# Going the Extra Mile in Face Image Quality Assessment: A Novel Database and Model

Shaolin Su[a,c], Hanhe Lin[b,c], Vlad Hosu[c], Oliver Wiedemann[c], Jinqiu Sun[a], Yu Zhu[a], Hantao Liu[d], Yanning Zhang[a] and Dietmar Saupe[c]

***Abstract*—An accurate computational model for image quality assessment (IQA) benefits many vision applications, such as image filtering, image processing, and image generation. Although the study of face images is an important subfield in computer vision research, the lack of face IQA data and models limits the precision of current IQA metrics on face image processing tasks such as face superresolution, face enhancement, and face editing. To narrow this gap, in this paper, we first introduce the largest annotated IQA database developed to date, which contains 20,000 human faces – an order of magnitude larger than all existing rated datasets of faces – of diverse individuals in highly varied circumstances. Based on the database, we further propose a novel deep learning model to accurately predict face image quality, which, for the first time, explores the use of generative priors for IQA. By taking advantage of rich statistics encoded in well pretrained off-the-shelf generative models, we obtain generative prior information and use it as latent references to facilitate blind IQA. The experimental results demonstrate both the value of the proposed dataset for face IQA and the superior performance of the proposed model.**

***Index Terms*—Image quality assessment, face quality, subjective study, GAN, generative priors**

## I. Introduction

The computer vision research on human faces includes the key area of multimedia processing. Since the human visual system (HVS) is especially sensitive to human faces [1], [2] in media content, dedicated processing tasks such as face superresolution, face enhancement, face generation and face editing have garnered growing interest over the past few decades. Although quality control in face image processing applications is a crucial factor determining user experience, the lack of face image quality metrics in the current research limits the precise measurement of these face-specific applications. Recently, blind image quality assessment (BIQA) approaches applied to broad-domain images have significantly improved; however, it is still unclear whether the methods are directly applicable to the face domain due to the following two factors: First, because of the specific processing mechanism dedicated to faces in the HVS [1], [3], [4], the perceptual representation and mapping pattern to face quality might be different from those to generically categorized images. Therefore, learning a dedicated face quality metric not only improves the quality prediction accuracy but also assists in understanding the perceptual mechanism of the HVS for human faces. Second, as existing IQA databases collect images of mostly generic categories, face image data are less often included. For example, only approximately 2% and 10% of images contain faces in two of the largest in-the-wild IQA datasets, KonIQ-10k [5] and SPAQ [6], respectively. As a result, either the image content shift or sample amount limits the ability of existing IQA models to draw correct quality mappings to face data.

Consequently, there is a need for IQA datasets that contain more subjectively rated face images to facilitate face IQA and processing. In this paper, we therefore introduce such a large-scale quality-annotated dataset and expect several applications to benefit from generic face image quality assessment (GFIQA). Examples of potential applications are as follows: 1. To improve the performance of face recognition, face images with quality scores below a predetermined threshold can be excluded during the acquisition phase, hence reducing the error rate of face recognition systems. 2. To improve general IQA model predictions, as the HVS is extremely sensitive to faces, the visual quality of face regions might be more critical in the perception process of the whole image; Therefore, an accurate face IQA metric could be advantageous for the general IQA task. 3. Other practical usages include album selection and optimization. When importing images to a photo album, face image quality can be used as a standard to determine acceptance or rejection.

Note that GFIQA differs from the definition of face image quality assessment in the biometrics community [7], [8], where quality is a form of utility for biometric systems such as identification of a face image. Recently, [9] also proposed a method to assess the visual quality of face images; however, they focused on GAN-generated face images, which address the quality assessment related to image synthesis models. Different from the above, for GFIQA, we aim to create predictive models (metrics) for in-the-wild face image quality assessment, where the quality relates to the degradation factors existing in the real world. The factors include the degradation introduced by an imaging system during capturing, processing, storage, compression, and display of face images [10], [11].

To compute accurate estimates of generic face IQA, we further proposed a novel model to fulfill the task. The recent successful use of deep generative priors in many image

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A05). Corresponding author: Hanhe Lin, E-mail: hlin001@dundee.ac.uk

[a]School of Computer Science and Engineering, Northwestern Polytechnical University, China.

[b]School of Science and Engineering, University of Dundee, DD1 4HN Dundee, United Kingdom.

[c]Department of Computer and Information Science, University of Konstanz, 78464 Konstanz, Germany.

[d]School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, United Kingdom.

restoration and editing tasks [12] has inspired us to explore the effectiveness of this powerful guidance in the field of IQA. In contrast to previous models using a vanilla encoder [13]–[15], we are the first to exploit deep generative priors in an image quality prediction model and develop an effective framework for utilizing these powerful priors in IQA. Rich statistics of natural images are encoded in pretrained generative models; by extracting intermediate generative features, we can utilize them as latent references corresponding to the distorted target images. The combination of distorted and latent reference features therefore facilitates the blind IQA task and allows for more accurate quality prediction results. Note that different from previous GAN-based IQA models [16], [17] that train their generators from scratch, the proposed model directly makes use of off-the-shelf GAN models to extract the prior information. The framework not only avoids the cumbersome training procedure of the generative models but also leverages the well pretrained GAN models on large-scale image data as an approximation to the natural image manifold, thus possessing more stability toward solving the challenging in-the-wild IQA problem.

The main contributions of this paper are as follows:

1) We created the largest IQA database of human faces in-the-wild, which is called the *Generic face image quality assessment 20k database* (GFIQA-20k). We collected 20,000 face images and ensured the diversity of the individuals, who are depicted in highly varied circumstances. We also validated the reliability of the collected dataset with gold-standard questions and self-consistency tests.
2) We proposed a novel quality prediction model that for the first time employs deep generative priors to facilitate the BIQA task. Using the rich statistics encoded in pretrained generative models, we obtain prior preserved images and use them as latent references to improve the IQA prediction accuracy.
3) The experimental results verified both the usefulness of the proposed dataset in evaluating face image quality and the effectiveness of the proposed model in achieving accurate predictions. The database and code will be made available at http://database.mmsp-kn.de/gfiqa-20k-database.html.

## II. Related Works

### A. Quality Assessment of Face Images

There are two main research areas that address the quality of face images. The first and most developed field stems from the biometrics community and aims at assessing face image quality for face recognition systems. This is most often referred to as face image quality assessment (FIQA). The second is GFIQA and relates to general image quality assessment dealing with perceptual image degradation. An in-depth discussion about the differences between the two fields has been presented by Schlett *et al.* [8].

FIQA has attracted increased attention in the face recognition community [7], [8]. Earlier works proposed measuring the quality of a face image in terms of its similarity to its reference face image with respect to multiple factors such as pose, expression, illumination, and occlusion. For example, Sellahewa and Jassim [18] measure image quality in terms of luminance distortion by comparing a face input image to a known reference image. However, such approaches are difficult to apply since they must consider every possible factor individually, and reference face images may not be available in an unconstrained environment.

In contrast, learning-based approaches, where the target face quality is defined in some manner to be indicative of face recognition performance, are more favorable. These learning-based approaches can be grouped according to the way the ground truth quality values are labeled. In most approaches, the ground truth quality values are determined computationally. For instance, Bharadwaj *et al.* [19] assigned qualities to face images by using two commercial off-the-shelf face recognition systems, where a face image is given a good quality value if it matches well. Chen *et al.* [20] assumed that the face images in dataset $A$ have better quality than those in dataset $B$ for a face recognition method if the recognition performance of this method on $A$ is better than that on $B$. Although there exists work that labels face quality manually, e.g., in binary classes (good or bad) [21], Best-Rowden and Jain [22] conducted the first subjective face quality assessment study. By conducting a study on a small set of pairwise comparisons of 13,233 face images taken from [23], the quality ratings of all images were inferred using matrix completion.

While FIQA is evolving significantly, no studies on GFIQA exist. To the best of our knowledge, except for the small number of face images present in existing IQA datasets, our work is the first study of this kind dedicated exclusively to face images.

### B. Generative Priors

With the rapid development of generative models, GANs have become capable of effectively learning the natural image manifold and synthesizing high-resolution images with pleasant visual quality [24]–[27]. With pretrained GAN models, the well-learned image manifold can be further explored to promote diverse image manipulation and restoration tasks [28]–[33], referred to as generative priors. To utilize the rich information encoded in generative models, target images are first mapped back to the intermediate features or latent space of pretrained GANs [12], and image manipulation or restoration tasks are then facilitated by feeding forward the inverted features or codes to generators.

There are typically two approaches to invert GANs and utilize the priors, optimization-based and learning-based. Optimization-based methods optimize the input code of the generator by minimizing the reconstruction error of the target image. By manipulating latent codes or modifying objective functions, image manipulation or restoration results can be obtained. Image2StyleGAN [29] and Image2StyleGAN++ [30] optimized latent codes in the $\mathcal{W}$ space of StyleGAN [26] and the $\mathcal{W}^+$ space of StyleGAN2 [27] and achieved image inpainting, morphing and style mixing results. mGANPrior [34] optimized multiple latent codes and adaptively fused them to achieve various image restoration results, including

colorization, superresolution, and denoising. Noticing the distribution gap between the training and testing data, DGP [31] proposed a method to fine-tune generator parameters on-the-fly to adapt the target images while maintaining the statistics of the GAN learned priors. Although they require no training procedures, optimization-based methods are usually time-consuming due to the large iteration numbers needed for each instance image.

Learning-based approaches train an extra encoder to map an image to its latent code. By modifying encoder architectures, various image manipulation and reconstruction results can be achieved. pSp [35] trained a multistage encoder to generate a series of style codes for StyleGAN2 [27] and handles various facial image translation tasks, including conditional image synthesis, facial frontalization, inpainting, *etc.*. GLEAN [36] proposed an encoder-generator-decoder design to fulfill the large-factor image superresolution task. GFP-GAN [32] and GPEN [37] fused target image features with generative prior features to restore real-world degraded face images. [33] warped and modulated generative prior features to achieve controllable image colorization results. Unlike optimization-based approaches, learning-based approaches obtain image restoration results by performing only one feed-forward pass at test time. However, extra data are usually needed to train these models.

In this paper, we investigate for the first time the potential application of generative priors to the task of IQA. Specifically, we employ rich statistics encoded by StyleGAN2 as latent references from the pristine image manifold to facilitate the solving of the blind IQA problem. To utilize generative priors efficiently and effectively, we propose a method to train a multistage encoder and take advantage of multilevel attributes controlled by the style codes to obtain the generative statistics. The proposed approach avoids both expensive optimization procedures and extra data training and shows its superiority in solving the objective face IQA problem.

Fig. 1. Example of duplicate image pairs or image pairs with the same identity, where the top images are face images in GFIQA-20k, and the bottom images are images excluded from GFIQA-20k using our sampling strategy.

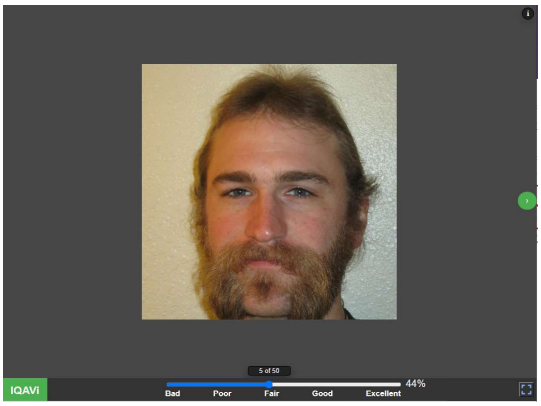



Fig. 2. UI for the subjective generic face IQA study. Each time participants were presented an image within a batch, they dragged a slider below the face to rate its visual quality on a scale ranging from Bad (1%) to Excellent (100%).

## III. The Creation of the GFIQA-20k Database

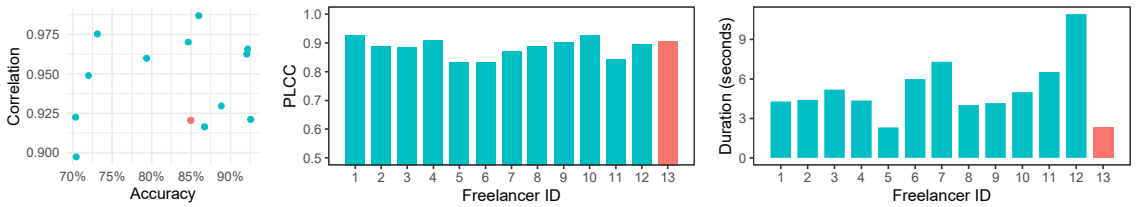
### A. Face Image Collection

Our goal is to build an ecologically valid face IQA database that includes a wide range of user-generated face images representing those in the real world and thus train a model that is of ecological validity.

To establish the dataset, we first collect face images from YFCC100M [38], a massive public multimedia database, to ensure diverse in-the-wild image qualities. We randomly selected and downloaded one million images, from which face images were extracted as follows. For a given image, we applied the MTCNN model [39] to detect faces and their corresponding key points, where the minimum size parameter of the face to detect was set as 400 pixels. Next, we aligned the image for each detected face according to the positions of the detected left and right eyes. The central point of a detected face was estimated to be the midpoint between the left and right eyes. Next, the detected face image was cropped such that both the width and height of the crop were equal to four times the distance between the left and right eyes. Finally, the crop was rescaled to $512\times512$ pixels. Using this procedure, we collected $86,026$ face images in the wild.

In practice, the face detection model cannot always guarantee the absence of incorrect face detection instances, such as false-positives (not human faces), or inaccurate key points. In light of this, we manually checked and removed incorrectly detected faces. This step reduced the number of samples to $53,058$.

In the final step, to ensure identity diversity, we removed duplicate identities from the collected face images. We used the FaceNet model [40] to extract the 512-dimensional deep features from the face images, which have been demonstrated to be effective in clustering face images into groups of people with the same identity. We next applied $k$-means clustering on the deep features to partition the $53,058$ images into 20,000 clusters. In this case, images in the same cluster will include duplicate images or face images with the same identity, as shown in Fig. 1. In each cluster, an image is randomly selected as a representative. With this step, the number of face images decreased to 20,000, which were the face images included in the GFIQA-20k. With our sampling strategy, duplicated face images or images with the same identity were removed, which guarantees identity diversity.




(a) (b) (c)

Fig. 3. Analysis of the subjective study. All participants submitted ratings for all images, except for the one shown in red, who submitted 153 batches (6,120 study images). (a) Reliability analysis of the subjective ratings. (b) Correlation between ratings of each individual freelancer and MOS. The PLCC values range from 0.832 to 0.928, all above the acceptance threshold of 0.75 PLCC, as recommended by [43]. (c) The average duration (second) of rating an image for each subject in our study. The average duration is 5.25 seconds.

### *B. Subjective Face Image Quality Assessment*

In the following, we performed a large-scale subjective study to assess the visual quality of 20,000 face images. The 20,000 images were randomly divided into 500 batches, where each batch initially contained 40 images. To better monitor and analyze participants' performance, two reliability mechanisms were used. One involved adding gold-standard or test data for which the correct answers are already known [41]. The other involved utilizing a consistency test by posing the same question multiple times [42]. In our study, we manually selected 100 high-quality and 100 low-quality face images as gold-standard images. Five images were randomly sampled (with replacement) from the 200 images and added to each batch. Moreover, five of the 40 study images were presented twice in each batch. Eventually, each batch contained 50 images to be rated.

Before carrying out the study, participants were first presented with a page of instructions containing four sections. In the first section, the definition of technical image quality was introduced. The hardware requirements and detailed study steps are explained in the second and third sections, respectively. In the final section, in addition to providing examples with different quality scales, we also provided some examples to differentiate between technical face image quality and face attractiveness. The user interface for our subjective face IQA study is illustrated in Fig. 2.

The standard 5-point absolute category rating (ACR) scale, i.e., Bad, Poor, Fair, Good, and Excellent, is used for subjective rating. Participants are presented with a batch of face images one at a time. Each time participants dragged a slider below the face image to rate its visual quality on a scale ranging from bad (1%) to excellent (100%). As participants were required to drag a slider on a scale, which we linearly mapped to the interval [0.01, 1]. To be more specific, let $x$ be the original 5-point ACR, and the mapped score is $y = (x-1)/4 \times 0.99 + 0.01$. As a result, the mapped 5-point ACR on the slider is Bad – 1%, Poor – 25.75%, Fair – 50.5%, Good – 75.25%, and Excellent – 100%.

To guide the freelancers on using the interface, we provided a training session for them. It contained 60 face images with given answers collected from the KonIQ-10k IQA database [5]. After giving a quality rating for an image, freelancers could click a button to proceed to the next image. However, if the assessment result was incorrect, they were informed of the incorrect assessment, and a range for the slider position was suggested. Freelancers could only proceed after having moved the slider into the suggested correct range. In the process of study, the duration of rating an image was recorded for further analysis.

A total of 13 freelancers were hired to participate in this study, 7 of whom are visual arts professionals such as designers, graphics artists, and photographers. More importantly, they all achieved excellent performance in a previous IQA contest of ours (not published), which demonstrated their expertise in IQA. One freelancer quit the study after submitting 153 batches, while the remaining freelancers completed the entire study.

### *C. Subjective study analysis*

We determined the reliability of the freelancers by measuring their accuracy on gold-standard test images and correlation on a self-consistency test. Before conducting the analysis, min-max normalization was applied for the rating of each subject. For a gold-standard test image, a freelancer's answer is counted as correct if his answer falls in the range of 1% to 35% when the image is labeled as low quality or in 65% to 100% when the image is labeled as high quality. For the self-consistency test, we used Spearman's rank correlation coefficient (SRCC) to calculate the reliability. The statistics of the freelancer reliability analysis are shown in Fig. 3(a). In Fig. 3(a), we plot the accuracy achieved on gold-standard images on the x-axis and the self-consistency on repeatedly presented images, expressed as the SRCC between the two scores provided for all images, on the y-axis. All participants achieved excellent self-consistency (mostly over 0.9 SRCC) and maintained a high level of accuracy relative to the gold-standard ratings (over 70% accuracy).

In addition to the reliability analysis, we report the Pearson linear correlation coefficient (PLCC) between individual ratings and MOS in Fig. 3(b). The ratings of each freelancer are highly correlated with the MOS, ranging from 0.832 to 0.928, all above the acceptance threshold of 0.75 PLCC, as recommended by [43]. The results demonstrate that all participants achieve high reliability regarding the subjective ratings, thus guaranteeing the reliability of the constructed dataset.


(a) (b)

Fig. 4. We divide the images into the categories of Excellent ((0.8,1]), Good ((0.6,0.8]), Fair ((0.4,0.6]), Poor ((0.2,0.4]), and Bad ([0,0.2]), where 10 images are randomly sampled from each category. (a) Sampled face images with their corresponding MOS (white digits) and categories in the GFIQA-20k dataset. (b) MOS distribution according to different categories.

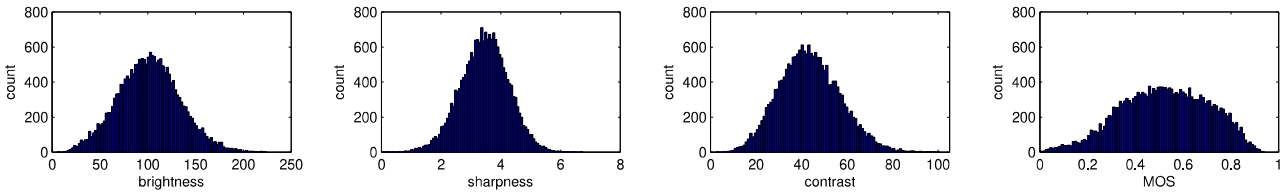



Fig. 5. We show the distributions of the dataset in regard to several aspects, including lighting, sharpness, contrast, and MOS distribution, to visualize the collected data attributes.

Fig. 3(c) shows the average duration of rating an image for each subject. The duration for each subject varies from 2.29 sec to 10.92 sec, with an average duration of 5.25 sec.

In the final step, to further improve the rater agreement, we perform outlier detection and removal procedures. We screen the ratings based on the assumption that the ratings provided by reliable participants lie in an interval around the mean of all the ratings of an image. To be more specific, the interval's length is twice the standard deviation of all ratings from an image; ratings outside the interval are removed, and the rest yield the mean opinion scores (MOSs). For each image, an average of 12 ACR ratings are obtained.

### D. Database overview

Finally, the collected MOSs with the corresponding 20,000 face images form the proposed GFIQA-20k dataset. In this subsection, we provide an overview and some analyses of the established dataset.

The database contains images with MOSs in the range of 0.005 to 0.941. In Fig. 4, we show image samples of different categories (Excellent, Good, Fair, Poor, and Bad) from the GFIQA-20k dataset. We sample 10 images from each category and plot their MOS distributions in Fig. 4 (b). As seen, the collected face images cover a diverse perceptual quality range, while images belonging to different categories are distinguishable from each other in the MOS distributions. In Fig. 5, we further show the distribution of the data in several dimensions, including brightness, sharpness, contrast and MOS distribution. In our calculation, brightness is estimated by the mean grayscale value, sharpness is calculated by taking the log of the image gradient magnitudes, and contrast can be calculated by $contrast(I) = std(I)/kurtosis(I)^{\frac{1}{4}}$, where $std(I)$ and $kurtosis(I)$ are the standard deviation and kurtosis of the image signal, respectively. It can be seen that the collected data is diverse in terms of different measurements, thus demonstrating the richness of the dataset.


(a) (b) (c) (d)

Fig. 6. We show some types of distortions that particularly affect face image quality in addition to common in-the-wild distortions. The distortions include occlusion/shadows on faces ((a) and (b)), underwater faces (c) and faces from old photos (d).

Although the proposed dataset collects data from the real world and thus contains distortions similar to those in the in-the-wild IQA dataset, we find that there exist other distortions that particularly affect face quality. We show some examples as well as their MOSs in Fig. 6; the distortions include occlusion/shadow on face regions, underwater faces and faces from old photos *etc.*. This indicates that the HVS perceives face quality differently from generic image quality and the need for constructing a specific face IQA dataset.

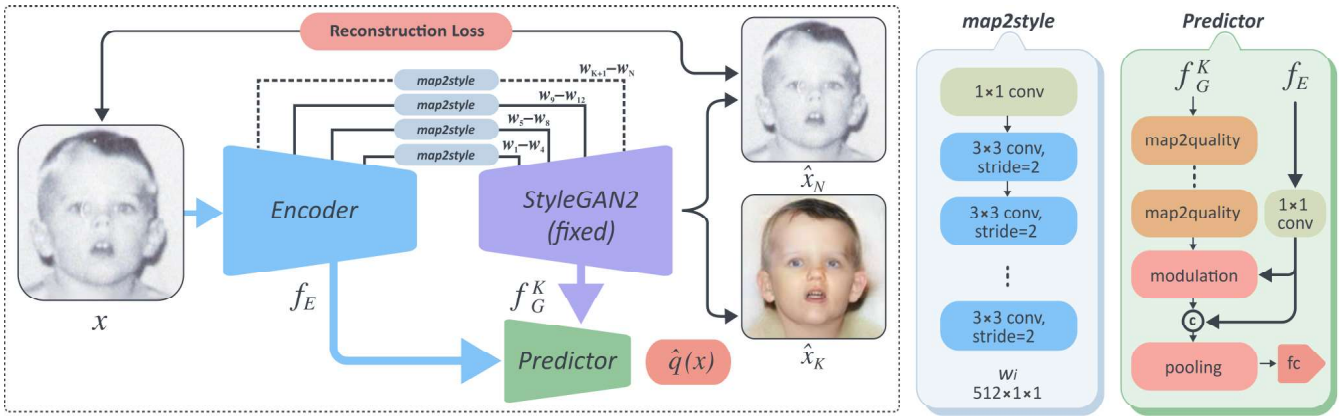



Fig. 7. The proposed face IQA model utilizing generative priors. Our framework consists of the following three parts: an encoder to both invert the target image and extract distortion features, a generator to produce latent reference features in a pretrained GAN space, and a predictor to make quality estimations by refining and fusing target image features and latent reference features.

## IV. Face Image Quality Assessment with Generative Priors

In this section, we provide a detailed description of the proposed objective face IQA model utilizing generative priors. As shown in Fig. 7, the overall framework consists of the following three parts: a multistage encoder to map the target image into the latent GAN space, a fixed pretrained GAN model to generate intermediate reference features, and a quality predictor to obtain objective face quality estimations by fusing both target image features and intermediate reference features. Compared with conventional IQA models, which mostly employ a single encoder architecture for quality score regression, utilizing generative priors in the proposed framework has two advantages. First, by restricting target image features to the GAN latent space, semantically meaningful and attribute-aware representations can be encoded. Second, by feeding the latent codes forward, intermediate GAN-encoded statistics can be obtained and used as latent references for the challenging no-reference quality prediction task.

### A. Obtaining GAN Encoded Statistics

Due to the lack of inference ability of the GAN model, we first invert the target image $x$ into $N$ latent codes $w_1, w_2, \ldots, w_N$ in the GAN input space. Specifically, we choose to train an encoder $E$ to map target images into the $\mathcal{W}^+$ space of StyleGAN2 [27], a state-of-the-art GAN model capable of generating diverse facial images with high resolution and visual quality. Similar to [35], we encode $N$ style codes $w_i \in \mathbb{R}^{512}$ from multiple stages of a ResNet50 [44] backbone, as follows:

$$(\mathbf{w}_N, f_E) = E(x; \theta_E), \quad \mathbf{w}_N = [w_1, w_2, \ldots, w_N] \tag{1}$$

where $f_E$ are intermediate features and $\theta_E$ are parameters of $E$. Latent codes $\mathbf{w}$ are then fed to the different scales of a fixed StyleGAN2 generator $G$ to produce a reconstructed result $\hat{x}_N$. During generation, we add $\mathbf{w}_N$ to the average latent code $\bar{\mathbf{w}} = [\bar{w_1}, \bar{w_2}, \ldots, \bar{w_N}]$ in the pretrained $\mathcal{W}^+$ space to achieve a good initialization, as follows:

$$\hat{x}_N = G(\mathbf{w}_N + \bar{\mathbf{w}}), \tag{2}$$

where $\hat{x}_N$ denotes the reconstructed result from $x$.

To train the encoder, we optimize $\theta_E$ over the reconstruction error between $\hat{x}_N$ and $x$, as follows:

$$\theta_E^* = \arg\min \mathcal{L}(\hat{x}_N, x), \tag{3}$$

where $\mathcal{L}$ denotes the loss function.

In this way, we train the encoder to map a target image $x$ into the GAN latent code space and obtain the intermediate features $f_E$ for quality prediction. However, to utilize rich generative priors, it is not enough to simply reconstruct the target image and extract the corresponding generative features. In the face IQA task, where target images are contaminated with distortions, the reconstructed results also contain degradation patterns and thus harm the GAN encoded statistics. To obtain facial statistics in the original GAN space, we take advantage of the interpretable and controllable attributes of the multiscale latent codes $\{w_i\}$. As latent codes at different scales are responsible for controlling level-specific facial attributes [26], [45], we observed that the low-level distortion attributes are inherently encoded in latent codes at finer scales, and the GAN statistics obtained in early stages are preserved. Therefore, during feed forward, we propose injecting the first $K, K < N$ codes into $G$ and discarding the last $N - K$ latent codes controlling the low-level details of $G$ to obtain generative representations $f_G^K$, which preserve the GAN encoded statistics.

$$(\hat{x}_K, f_G^K) = G(\mathbf{w}_K + \bar{\mathbf{w}}), \quad \mathbf{w}_K = [w_1, w_2, \ldots, w_K] \tag{4}$$

where $\hat{x}_K$ is the reconstructed image with only the first $K$ codes injected, and $f_G^K$ denotes the intermediate generative features.

In this framework, the first $K$ codes are mainly responsible for reconstructing high-level facial attributes such as facial contours and organ shapes resembling the target image, and generative statistics are preserved since distortion patterns

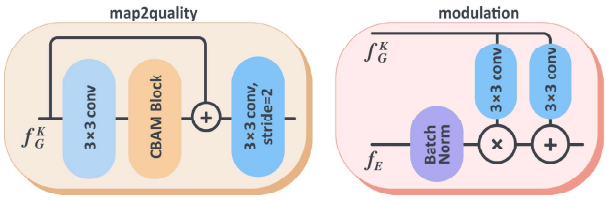



Fig. 8. The detailed architecture of map2quality and modulation blocks.

encoded in low-level codes are discarded. Notably, by directly discarding $N-K$ latent codes, we do not obtain reconstruction results that precisely match target images. However, in our IQA task, we do not need such perfect reconstruction results, and the results already obtain sufficient statistical priors to serve as latent references for solving the IQA problem. By further refining reference features to target image features, we combine them to make objective quality predictions.

### B. Quality Assessment with Generative References

After obtaining the target image features $f_E$ and the generative reference features $f_G^K$, we refine and fuse them for quality prediction. Specifically, we extract high-level representations $f_E \in \mathbb{R}^{2048\times16\times16}$ from the last stage of $E$ to avoid the need for another encoding process and $f_G^K \in \mathbb{R}^{32\times256\times256}$ from the last scale of $G$ since it contains most of the generative information. We then apply an $1\times1$ convolution to $f_E$ and a series of map2quality blocks to $f_G^K$ for feature refinement. As mentioned in Section IV-A, since the reconstructed structures do not perfectly match the target image, to refine the reference features, we use a CBAM [46] with a residual connection inside each block to gradually adjust the features and a $3\times3$ convolution with stride 2 and doubled channel numbers to resize the features, as shown in Figure 8. We then modulate $f_E$ by $f_G^K$, following the spatially adaptive denormalization operation proposed in [47], as follows:

$$f_{mod} = \gamma^{n,c,y,x}(f_G^K)\frac{f_E^{n,c,y,x} - \mu_E^c}{\sigma_E^c} + \beta^{n,c,y,x}(f_G^K), \quad (5)$$

where $\gamma^{n,c,y,x}(f_G^K)$ and $\beta^{n,c,y,x}(f_G^K)$ are elementwise modulation parameters after convolving $f_G^K$ with $3\times3$ kernels and $n, c, y, x$ are batch, channel and spatial indices, respectively. $\mu_E^c$ and $\sigma_E^c$ denote the channelwise mean and standard deviation values of $f_E$.

The operation modulates the distribution of target image features $f_E$ from its original distorted space to a generative reference space, thus serving refined reference features to the input. Finally, we concatenate $f_E$ with $f_{mod}$ and apply global average pooling followed by three fully connected layers to regress the features to the quality prediction score $\hat{q}(x)$.

### C. Objective Functions

We use three loss functions, *i.e.*, image reconstruction loss, regularization loss and quality prediction loss, to train our model. Image reconstruction loss ensures accurate GAN inversion results, which contain an $\mathcal{L}_2$ loss, a perceptual loss $\mathcal{L}_{\text{percep}}$ and a face identity loss $\mathcal{L}_{\text{ID}}$, represented as follows:

$$\mathcal{L}_2(x) = \|x - \hat{x}_N\|_2 \quad (6)$$

$$\mathcal{L}_{\text{percep}}(x) = \|f_{\text{percep}}(x) - f_{\text{percep}}(\hat{x}_N)\|_2 \quad (7)$$

$$\mathcal{L}_{\text{ID}}(x) = 1 - \langle R(x), R(\hat{x}_N)\rangle, \quad (8)$$

where $f_{\text{percep}}(\cdot)$ extracts perceptual features from a pretrained VGG [48] model, and $R(\cdot)$ extracts identity vectors from a pretrained ArcFace [49] model.

The regularization loss constrains encoder $E$ to output $\{w_i\}$ distributed within the latent generator space to avoid harming generative encoded statistics, as follows:

$$\mathcal{L}_{\text{reg}}(x) = \|\{w_i\} - \bar{\mathbf{w}}\|_2. \quad (9)$$

The quality prediction loss further optimizes the parameters in the predictor, and we calculate the $\mathcal{L}_1$ loss between the prediction result and subjective labels $q(x)$ as follows:

$$\mathcal{L}_{\text{q}}(x) = |q(x) - \hat{q}(x)|. \quad (10)$$

Finally, we sum the above loss functions with weights $\lambda_i, i = 1, 2, \ldots, 5$ and jointly train the proposed model as follows:

$$\begin{aligned}\mathcal{L}(x) =& \lambda_1\mathcal{L}_2(x) + \lambda_2\mathcal{L}_{\text{percep}}(x) + \lambda_3\mathcal{L}_{\text{ID}}(x) \\ &+ \lambda_4\mathcal{L}_{\text{reg}}(x) + \lambda_5\mathcal{L}_{\text{q}}(x).\end{aligned} \quad (11)$$

## V. Implementation Details

In Table I, we show the architectural details of the proposed model, including each module operation with its source input and output settings. The output size is shown in the order of $Channels \times Height \times Width$. It is worth noting that for a pretrained StyleGAN2 generating $512\times512$ resolution images, a total of 8 stages (1 stage without and 7 stages with upsampling) are included, and we combined every two stages in Table I for simplicity.

TABLE I
Detailed Architecture of Our Proposed Model.

| Module | Operation | Input | Output Size |
|---|---|---|---|
| Encoder | ResNet Stage1 | $3\times512\times512$ target image | $256\times128\times128$ |
| | ResNet Stage2 | ResNet Stage1 | $512\times64\times64$ |
| | ResNet Stage3 | ResNet Stage2 | $1024\times32\times32$ |
| | ResNet Stage4 | ResNet Stage3 | $2048\times16\times16$ |
| | map2style1 | ResNet Stage1 | $512\times1\times1$ |
| | map2style2 | ResNet Stage2 | $512\times1\times1$ |
| | map2style3 | ResNet Stage3 | $512\times1\times1$ |
| | map2style4 | ResNet Stage4 | $512\times1\times1$ |
| Generator | StyleGAN2 Stage1 | map2style4, $4\times4$ constant | $512\times8\times8$ |
| | StyleGAN2 Stage2 | map2style3, StyleGAN2 Stage1 | $512\times32\times32$ |
| | StyleGAN2 Stage3 | map2style2, StyleGAN2 Stage2 | $128\times128\times128$ |
| | StyleGAN2 Stage4 | map2style1, StyleGAN2 Stage3 | $32\times512\times512$ |
| Predictor | map2quality $\times$ 5 | StyleGAN2 Stage4 | $1024\times16\times16$ |
| | modulation | ResNet Stage4, map2style | $1024\times16\times16$ |
| | concat | ResNet Stage4, modulation | $2048\times16\times16$ |
| | global average pool | concat | 2048 |
| | fully connection1 | global average pool | 1024 |
| | fully connection2 | fully connection1 | 512 |
| | fully connection3 | fully connection2 | 1 |

We implemented our model with Pytorch, and StyleGAN2 is implemented based on its Pytorch version reimplementation. The pretrained StyleGAN2 model is taken from GFPGAN

TABLE II
PERFORMANCE COMPARISONS BY TRAINING ON PREVIOUS GENERIC IQA DATASETS WITH A SPECIFIED MODEL AND TESTED ON THE GFIQA-20K TEST SUBSET.

| Dataset/Model | SRCC↑ | PLCC↑ | RMSE↓ |
|---|---|---|---|
| LIVE/MEON | 0.6603 | 0.6371 | 0.1593 |
| LIVEC/HyperIQA | 0.7501 | 0.7314 | 0.1055 |
| KonIQ-10k/Koncept512 | 0.8968 | 0.8925 | 0.0826 |
| SPAQ/MT-A | 0.6980 | 0.7144 | 0.1282 |
| KonIQ++/BIQA | 0.9225 | 0.9196 | 0.0720 |

[32], where they provided the parameters for a $512 \times 512$ generator model. We selected $K = 12$ for our model. During training, the batch size is set to 16, and the learning rate is set to $5 \times 10^{-5}$ and then decayed by a factor of 10 every 10 epochs. During training, the StyleGAN2 decoder is fixed, and only the encoder and the predictor are optimized. We trained the model with the Adam optimizer [50] for a total of 25 epochs to report the final results. The whole model is trained using eight NVIDIA 1080Ti GPUs.

## VI. EXPERIMENTS

### A. Setup

We first randomly split the proposed GFIQA-20k dataset into a training subset (70%, 14,000 images), a validation subset (10%, 2,000 images) and a test subset (20%, 4,000 images). For testing, we selected the best performing model with the highest SRCC on the validation set for performance comparisons. We use the SRCC, Pearson Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE) to evaluate the model prediction accuracy and monotonicity.

### B. How Does the Generic IQA Perform on the FaceIQA Task?

To reveal the quality properties of face data, we first tested a cross database to observe how previous generic IQA datasets and models performed on the face IQA task. Specifically, we selected one synthetic IQA dataset LIVE [51] and four in-the-wild IQA datasets, including LIVE Challenge (LIVEC) [52], KonIQ-10k [5], SPAQ [6] and KonIQ++ [53], for cross testing. We trained IQA models MEON [54], HyperIQA [55], Koncept512 [5], MT-A [6], and BIQA [53] on the five datasets. Among the testing models, MEON [54] and HyperIQA [55] are state-of-the-art (SOTA) IQA methods that perform well on synthetic and authentic distortions, respectively, and the other models are proposed along with their training datasets. We tested the performance on the GFIQA-20k test subset and show the results in Table II.

From Table II, we observe that training on the synthetic IQA dataset LIVE did not give good predictions for in-the-wild face data. This result was foreseeable because of the domain gap between real world degradation and laboratory simulated distortions. The two authentic IQA datasets LIVEC and SPAQ also yielded relatively poor performances on the face data. This is probably because of the small number of training samples (1,162 images) contained in LIVEC and because of the bias in the smartphone photography images contained in SPAQ. Surprisingly, we found that KonIQ-10k and its extension KonIQ++ both performed relatively well (approximately 0.90 SRCC). The possible reason is that images from the KonIQ-10k dataset and the proposed GFIQA-20k dataset are both selected from YFCC100M [38], resulting in a smaller domain gap. Despite this, there is still room for further performance improvement, and the following section discusses the development of face IQA models.

### C. Performance Evaluation with Competing Models

In this subsection, we conduct performance comparisons of models trained with GFIQA-20k data. Due to the lack of baselines in the face IQA task, we first created diverse baseline models from different upstream tasks. Specifically, we selected ArcFace [49] pretrained on the refined face recognition dataset MS1M [56], Koncept512 [5] pretrained on the general IQA dataset KonIQ-10k [5], and ResNet50 [44] pretrained on the image classification dataset ImageNet [57]. We finetuned these models on the GFIQA-20k training subset and reported the results in Table III. As seen, by simple transfer learning, all the baseline models achieved high performance (over 0.95 SRCC). The results demonstrate the effectiveness of the collected data in handling face IQA tasks.

TABLE III
PERFORMANCE COMPARISONS OF TRANSFER LEARNING OF BASELINE MODELS, SOTA IQA MODELS AND THE PROPOSED MODEL.

| Model | SRCC↑ | PLCC↑ | RMSE↓ |
|---|---|---|---|
| ArcFace | 0.9505 | 0.9503 | 0.0588 |
| Koncept512 | 0.9520 | 0.9512 | 0.0572 |
| ResNet50 | 0.9629 | 0.9635 | 0.0504 |
| BRISQUE | 0.7824 | 0.8055 | 0.1793 |
| CORNIA | 0.8547 | 0.8616 | 0.1001 |
| HOSA | 0.8861 | 0.8997 | 0.0945 |
| PQR | 0.9519 | 0.9534 | 0.0551 |
| DBCNN | 0.9609 | 0.9611 | 0.0520 |
| HyperIQA | 0.9627 | 0.9635 | 0.0505 |
| MUSIQ | 0.9630 | 0.9637 | 0.0503 |
| TRes | 0.9632 | 0.9638 | 0.0498 |
| Proposed | **0.9643** | **0.9652** | **0.0486** |

We further compared the proposed model with the following eight general IQA models: BRISQUE [59], CORNIA [60], HOSA [61], PQR [62], DBCNN [63], HyperIQA [55], MUSIQ [58] and TRes [64]. Here, BRISQUE, CORNIA and HOSA are traditional IQA methods, PQR, DBCNN and HyperIQA are CNN-based deep learning models, and MUSIQ and TRes are transformer-based deep learning models. The selected deep learning models are all SOTA in-the-wild IQA models. Except for in the case of the traditional models, the training and testing runs are all repeated 10 times with random weight initialization for competing deep learning models, and the median results are reported in Table III. Among all the competing models, the proposed model outperformed the others on all three criteria. Statistical analysis also demonstrates the superior performance of the proposed model: by conducting a Student's t-test, the $p$ values between the proposed model and MUSIQ [58] are 0.0100 for SRCC and 0.0032 for PLCC, and 0.0176 for SRCC and 0.0081 for PLCC against TRes [64], where $p < 0.05$ indicates statistical significance.

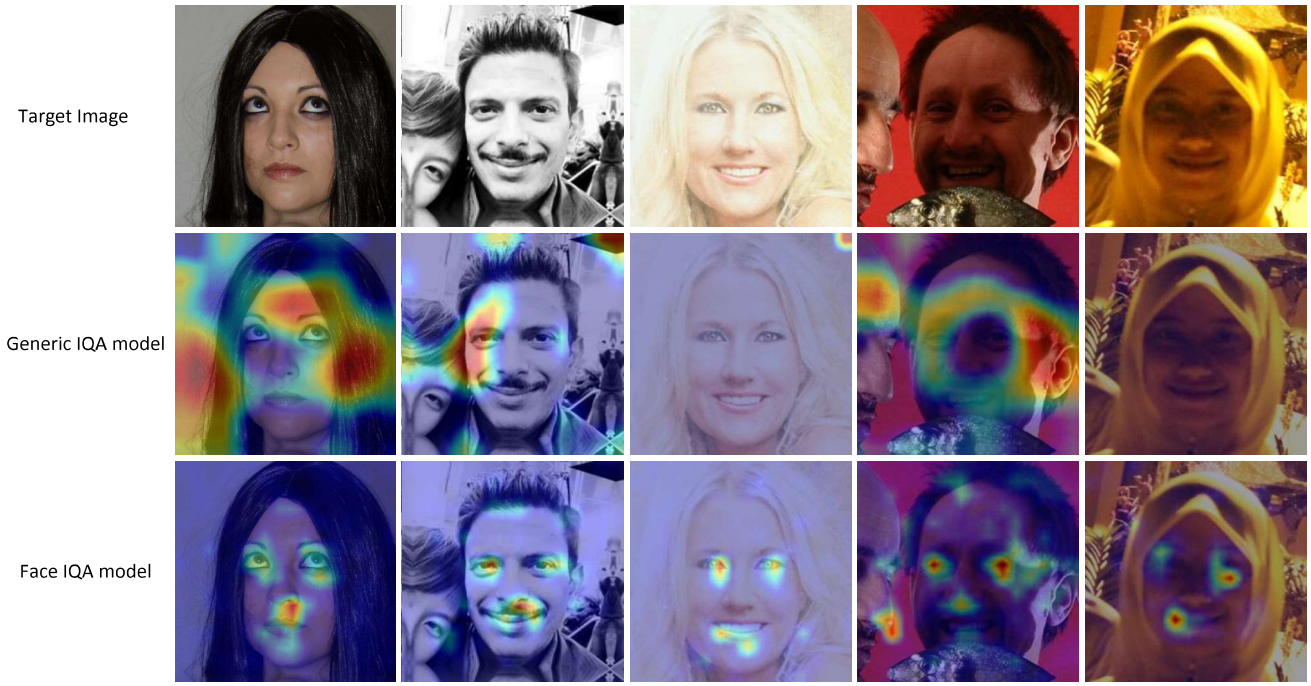



Fig. 9. We compare heatmaps of the generic IQA model [53] and face IQA model to visualize their perceptual differences regarding faces. It is clearly shown that the generic IQA model does not make quality predictions from faces, while the face model learns to consistently concentrate on perceptual critical regions such as eyes, noses and mouths to determine a face image's quality, regardless of various distortions.

TABLE IV
PERFORMANCE COMPARISONS FOR DIFFERENT TRAINING SAMPLE AMOUNTS.

| Criterion | Model | 10% | 20% | 30% | 40% | 50% | 60% | 70% |
|---|---|---|---|---|---|---|---|---|
| SRCC↑ | ResNet50 | 0.9480 | 0.9542 | 0.9581 | 0.9612 | 0.9626 | 0.9628 | 0.9629 |
| | Proposed | **0.9484** | **0.9565** | **0.9609** | **0.9625** | **0.9632** | **0.9638** | **0.9639** |
| PLCC↑ | ResNet50 | 0.9474 | 0.9537 | 0.9586 | 0.9618 | 0.9625 | 0.9632 | 0.9635 |
| | Proposed | **0.9478** | **0.9570** | **0.9608** | **0.9623** | **0.9635** | **0.9643** | **0.9644** |
| RMSE↓ | ResNet50 | 0.0603 | 0.0550 | 0.0524 | 0.0521 | 0.0514 | 0.0507 | 0.0504 |
| | Proposed | **0.0586** | **0.0538** | **0.0514** | **0.0503** | **0.0501** | **0.0489** | **0.0489** |

To further validate the effectiveness of the proposed model, in Table IV, we evaluated how the model performed with different training sample amounts. We compared the proposed model with the well-performing ResNet50 baseline and varied the training sample size from 10% to 70% of the images in the GFIQA-20k dataset, leaving the remaining images, except for the validation subset, for testing. Similarly, the proposed model showed consistently superior prediction accuracy for variable training sample sizes.

### *D. Perceptual Comparison between Generic IQA and FaceIQA Data*

To understand the perceptual mechanism of deep models on face images and to reveal their difference from generic IQA, in this subsection, we visualize how models make their predictions when trained for generic IQA and for the GFIQA task. Specifically, we select the BIQA model from [53] trained on the generic IQA dataset KonIQ++ and the proposed model trained on the GFIQA-20k dataset for comparison and draw their heatmaps [65] to understand how they perceive the face image quality. We show the results in Figure 9 and make the following observations. First, although the generic IQA model [53] performed relatively well on the GFIQA task in Table II, the predictions are not correctly made from the regions of the faces. It extracts features from different regions for different faces and thus might not be robust to diverse images. Second, as a comparison, when trained on the GFIQA-20k dataset, the model learns to capture critical face regions as quality representations, including eyes, noses and mouths. The result indicates the value and importance of the constructed dataset, which converges deep models on critical perceptual regions for the GFIQA task. Third, it is also interesting to find that although not explicitly constrained, the model learns on its own to consistently focus on the fixed regions when making quality predictions. We attribute the phenomenon to the subjective MOS being intrinsically perceptually biased toward these face organs, resulting in the model being optimized on these regions. The hypothesis further assists us in understanding some perceptual HVS mechanisms regarding faces. Since the deep model fits its perceptual mapping to the HVS, by examining how the model perceives face quality,

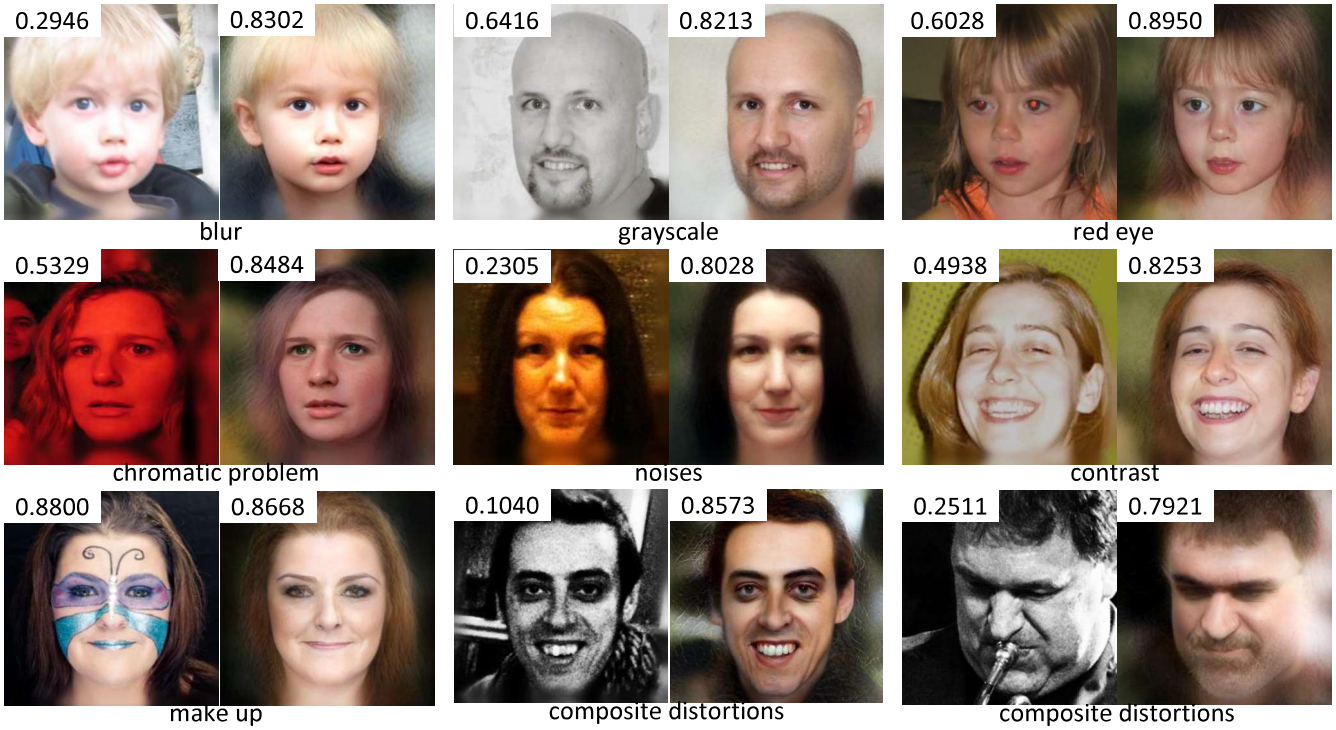



Fig. 10. We visualize some generated latent reference images (right) with respect to their distorted images (left). Thanks to the powerful generative prior information, latent reference images are constructed despite various distortions contained in inputs and are able to facilitate the implementation of our quality prediction task. For each image, we also show the predicted quality score by MUSIQ [58] for comparison.

we also assume that the HVS perceives face quality prior to perceiving facial organs. The finding is not only consistent with the neuroscience study that refrontal neurons in the HVS are selectively biased to recognize face identity [2] but also unravels the precise and concrete regions to which the neuron cortex is particularly sensitive.

### E. Visualizing Generative References

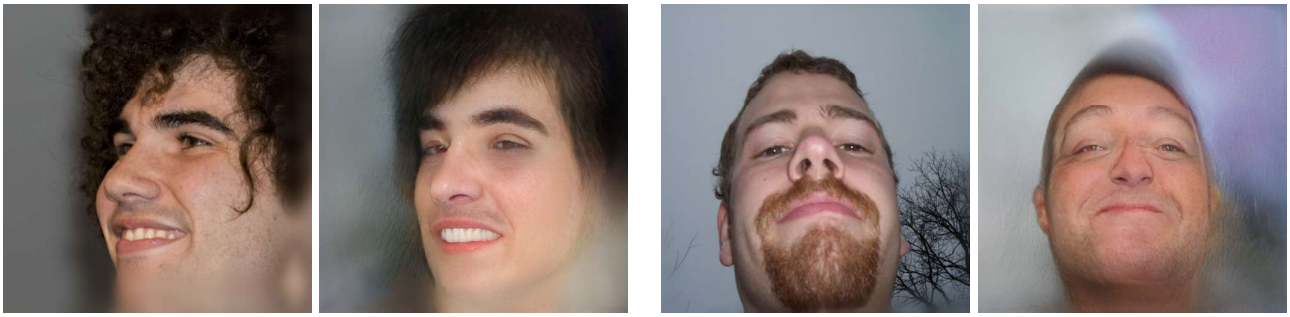
One of the benefits facilitated by utilizing generative priors is producing latent reference face images with preserved GAN statistics. In this subsection, we visualize the reconstructed reference images to illustrate the effectiveness. In Fig. 10, we show pairs of distorted images $x$ and the reconstructed latent references $\hat{x}(K)$, as well as quality scores predicted by MUSIQ [58] for both images. We include various in-the-wild distortions, including blur, color, contrast, noise, and composite distortions. Thanks to the rich prior information encoded in generative models, the reconstructed images are of high quality, *i.e.*, mostly approximately 0.8-0.85 by MUSIQ scores, thus serving as latent references to the blind face IQA task. Since we impose loss constraints mainly on face regions, the generated latent reference images might not precisely match the target images in the hair or background regions. However, since the HVS is extremely sensitive to face regions such as eyes and mouths, the difference in hair or background regions does not contribute critically to the perceptual quality, which is also shown in the predicted reference image scores.

### F. Ablation Study

In this subsection, we conducted several ablation experiments to evaluate the effectiveness of the model design. We first compared our model configured with the baseline encoder ResNet50 trained by only $\mathcal{L}_{\mathrm{q}}$ loss. We then added StyleGAN2 and reconstruction loss to the model but did not fuse the latent reference features (w/o ref) to observe if encoding in generative latent space benefits model performance. Next, we evaluated how different values of $K$ affected the performance. We selected $K = 4, 8, 12, 16$ while keeping the other components fixed. Finally, we validated the designation of the quality predictor. We substituted the map2quality module with ordinary convolution blocks (w/o map) while the other parts remained fixed. We also removed the modulation block and simply concatenated $F_E$ and $f_G^K$ (w/o mod) to observe the effectiveness of the feature modulation block. The results are shown in Table V.

TABLE V
ABLATION STUDIES ON DIFFERENT MODEL CONFIGURATIONS.

| model | SRCC↑ | PLCC↑ | RMSE↓ |
|---|---|---|---|
| baseline | 0.9629 | 0.9635 | 0.0504 |
| w/o ref | 0.9629 | 0.9636 | 0.0504 |
| K=4 | 0.9624 | 0.9630 | 0.0504 |
| K=8 | 0.9627 | 0.9635 | 0.0505 |
| K=12 | 0.9639 | 0.9644 | 0.0489 |
| K=16 | 0.9637 | 0.9644 | 0.0492 |
| w/o map | 0.9631 | 0.9635 | 0.0503 |
| w/o mod | 0.9634 | 0.9640 | 0.0495 |
| full | **0.9643** | **0.9652** | **0.0486** |



(a) MOS = 0.752, prediction = 0.647 (b) MOS = 0.705, prediction = 0.571

Fig. 11. We show some image cases where quality prediction errors are larger than 0.1. The faces (left) are viewed at rotated angles, leading to inaccurate generative references (right) and, thus, inaccurate quality predictions. Including images with more diverse rotation angles to train more powerful generative priors might be a solution.

From Table V, we make several observations. First, although not evident, encoding in generative latent space (w/o ref) slightly improved the model performance. Second, when extracting latent reference features from earlier generator stages (small $K$ values), the model showed inferior performance compared with the baseline. This is probably because, in the early stages, the generator is not able to encode enough statistics as a reference. However, when we extracted features from the latter stages, they outperformed the baseline model. Third, removing the map2quality or modulation module reduced the model performance, indicating the effectiveness of the proposed architecture of the quality predictor.

### G. Complexity Analysis

In this subsection, we compared the complexity of our model with that of two models, HyperIQA [55] and MUSIQ [58], in terms of computation complexity (FLOPs) and running time. The results are shown in Table VI.

TABLE VI
COMPLEXITY COMPARISONS BETWEEN THE PROPOSED MODEL AND OTHER MODELS.

| Model | FLOPs(G) | Time(s) |
|---|---|---|
| HyperIQA | 108.38 | 0.092 |
| MUSIQ | 72.78 | 0.068 |
| Proposed | 240.75 | 0.205 |

Although the computation complexity and running time of the proposed model are slightly larger than those of the two competitors due to the extra generative model, it achieves better prediction accuracy. Compared to other models, another benefit of introducing generative priors is that a totally training-free IQA model could be developed, as we will explain in Section VI-H.

### H. Developing a Total Training-free GFIQA Metric

As shown in Section VI-E, thanks to generative priors, the model is able to produce images that are distributed close to the pristine image space as latent references. In this subsection, we further asked, with the latent references, are we able to develop a training-free GFIQA metric that fulfills the NR-IQA task without any training requirements? To answer this question, we modified the proposed model to a training-free model (proposed-TF) and evaluated its performance on the GFIQA-20k test set. Specifically, we extracted the generated image as a reference and calculated its LPIPS [66] distance to the target image as the quality measurement. Our underlying hypothesis is that since the generated images can serve as high-quality references, why not directly use FR-IQA models for quality prediction? Since both the StyleGAN2 and LPIPS models are off-the-shelf models, following the proposed framework, we are able to avoid the extra training process.

TABLE VII
PERFORMANCE COMPARISON OF THE PROPOSED TRAINING-FREE GFIQA METRIC WITH OTHER OPINION-UNAWARE IQA METRICS.

| Model | SRCC↑ | PLCC↑ | RMSE↓ |
|---|---|---|---|
| NIQE | 0.5549 | 0.5612 | 0.5824 |
| IL-NIQE | 0.5806 | 0.5830 | 0.5538 |
| LPSI | 0.2160 | 0.2504 | 0.9229 |
| QAC | 0.2844 | 0.2932 | 0.8483 |
| dipIQ | 0.4582 | 0.4847 | 0.6470 |
| RankIQA | 0.5262 | 0.5435 | 0.4718 |
| Proposed-TF | **0.7012** | **0.7236** | **0.3407** |

We compare the proposed training-free metric with other opinion-unaware IQA models, including NIQE [67], IL-NIQE [68], LPSI [69], QAC [70], dipIQ [71] and RankIQA [72]. Among the compared methods, NIQE [67], IL-NIQE [68] and LPSI [69] are totally blind IQA estimators, and the rest are trained on pseudo image quality labels. We show the results in Table VII, and we make the following observations. First, as shown, the proposed metric outperformed all the competitors by a large margin on the GFIQA-20k test set. Since the competing models mainly focus on the synthetic IQA task, the proposed model showed its superior effectiveness on the more challenging in-the-wild IQA task. Second, we also found that the two deep learning-based models dipIQ [71] and RankIQA [72] actually performed worse than the totally training-free methods NIQE [67] and IL-NIQE [68]. The result indicates that the potential risk of overfitting exists in the two deep learning-based models, which are trained with images containing synthetic distortions. Third, compared with the training-based model, the training-free framework

(a) (b) (c) (d)

Fig. 12. We show face examples with various technical qualities and attractiveness. (a) High quality (MOS = 0.910) and high attractiveness, (b) High quality (MOS = 0.901) but low attractiveness, (c) Low quality (MOS = 0.130) but high attractiveness, and (d) Low quality (MOS = 0.138) and low attractiveness.

still performed worse. This is probably due to the discrepant perceptual mapping learned by LPIPS to the face perceptual domain and the imperfect generative references of some flawed face reconstruction cases, as will be explained in Section. VII. Nevertheless, training-free models are commonly agreed to be more robust to unseen data [73] [74]; thus, we expect the proposed training-free IQA framework to be applied to challenging real-world IQA applications.

## VII. Discussion

Although the proposed model showed its superiority in face quality prediction, we find some limitations regarding the generative priors. By selecting test images with prediction errors greater than 0.1 (10% of the quality score range), we find that images with rotated faces usually lead to unsatisfactory reconstructed results and poorer quality predictions, as shown in Fig. 11. We assume this is because the generative prior model StyleGAN2 was mostly trained with frontally viewed samples while few images of rotated faces were included. Thus, the generator can produce frontally viewed faces but underperforms otherwise. To address this issue, training the generative model with faces viewed from diverse view angles to provide more powerful priors might be a solution, and we leave the task for future work. It is also worth noting that in our proposed framework, the prior model could be substituted by others; therefore, with the development of more powerful generative models, the proposed IQA model should also benefit and perform better, which we expect to observe in the future.

We also clarify that technical face image quality should not be confused with face attractiveness, *i.e.*, aesthetic. In the subjective study, apart from giving a precise definition of technical image quality in the instruction, we provided a few examples to teach freelancers how to differentiate them. Fig. 12 shows some face examples in GFIQA-20k with various qualities and attractiveness. It shows the MOSes are consistent with technical quality rather than attractiveness, which demonstrates the effect of face attractiveness was minimized in the proposed dataset.

## VIII. Conclusion

We created GFIQA-20k, the largest annotated in-the-wild database for face image quality prediction. The dataset contains 20,000 faces of diverse individuals in various circumstances. Furthermore, to accurately predict face image quality, we introduce generative prior information to the IQA task for the first time. The proposed model makes use of rich statistics encoded in pretrained deep generative models. Our experiments validated its superiority relative to existing works. We expect both the dataset and the model will be valuable in face processing and general IQA research.

## References


[1] D. I. Perrett, A. J. Mistlin, and A. J. Chitty, "Visual neurones responsive to faces," *Trends in Neurosciences*, vol. 10, no. 9, pp. 358–364, 1987.

[2] S. P. O Scalaidhe, F. A. Wilson, and P. S. Goldman-Rakic, "Areal segregation of face-processing neurons in prefrontal cortex," *Science*, vol. 278, no. 5340, pp. 1135–1138, 1997.

[3] T. Ro, C. Russell, and N. Lavie, "Changing faces: A detection advantage in the flicker paradigm," *Psychological Science*, vol. 12, no. 1, pp. 94–99, 2001.

[4] J. Theeuwes and S. Van der Stigchel, "Faces capture attention: Evidence from inhibition of return," *Visual Cognition*, vol. 13, no. 6, pp. 657–665, 2006.

[5] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[6] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3677–3686.

[7] P. Grother, A. Hom, M. Ngan, and K. Hanaoka, "Ongoing face recognition vendor test (FRVT) part 5: Face image quality assessment," *NIST Interagency Report*, 2020.

[8] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: A literature survey," *ACM Computing Surveys*, 2021.

[9] Y. Tian, Z. Ni, B. Chen, S. Wang, H. Wang, and S. Kwong, "Generalized visual quality assessment of gan-generated face images," *arXiv preprint arXiv:2201.11975*, 2022.

[10] L. Liu, T. Wang, and H. Huang, "Pre-attention and spatial dependency driven no-reference image quality assessment," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2305–2318, 2019.

[11] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "GraphIQA: Learning distortion graph representations for blind image quality assessment," *IEEE Transactions on Multimedia*, 2022.

[12] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "GAN inversion: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3121–3138, 2023.

[13] X. Yang, F. Li, and H. Liu, "TTL-IQA: Transitive transfer learning based no-reference image quality assessment," *IEEE Transactions on Multimedia*, vol. 23, pp. 4326–4340, 2020.

[14] C. Yang, X. Zhang, P. An, L. Shen, and C.-C. J. Kuo, "Blind image quality assessment based on multi-scale klt," *IEEE Transactions on Multimedia*, vol. 23, pp. 1557–1566, 2020.

[15] F.-Z. Ou, Y.-G. Wang, J. Li, G. Zhu, and S. Kwong, "A novel rank learning based no-reference image quality assessment method," *IEEE Transactions on Multimedia*, 2021.

[16] K.-Y. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 732–741.

[17] H. Ko, D. Y. Lee, S. Cho, and A. C. Bovik, "Quality prediction on deep generative images," *IEEE Transactions on Image Processing*, vol. 29, pp. 5964–5979, 2020.

[18] H. Sellahewa and S. A. Jassim, "Image-quality-based adaptive face recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 805–813, 2010.

[19] S. Bharadwaj, M. Vatsa, and R. Singh, "Can holistic representations be used for face biometric quality assessment?" in *IEEE International Conference on Image Processing*, 2013, pp. 2792–2796.

[20] J. Chen, Y. Deng, G. Bai, and G. Su, "Face image quality assessment based on learning to rank," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 90–94, 2014.

[21] X. Zhao, Y. Li, and S. Wang, "Face quality assessment via semi-supervised learning," in *International Conference on Computing and Pattern Recognition*, 2019, pp. 288–293.

[22] L. Best-Rowden and A. K. Jain, "Learning face image quality from human assessments," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 12, pp. 3064–3077, 2018.

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

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[25] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2018.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[28] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[29] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the stylegan latent space?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4432–4441.

[30] ——, "Image2StyleGAN++: How to edit the embedded images?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8296–8305.

[31] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, "Exploiting deep generative prior for versatile image restoration and manipulation," in *European Conference on Computer Vision*. Springer, 2020, pp. 262–277.

[32] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9168–9178.

[33] Y. Wu, X. Wang, Y. Li, H. Zhang, X. Zhao, and Y. Shan, "Towards vivid and diverse image colorization with generative color prior," in *IEEE International Conference on Computer Vision*, 2021, pp. 14 377–14 386.

[34] J. Gu, Y. Shen, and B. Zhou, "Image processing using multi-code GAN prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3012–3021.

[35] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A StyleGAN encoder for image-to-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2287–2296.

[36] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for large-factor image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 245–14 254.

[37] T. Yang, P. Ren, X. Xie, and L. Zhang, "GAN prior embedded network for blind face restoration in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 672–681.

[38] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[39] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[41] J. Le, A. Edmonds, V. Hester, and L. Biewald, "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution," in *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, vol. 2126, 2010, pp. 22–32.

[42] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *IEEE International Symposium on Multimedia*, 2011, pp. 494–499.

[43] ITUT, "Recommendation itut p.913 (06/2021) methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," p. 52, 2021.

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

[45] Y. Alharbi and P. Wonka, "Disentangled image generation through structured noise injection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5134–5142.

[46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision*. Springer, 2018, pp. 3–19.

[47] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[49] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[52] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[53] S. Su, V. Hosu, H. Lin, Y. Zhang, and D. Saupe, "KonIQ++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects," *The 32nd British Machine Vision Conference*, 2021.

[54] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.

[55] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[56] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 87–102.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[58] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *IEEE International Conference on Computer Vision*, 2021, pp. 5148–5157.

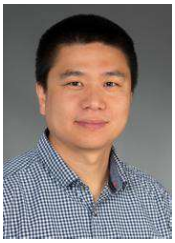
[59] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[60] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[61] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[62] H. Zeng, L. Zhang, and A. C. Bovik, "A probabilistic quality representation approach to deep blind image quality prediction," *arXiv preprint arXiv:1708.08190*, 2017.
[63] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.
[64] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 1220–1230.
[65] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
[67] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.
[68] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.
[69] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *IEEE International Conference on Image Processing*. IEEE, 2015, pp. 339–343.
[70] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 995–1002.
[71] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, 2017.
[72] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *IEEE International Conference on Computer Vision*, 2017, pp. 1040–1049.
[73] P. Kancharla and S. S. Channappayya, "Completely blind quality assessment of user generated video content," *IEEE Transactions on Image Processing*, vol. 31, pp. 263–274, 2021.
[74] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

**Shaolin Su** received his M.S. degree from Harbin Institute of Technology, China in 2014 and Ph.D degree from Northwestern Polytechnical University, China in 2023. He is currently a visiting researcher at the University of Konstanz, Germany. His research interests include image quality assessment, image restoration and image synthesis.



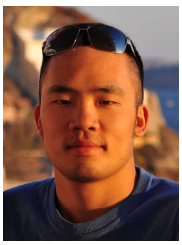
**Hanhe Lin** received his Ph.D. at the Department of Information Science, University of Otago, New Zealand in 2016. From 2016 to 2021, he was a postdoc at the Department of Computer and Information Science at the University of Konstanz, Germany, where he was working on project A05 (visual quality assessment) of SFB-TRR 161, funded by the German Research Foundation (DFG). Currently, he is Lecturer in Computing at University of Dundee, UK. His research interests include image processing, computer vision, machine learning, deep learning, and visual quality assessment. He serves as a member of technical program committee or a reviewer in a number of conferences/journals, e.g. QoMEX, IEEE TPAMI/TIP.

**Vlad Hosu** received a Ph.D. in Computer Vision from the National University of Singapore, with his dissertation on computational aesthetics in photography. Currently, he explores the intersection of technical and aesthetic quality assessment by studying human and machine visual perception. Vlad is developing reliable visual quality models using deep learning and crowdsourcing. His contributions include co-authoring several databases central to the field and significantly improving state-of-the-art predictive models, leading to many applications.

**Oliver Wiedemann** received his B.Sc. and M.Sc. in Computer and Information Science from the University of Konstanz, Germany in 2018 and 2020. He is currently a Ph.D. student working in the field of image quality assessment. His focus is on no-reference metrics in the context of scale-invariance, processing of high-resolution images and unsupervised representation learning.

**Jinqiu Sun** received her B.S. degree from Northwestern Polytechnical University in 1999, M.S. and Ph.D. Degree from Northwestern Polytechnical University in 2004 and 2005 respectively. She is presently a Professor of School of astronomy, Northwestern Polytechnical University. Her research work focuses on signal and image processing, computer vision and pattern recognition.

**Yu Zhu** received the Ph. D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2017. Currently he is an assistant researcher in School of Computer Science, NPU. His current research interests include image debluring, Image enhancement and image super-resolution.



**Hantao Liu** received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and IEEE Signal Processing Letters.

**Yanning Zhang** received the B.S. degree from the Dalian University of Technology, Dalian, China, in 1988, the M.S. degree from the School of Electronic Engineering, Northwestern Polytechnical University, Xi'an, China, in 1993, and the Ph.D. degree from the School of Marine Engineering, Northwestern Polytechnical University, Xian, China, in 1996. She is currently a Professor with the School of Computer Science, Northwestern Polytechnical University. She is also a Cheung Kong Professor of Ministry of Education, China. Her current research interests include remote sensing image analysis, computer vision and pattern recognition and etc.

**Dietmar Saupe** was born in Bremen, Germany, in 1954. He received the Dr.rer.nat. degree in mathematics from the University of Bremen, Germany, in 1982. From 1985 to 1993, he was an Assistant Professor with the Departments of Mathematics, first at the University of California, Santa Cruz, USA, and then at the University of Bremen, resulting in his habilitation. From 1993 to 1998, he was a Professor of computer science with the University of Freiburg, Germany, the University of Leipzig, Germany, until 2002, and since then, the University of Konstanz, Germany. He is the coauthor of the book Chaos and Fractals (Springer-Verlag, 1992), which won the Association of American Publishers Award for Best Mathematics Book of the Year, the book The Science of Fractal Images (Springer-Verlag, 1988), and well over 100 research articles. His research interests include image and video processing, computer graphics, scientific visualisation, dynamical systems, and sport informatics.