaches. The former comprises methods for finding interpretable factors, such as blur and sharpness, which could help an operator to avoid inadequate face images when recapturing. The latter outputs an overall FIQ value leading to comparatively opaque assessments/quality scores.

Factor-specific approaches to FIQ assessment are based on either facial attributes (e.g. inter-eye distance, pose) or environmental attributes (e.g. illumination, blur) [9], [25], [26], [27]. For example, in [25], the authors trained and compared ten features of quality estimates of a single human to assess general image quality. In [26], the authors estimate only the pose angle without producing a normalized quality score, demonstrating that pose estimation can be used for FIQ assessment. In [27], 17 parameters based on ICAO Doc 9303 requirements are used to evaluate FIQ, resulting in an 88% correct classification rate. In [9], FaceQvec is proposed as a method to estimate the conformity of facial images with ISO/IEC 19794-5, a quality standard for face images in official documents. The method consists of 25 individual tests related to the standard and other
image characteristics, with accurate evaluation results. However, these methods tend to use high-quality images for evaluation and little attention is given to forensic applications.

The monolithic approaches are divided into: human ground truth training, FR-based ground truth training, FR-based inference and FR-integration. Hernandez-Ortega et al. proposes the FaceQnet model [17] with versions v0 and v1. For both versions, as part of the training data preparation, the BioLab-ICAO framework from [28] is employed to select suitable high-quality images per subject, which are used to compute the ground truth quality scores for the subjects’ remaining training images. This ground truth quality score computation consists of the normalized Euclidean distances of embeddings produced by a number of FR feature extractors (three for v1; and only one, FaceNet [29], for v0). Both FaceQnet versions were based on a ResNet50 [30] model pretrained for FR using the VGGFace2 dataset [10], replacing the final output layer with two fully connected layers which are used for finetuning while the rest of the network weights were frozen.

SER-FIQ is a model proposed by Terhörst et al. with two variants: “same model” and “on-top model.” Both variants estimate the quality of FR by comparing the embeddings of randomly chosen subnetworks without ground truth quality labels. The “same model” variant can be used on FR networks trained with dropout and the “on-top model” variant uses a small additional network trained with dropout on top of the FR model. The evaluations showed that the “on-top model” variant mostly outperformed the baseline approaches and the “same model” variant showed strong FNMR performance improvement for a fixed FMR of 0.001.

The MagFace model [8] integrates quality and FR, with quality directly indicated by the magnitude of the FR feature vector. The model extends the ArcFace [31] training loss with a magnitude-aware angular margin and magnitude regularizer, resulting in larger magnitudes for higher quality images and smaller magnitudes for lower quality images. The magnitude is bounded during training, and a normalized quality score can be derived through linear scaling. The FR function after training remains unchanged from ArcFace.

The Pixel-Level FIQ approach [12] allows evaluating the pixel-level attributes of a face picture given an arbitrary FR network that does not require any training. A model-specific quality value of the input picture is computed and utilized to develop a sample-specific quality regression model to do this. Using this technique, quality-based gradients are back-propagated and translated into pixel-level quality estimates. They evaluate the significance of their suggested pixel-level features subjectively and quantitatively using actual and fake disruptions and compare explanation maps on faces that do not meet ICAO rules. The findings show that the suggested method creates meaningful pixel-level characteristics that improve the interpretability of the full facial picture quality in all cases.

Ou et al. proposed SDD-FIQA [16], a FIQ method that considers both the intrinsic properties and the recognizability of the face image. They argue that a high-quality face image should be similar to its intra-class samples and dissimilar to its inter-class samples. Thus, their method generates quality pseudo-labels by calculating the Wasserstein distance between the intra-class similarity distributions and inter-class similarity distributions. With these quality pseudo-labels, they are capable of training a regression network for quality prediction. Their method shows good generalization across different recognition systems. However, they do not provide the set of attributes that would affect the high or low quality of an image.

MTL has been applied to FR tasks such as landmark detection and anti-spoofing, but not to FIQ estimation. In [32], MTL is used for landmark detection and improves performance for faces with severe occlusion and pose variation. In [33], AENet uses rich semantic annotations as auxiliary tasks to boost the performance of face anti-spoofing. In [34], MHCNN is proposed for joint face detection, landmark detection, facial quality, and attribute analysis, but it is only used for face detection, not recognition. To date, there are no MTL methods applied to FIQ estimation.

III. METHODOLOGY

The proposed multi-task learning model consists of 3 steps. First, for a given facial image, the face must be pre-processed to obtain an input image of (3, 112, 112). Second, the input image is processed by the network, and third, the vector with facial and environmental attributes output will be obtained. The framework of the proposed multitask learning model XQNet is shown in Fig. 2. The model consists of a body with several heads. Each head processes one of the facial or environmental attributes. The loss function combines all of the head attribute outputs assigning a weight to each. Finally, the model outputs the target FIQ together with facial and environmental attributes that contribute to the decision for such quality.

To choose a backbone for the network, one has to take into account that better accuracies tend to be obtained by more complex DNN architectures that require significant computational resources. In our context, it is desirable that the network is able to process a large amount of images in a short time, and has a good trade-off between accuracy and performance. It is for this reason that the backbones chosen are EfficientNet [35] and ConvNext [36]. Vision transformers (ViTs) are another type of DNNs that are receiving the attention of the computer vision community recently, as they have demonstrated superiority in accuracy over CNNs [37]. But they currently have higher computational costs and therefore require further research on optimization techniques to efficiently deploy them in resource-constrained processors [38].

For MTL we use the hard parameter sharing approach [39]. It is generally applied by sharing the hidden layers between all tasks, while keeping several task-specific output layers. Hard parameter sharing greatly reduces the risk of overfitting [40]. The more tasks we are learning simultaneously, the more our model has to find a representation that captures all of the tasks and the less chance of overfitting on our original task, i.e. FIQ.

A. Training

The training pipeline is as follows: the landmarks of the facial image are first detected. In the case of the face not being detected,
the image is not considered for training. Then the landmarks are used to crop and align the face in order to yield the appropriate shape (3, 112, 112) to be inserted into the network. In our method, we rely on detecting and aligning faces as the input to our model because the quality assessment should be performed on an aligned face in order to ensure accurate results. The model expects an aligned face as input in order to process the quality assessment in an adequate manner. The network itself is formed by a backbone and several heads. The outputs of the heads go to the loss function. There are seven regression tasks to estimate the following variables: FIQ, number of hotspots, sharpness, deviation from uniform lightning, deviation from frontal pose, and age. Regression tasks are those which output a continuous variable. Moreover, there are two categorical variables to learn: ethnicity and gender. Category tasks are those which output a class value. Each of these tasks are represented in the network as a head that comes from the backbone.

The cost function is based on the work of [23] and has the weights as trainable parameters of the network. The loss function is defined as:

$$L(W, \sigma_1, \sigma_2) = \frac{1}{2\sigma_1^2}L_1(W) + \frac{1}{2\sigma_2^2}L_2(W) + \log(\sigma_1) + \log(\sigma_2),$$

where $\sigma_1$ is the observation noise. The variable $\sigma_1$ represents the noise parameter for the model output $y_1$ (regression) and $\sigma_2$ represents the noise parameter for the model output $y_2$ (categorical). The losses $L_1$ and $L_2$ are defined by:

$$L_1(W) = \| y_1 - f^W(x) \|^2,$$

and:

$$L_2(W) = -\log \text{Softmax}(y_2, f^W(x)).$$

The equations 2 and 3 can be used for each of the regression and categorical tasks of our model, allowing us to learn the relative weights. This loss is smoothly differentiable, and it ensures that the task weights will not converge to zero. In addition to the model of [23], we apply a GELU [41] function before introducing the weights in the valid loss, both to ensure that the valid loss does not get in negative values and resulting in better training results.

The model is trained using [42], which demonstrates an improvement in the learning speed with regards to a cycle in which the learning rate ($lr$) and momentum are kept constant. It consists of the following steps: first, we progressively increase our $lr$ from $lr_{\text{max}}/f$ to $lr_{\text{max}}$ and at the same time we progressively decrease our momentum from $\text{mom}_{\text{max}}$ to $\text{mom}_{\text{min}}$. Second, we do the exact opposite: we progressively decrease our $lr$ from $lr_{\text{max}}$ to $lr_{\text{max}}/f$ and at the same time we progressively increase our momentum from $\text{mom}_{\text{min}}$ to $\text{mom}_{\text{max}}$. Thirdly, we further decrease our $lr$ from $lr_{\text{max}}/f$ to $lr_{\text{max}}/(f \times 100)$ and we keep momentum steady at $\text{mom}_{\text{max}}$.

### B. Evaluation Protocols

As done in [7], [8], [19], we use EVRC to evaluate the performance of FIQ assessment methods. The EVRC uses the partial area under the curve defined as:

$$\text{pAUC} = \int_0^\alpha \text{FNMR}(\varphi) d\varphi,$$

where $\varphi$ is defined as the percentage of images which are not considered and FNMR is the False Negative Match Rate at the given $\varphi$. For convenience and to be able to compare with The FNMR is defined as the number of false negatives (negative facial recognition claims which should have been accepted) divided by the total amount of real positives (false negatives or FN + true positives or TP) i.e.:

$$\text{FNMR} = \frac{FN}{FN + TP}.\quad (5)$$

FNMR is a useful metric for evaluating the performance of a face recognition system. In face recognition, false negative errors refer to the situation where the system fails to match a pair of face images that belong to the same person, i.e., it wrongly classifies them as different persons. A low FNMR indicates that the system is able to accurately match faces that belong to the same person. FNMR is particularly relevant in security and surveillance scenarios where failing to recognize a person can have serious consequences. For example, a false negative
error could result in a person being incorrectly denied access, while in a criminal investigation, it could result in a suspect going undetected.

The use of the EVRC is beneficial because it demonstrates the effect of discarding low-quality face images on FR performance, as measured by FNMR. This curve shows the relationship between FNMR and reject rates, allowing us to understand how FNMR changes as an increasing amount of low-quality data is discarded. Using the EVRC curve is a fair method to compare the performance of different FIQ assessment algorithms, as it is independent of the absolute quality score values and their range. Additionally, the use of the ERVC provides a clear, visual representation of the relationship between FIQ and FR performance, making it an informative and effective evaluation tool.

When evaluating LFW, 6000 randomly generated pairs were used as a benchmark. We also used the 6000 pairs provided for XQLFW [4]. From each pair, the image with the minimum quality is taken as the ‘pair quality’. Performing FR in all the pairs, FNMR is computed. Then 5% of the worse quality pairs are removed and FNMR recomputed. The process is iterated until no more pairs are left. However, this method does not contemplate all the qualities in the databases nor can it be extended to other databases that do not have a standard set pairs to evaluate such as LFW or XQLFW. If a new database arises such as SCFace [1], and it does not come with a preset evaluation of pairs. Should we generate one ourselves randomly? Even if the pair set is already available such as in LFW. Is it the most suitable for quality evaluation? XQLFW [4] proved that it is not, but they used both SER-FIQ and BRISQUE quality to generate the pairs, which can make the method biased. It is for this that we propose a new method which allows to extend the evaluation to other databases that do not have pre-established pairs.

First, we compute the quality for each image in the dataset. After that, we compute a histogram with \( n = 20 \) bins. For each bin, we compute the maximum number of pairs of the same quality that are available. In this way, we make sure that when performing face comparison pairs of similar quality are compared and not pairs that have very different qualities. Another way to form these pairs would be to perform cross-bin comparisons so that very high-quality images are compared against very low qualities, but as we are removing pairs with the lowest quality, and to be fair in the comparison (the quality of the pair is the minimum of the two image qualities), we decide to adopt this criterion of making pairs of similar quality and not maximizing the difference.

### Algorithm 1: Dataset-Agnostic Quality Pairing (DAQP).

**Input:** Set of images annotated by quality \( I \)

**Output:** Set of evaluation pairs \( P \)

1. **DAQP** \( (I, n = 20) \)
2. Distribute \( I \) in \( n \) quantiles based on quality
3. **for** \( \text{quantile} = 1 \) to \( n \) **do**
   - \( si \leftarrow \) all pair combinations of same identity
   - \( nsi = \text{len}(si) \)
   - \( di \leftarrow nsi \) pair combinations of different identity
4. **end for**
5. \( P \leftarrow \) empty list
6. \( \text{minp} \leftarrow \text{find min}(nssi) \) in all the quantiles
7. \( P = P.add(si, di) \) where quartile\((si, di)\) has minp
8. **iterate over the rest of quantiles**
9. **for** \( \text{quantile} = 1 \) to \((n-1)\) **do**
   - \( si = si.randomselect(minp) \)
   - \( di = di.randomselect(minp) \)
   - \( P = P.add(si, di) \)
10. **end for**
11. **return** \( P \)

Once the quality pairs are computed, the same number of different identity pairs are obtained. Once this is done for all the bins in the histogram, we compute the FNMR 20 times each time removing the bin, the lowest quality. A summary of the DAQP algorithm is found in Algorithm 1. A graph showing the pair-quality distribution for the datasets of LFW, XQLFW and DAQP is shown in Fig. 4. The FNMR performance of the FIQ model using DAQP (now renamed DAQP) provides a more comprehensive understanding of the model’s performance across a diverse range of image quality, as compared to evaluating the model using randomly selected pairs. This is a crucial aspect in assessing the generalization capability of the FIQ method. Samples of different evaluation pairs are shown in Fig. 3.

### IV. Experiments

We use two types of databases: forensic-oriented, where the images have low-resolution, are taken at a distance, or the subjects have very different poses such as SCFace [1] and ForenFace [21], and standard databases used commonly in literature to test FIQ algorithms, such as LFW [20], XQLFW [4] or UTKFace [44]. We use UTK Face for training and for testing we use LFW, XQLFW, SCFace and ForenFace. LFW and XQLFW are both
public datasets, and although LFW almost reaches saturation in most FR systems [4], it is still widely used in literature for FIQ evaluation. On the other hand, SCFace and ForenFace are closed datasets but much more focused in forensics, proposing more challenging images for quality evaluation.

LFW [20] is a database of 13,000 images of faces collected from the web. 1680 of the people pictured have two or more distinct photos in the data set. SCFace [1] is a database in which images were taken in uncontrolled indoor environment using five video surveillance cameras of various qualities. The database contains 4160 static images (in the visible and infrared spectrum) of 130 subjects. ForenFace [21] contains video sequences and extracted images of 98 subjects recorded with six different surveillance camera of various types. Moreover, it also contains high resolution images and 3D scans for these subjects. A subset of 435 images (87 subjects, five images per subject) has been manually annotated, yielding a unique and very rich annotation containing almost 19.000 entries. It also contains a training/testing protocol. The detector used is MTCNN [45] and the face recognizers were ArcFace [31] and FaceNet512 [29]. These face recognizers where implemented using the library DeepFace [3]. The annotations used are as follows: the FIQ uses SDD-FIQA [43] for ground truth annotation, due to being to our knowledge the state of the art in quality estimation. Gender and Ethnicity are taken from the annotated UTK face database [44]. In case the ethnicity output in UTK dataset was “others”, the software annotation used was [46]. The rest of the annotations, which are: number of hotspots, sharpness, deviation from uniform lightning and deviation from frontal pose are estimated using the commercial software library FaceVACs [47]. A sample of these annotations is shown in Fig. 5. A summary of the attribute definitions is in Table I. The purpose of XQNet is to be open software, so new training with other annotations and other databases is possible, making it less of the black-box that FaceVACs commercial software is.

V. RESULTS

The proposed explainable face quality estimation is analysed in four ways. It’s important to note that determining what

TABLE I

<table>
<thead>
<tr>
<th>Type</th>
<th>Attribute</th>
<th>Description</th>
<th>Range / Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>Hot Spots</td>
<td>Bright areas of light reflected from the face</td>
<td>0-12000</td>
</tr>
<tr>
<td></td>
<td>Sharpness</td>
<td>Focus and depth of field according to specification of ISO 19794-5:2005 section 7.3.3.</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Deviation from uniform lightning</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Deviation from frontal pose</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-</td>
<td>0-116</td>
</tr>
<tr>
<td></td>
<td>Ethnicity</td>
<td>-</td>
<td>White, Black, Asian, Indian, Hispanic, Middle Eastern</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>-</td>
<td>Restricted to Male/Female</td>
</tr>
</tbody>
</table>

Fig. 4. Set of qualities for all the state-of-the-art methods used, including our own proposed method XQNet-ConvNext and XQnet-EfficientNet.

Fig. 5. Sample of dataset annotations.
constitutes a “good enough” image is not an absolute term and depends on the desired FNMR and the specific dataset being evaluated. A higher FIQ score generally indicates a higher quality image and a higher likelihood of successful person identification, but the appropriate threshold will depend on the user’s specific requirements and desired trade-off between false negatives and false positives. First, we used the pair-generation method DAQP algorithm 1 to compare the FNMR results in three datasets. Second, we compare our pair-generation method DAQP against 6000 randomly generated pairs from LFW and the algorithm of 6000 pairs used in XQLFW. Thirdly, as our method is intended for forensic large-scale usage, we compare both CPU and GPU performance times of different FIQ algorithms and fourthly, we show the attribute distributions produced by our explainability method in the different datasets.

Our second analysis consists of comparing our pair-generation algorithm DAQP method against the 6000 randomly generated pairs in LFW [20] and the algorithm for generating the qualitative results are shown in Fig. 6, whereas the quantitative results are shown in Table II. the numbers in the table represent the evaluation in terms of partial area under the curve (pAUC) for reject fraction ranges from 5%, 15% and 35%. Lower values indicate lower FNMR, and thus, better performance of the model. For each rejection range, we have marked in bold the minimum FNMR, which indicates the best performance at that percentage of discarded images (see (5) and (4)).

Our pair-generation method DAQP Algorithm 1 is used to compare the FNMR results in three datasets, (SCFace, ForenFace and LFW). Moreover, to assess the generalization against different Face Recognition systems, we perform the pair verification with ArcFace, FaceNet512 and SFace. The qualitative results are shown in Fig. 6, whereas the quantitative results are shown in Table II. the numbers in the table represent the evaluation in terms of partial area under the curve (pAUC) for reject fraction ranges from 5%, 15% and 35%. Lower values indicate lower FNMR, and thus, better performance of the model. For each rejection range, we have marked in bold the minimum FNMR, which indicates the best performance at that percentage of discarded images (see (5) and (4)).

Our second analysis consists of comparing our pair-generation algorithm DAQP method against the 6000 randomly generated pairs in LFW [20] and the algorithm for generating the qualitative results are shown in Fig. 6, whereas the quantitative results are shown in Table II. the numbers in the table represent the evaluation in terms of partial area under the curve (pAUC) for reject fraction ranges from 5%, 15% and 35%. Lower values indicate lower FNMR, and thus, better performance of the model. For each rejection range, we have marked in bold the minimum FNMR, which indicates the best performance at that percentage of discarded images (see (5) and (4)).

Our second analysis consists of comparing our pair-generation algorithm DAQP method against the 6000 randomly generated pairs in LFW [20] and the algorithm for generating the qualitative results are shown in Fig. 6, whereas the quantitative results are shown in Table II. the numbers in the table represent the evaluation in terms of partial area under the curve (pAUC) for reject fraction ranges from 5%, 15% and 35%. Lower values indicate lower FNMR, and thus, better performance of the model. For each rejection range, we have marked in bold the minimum FNMR, which indicates the best performance at that percentage of discarded images (see (5) and (4)).

Our second analysis consists of comparing our pair-generation algorithm DAQP method against the 6000 randomly generated pairs in LFW [20] and the algorithm for generating the qualitative results are shown in Fig. 6, whereas the quantitative results are shown in Table II. the numbers in the table represent the evaluation in terms of partial area under the curve (pAUC) for reject fraction ranges from 5%, 15% and 35%. Lower values indicate lower FNMR, and thus, better performance of the model. For each rejection range, we have marked in bold the minimum FNMR, which indicates the best performance at that percentage of discarded images (see (5) and (4)).
6000 pairs in XQLFW [4]. Equally to the first analysis, the quantitative results are shown in Table III.

In both tables, the fraction of images that were deemed “unconsidered” refers to the percentage of images that were excluded (based on their quality scores) from the calculation of FNMR. If the quality model is accurately assigning quality scores, then discarding lower quality images for the computation of FNMR should result in improved performance and thus, lower FNMR. Conversely, if the FNMR deteriorates upon discarding low quality images, it suggests that these images were not in fact of low quality. In Tables II and III best results for each database have been marked in bold. We observe that SDD-FIQA [43] and EQFace [49] tend to have better performance on LFW and XQLFW benchmarks, whereas XQNET (ours) has better performance with the DAQP evaluation protocol.

Finally, we analyze the explainability component of the XQNets. The set of attributes is predicted and plotted in (7) and (8). In Fig. 7, the correlation between pairs of variables is depicted for each database. The visual representation of the correlation between the variables can provide important insights into the relationship between the variables under study. The more concentrated and circular the curves are, the greater the correlation between the pair of variables. This can be seen in the quality/sharpness graph, where the dataset LFW lines occupy a very small area of the plane, indicating a strong correlation between these two variables. Conversely, in the hotspots/quality graph, the red curves (SCFace database) are widely dispersed, covering a fairly extensive area, indicating a weaker correlation both with CPU and GPU. The CPU used was AMD EPYC 7B12, and the GPU used was Tesla T4. As seen in the table, our method performs competitively against other state-of-the-art algorithms such as SER-FIQ [7], EQFace [49] and SDD-FIQA [43].
between the two variables. This information can be used to make informed decisions about the variables that are most important to focus on for a given study or analysis. In addition to the correlation between variables, the diagonal of the figure also provides information about the distribution of each individual variable. Some variables have a pointed distribution, with a greater concentration around a dominant value, such as quality or hotspots in LFW database. Other variables have a flattened distribution, showing an almost uniform distribution within a range, such as age in LFW. Understanding the distribution of

Fig. 7. Attribute regression predictions by XQNet.

Fig. 8. Attribute categorical predictions by XQNet.
each variable can inform data preprocessing and modelling decisions, as well as provide insight into the underlying structure of the data.

Fig. 8 displays the distribution of each categorical parameter value for each database. For example, in the LFW database, with regards to ethnicity and gender, it can be observed that there are no samples of women from the Middle East, while the quality of those for the same ethnicity for men is highly concentrated. The elongated and thick graphs with little variation indicate a wide dispersion of the variable under study (quality), as can be seen with the Indian ethnicity in the SCFace database. In addition to the information about the distribution of each parameter value, Fig. 8 can also provide valuable insights into the underlying structure of the data. The distribution of values for each parameter can reveal the presence of biases or imbalances in the data, which can affect the results of further analysis and modelling. Understanding these patterns can inform the development of data preprocessing and balancing techniques, as well as provide guidance for future data collection efforts. Furthermore, the distribution of values for each parameter can also inform decision-making in terms of model selection and performance evaluation.

VI. DISCUSSION AND CONCLUSION

Current definitions of face quality assessment are based on the suitability of a face image for the task of face comparison. However if these FIQ applications are meant to be used by humans, such as in forensics, this suitability score has to be accompanied with a sufficient degree of explainability. This explainability can be achieved through pixel values, such as the work of [12], or by measuring a set of standard attributes and weighting the contribution of each of them such as the work of [9]. In this work, we have chosen to develop a multi-task learning model that jointly learns the suitability score with the facial and environmental attributes that contribute to it. The results show that FIQ highly correlates with sharpness, frontal pose and age. This can help the user to get real-time feedback on how to improve the quality of the image before further processing. Also, for forensic purposes in large databases, clusters of images with different qualities and different attributes can be produced to facilitate the investigation. As a caveat, it has to be mentioned that in cases where proper face detection and alignment are not possible, these images cannot be considered for computation. However, manual detection and alignment may be performed with the use of adequate software. This is important because XQNet relies on detecting and aligning faces as the input to our model because the quality assessment should be performed on an aligned face in order to ensure accurate results. The model expects an aligned face as input in order to process the quality assessment in an adequate manner. Another limitation of our current work is that it depends on the availability of labeled data for training and evaluating the model. This implies that the balance of the datasets used of training can have an impact on the results, and the results obtained from our work only apply to the datasets and embedding models used in the study and may not generalize to other datasets or models. Another aspect of training the score together with the attributes is that if the attributes are carefully chosen, unintended bias can be avoided. When the suitability estimation (i.e. facial image quality) is built on the deployed face recognition algorithm, unintended bias can happen. Training several attributes can avoid this bias both in the training and the datasets chosen. Additionally, we could consider handling non-categorical variables that cannot be regressed directly. For example, continuous attributes could be discretized into ordinal bins.

As a conclusion, this paper proposes a novel FIQ assessment approach, which adds explainability as FIQ annotation. The novelties of our algorithm are three-fold: First, we are the first to train a Multi-Task learning model considering several attributes that affect quality estimation of a face image. Second, we propose a new protocol to evaluate the traditional benchmarks such as LFW, but with a larger number of pairs and equal distribution of qualities. Third, an efficient implementation of multilatask learning model shows that it speeds up the label generation and has competitive inference times. Our proposed method combines regression and classification, allowing it to be retrained for different labels (e.g., from quality to another type of float) or classes (e.g., from gender to another binary classification). This adaptability makes it suitable for tasks like person re-identification by adjusting labels and classes accordingly.

REFERENCES

Luis Unzueta received the M.S. and Ph.D. degrees in mechanical engineering from Tecnun, University of Navarra, Donostia-San Sebastián, Spain, in 2002 and 2009, respectively. He is currently a Senior Researcher with the Intelligent Security Video Analytics Department of Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, Spain. His research interests include video-surveillance systems and human-computer interaction.

Zeno Geradts (Senior Member, IEEE) is currently a Senior Forensic Scientist with the Netherlands Forensic Institute, Netherlands. He is an expert witness in the area of forensic (video) image processing and biometrics. He was the President of the American Academy of Forensic Science from 2019 to 2020 and has been the Chairman of the ENFSI Forensic IT Working Group since 2018. From September first 2014, he is a Full Professor on Forensic Data Science by special appointment with the University of Amsterdam, Amsterdam, Netherlands. He was the recipient of several awards including the Distinguished Forensic Scientist Award from ENFSI in 2012.

Marcel Worring (Senior Member, IEEE) received the M.Sc. degree (Hons.) in computer science from Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, in 1988, and the Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, in 1993. He is currently a Full Professor within the Informatics Institute, University of Amsterdam. He has coauthored more than 200 scientific papers covering a broad range of topics. His research interests include integration of multimedia analysis, multimedia mining, information visualization, and multimedia interaction into a coherent framework yielding more than its constituent components. Dr. Worring was an Associate Editor for the IEEE Transactions on Multimedia, the ACM Transactions on Multimedia, and IEEE Multimedia.

Unai Elordi received the M.Sc. degree in computational engineering and intelligent systems in 2016 from the Basque Country University (EHU/UPV) Donostia-San Sebastián, Spain, where he is currently working toward the Ph.D. degree. He is currently a Researcher with the Intelligent Security Video Analytics Department of Fundación Vicomtech, Basque Research and Technology Alliance, Donostia-San Sebastián, Spain. His research interests include optimizing the deployment of deep neural networks in edge-to-cloud environments for video-surveillance applications.