

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

Shaolin Su^a, 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—Computer vision models for image quality assessment (IQA) predict the subjective effect of generic image degradation, such as artefacts, blurs, bad exposure, or colors. The scarcity of face images in existing IQA datasets (below 10%) is limiting the precision of IQA required for accurately filtering low-quality face images or guiding CV models for face image processing, such as super-resolution, image enhancement, and generation. In this paper, we first introduce the largest annotated IQA database to date that contains 20,000 human faces (an order of magnitude larger than all existing rated datasets of faces), of diverse individuals, in highly varied circumstances, quality levels, and distortion types. Based on the database, we further propose a novel deep learning model, which re-purposes generative prior features for predicting subjective face quality. By exploiting rich statistics encoded in well-trained generative models, we obtain generative prior information of the images and serve them as latent references to facilitate the blind IQA task. Experimental results demonstrate the superior prediction accuracy of the proposed model on the face IQA task.

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

I. INTRODUCTION

WHILE the performance of blind image quality assessment (BIQA) methods on broad-domain user-generated images (authentic distortions) has improved significantly in recent years with in-the-wild datasets such as KonIQ-10k [1] and SPAQ [2], their performance on face images has not been fully explored yet. We assume that the performance of existing IQA methods on images of faces is limited due to the following factors:

- 1) Existing IQA databases contain few images of faces, for instance, two of the largest datasets, KonIQ-10k [1], and SPAQ [2] contain around 2% and 10% faces, respectively.
- 2) People are more sensitive to image degradations that affect faces than general image categories, due to the specialized processing dedicated to faces in the human visual system [3], [4].

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A05) and DFG Project-ID 496858717. Corresponding author: Hanhe Lin, E-mail: hlin001@dundee.ac.uk

^aSchool of Computer Science and Engineering, Northwestern Polytechnical University, China.

^bSchool of Science and Engineering, University of Dundee, DD1 4HN Dundee, United Kingdom.

^cDepartment of Computer and Information Science, University of Konstanz, 78464 Konstanz, Germany.

^dSchool of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, United Kingdom.

- 3) There is inherent perceptual feature shift when training IQA models on general image categories and testing on faces.

Consequently, there is a need for IQA datasets that contain more subjectively rated face images. Note that our IQA on human faces is different from face image quality assessment in biometrics community [5], [6], where quality is a form of utility for biometric systems, i.e., recognition or identification of a face image. We aim to create predictive models (metrics) for generic face image quality assessment, where the quality relates to the level of degradation introduced by an imaging system during capture, processing, storage, compression, and display of face images [7], [8].

There are several potential applications of generic face image quality assessment (GFIQA):

- 1) In terms of generalized IQA, GFIQA can be used to improve model predictions. It has been demonstrated that the human visual system (HVS) is extremely sensitive to faces [9], [10]. An accurate face IQA metric could benefit the generalized IQA task.
- 2) In terms of face recognition, GFIQA can be used to boost face recognition performance. Face images with quality scores below a set threshold can be rejected during the acquisition process, which decreases the error rate in face recognition applications.
- 3) GFIQA can also be used in practical applications such as album optimization. For example, when importing photos from a digital camera, memory card, scanner, or computer hard disk to the digital photo album, face image quality can be used as a reference to make acceptance or rejection decisions.

In order to make accurate objective computations on generic face IQA, we further proposed a novel model to fulfill the task. Distinct from previous models [11]–[13], we in the first exploited deep generative priors to facilitate quality prediction. The motivation lying behind is that in pre-trained generative models, rich statistics of natural images are encoded, which could be utilized as latent references to the blind IQA task. By combining both distorted and latent reference features, accurate quality prediction results can be achieved.

The main contributions of this paper are:

- 1) We created the largest IQA database of human faces in the wild called the *Generic face image quality assessment 20k database* (GFIQA-20k). We extracted 20,000 faces from 1 million YFCC100M [14] images to ensure the diversity of (1) the individuals depicted, (2) the

image quality, and (3) distortion types, as well as (4) the contexts in which the images were captured. Selected freelancers with expert domain knowledge rated the quality of each image on average 12 times.

- 2) We validated the reliability of freelancers by analyzing their accuracy on gold-standard questions and self-consistency correlation on repeated questions. We validated the effectiveness of our dataset GFIQA-20k by transfer learning. High prediction correlation (over 0.95 SRCC) can be achieved using different baseline models.
- 3) We proposed a novel quality prediction model which is the first to exploit deep generative priors to facilitate the BIQA task. Using the rich statistics encoded in pre-trained generative models, we obtain prior preserved images and serve them as latent references to further improve prediction accuracy. Experimental results verified the effectiveness of the proposed model. The database and code will be made available at <http://database.mmsp-kn.de/gfqa-20k-database.html>.

II. RELATED WORKS

A. Quality Assessment of Face Images

There are two main research areas that deal with the quality of face images. The first, and most developed field, comes from the biometrics community for assessing the face quality for face recognition systems. This is most often denoted by face image quality assessment (FIQA). The second is GFIQA and relates to general image quality assessment dealing with perceptual image degradation. A in-depth discussion about the differences between the two fields has been presented by Schlett *et al.* [6].

FIQA has attracted increased attention in the face recognition community [5], [6]. Earlier works proposed the quality of a face image as its similarity to its reference face image with respect to multiple factors such as pose, expression, illumination, occlusion. For example, Sellahewa and Jassim [15] measures image quality in terms of luminance distortion by comparing a face input image against a known reference image. However, such approaches are hard to apply since they have to consider every possible factor manually 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.* [16] assigned qualities to face images by using two commercial off-the-shelf face recognition systems, where a face image is labeled by a good quality value if it matches well. Chen *et al.* [17] assumed 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 on *B*. Although there exists work that labels face quality manually, e.g., in binary class (good or bad) [18], Best-Rowden and Jain [19] conducted the

first subjective face quality assessment study. By conducting a small set of pairwise comparison of 13,233 face images, taken from [20], the quality ratings all images were inferred with matrix completion.

While there is significant development of FIQA, there is no study on GFIQA. To the best 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 become capable of effectively learning the natural image manifold and synthesizing high resolution images with pleasant visual quality [21]–[24]. With pre-trained GAN models, the well-learned image manifold is further explored to promote image manipulation and restoration tasks [25]–[30], referred to as generative priors. In order to utilize the rich semantic information encoded in generative models, input images are first mapped back to the intermediate features or latent space of pre-trained GANs [31], image manipulation or restoration tasks are then facilitated by feeding forward the inverted feature or code to generators.

There are typically two approaches to invert GAN models, 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 and modifying objective functions, image manipulation or restoration results can be obtained. In Image2StyleGAN [26] and Image2StyleGAN++ [27], latent codes are optimized and manipulated over the intermediate latent \mathcal{W} space of StyleGAN [23] and \mathcal{W}^+ space of StyleGAN2 [24], to achieve image inpainting, morphing and style mixing results. mGANPrior [32] optimized multiple latent codes and adaptively fused them to achieve various image restoration results, including image colorization, super-resolution, and image denoising. Noticing the distribution gap between the training and testing data, DGP [28] further proposed to fine-tune generator parameters on-the-fly to adapt the target images while maintaining statistics of the GAN learned priors. Though they require no training procedure, optimizing-based methods are usually time-consuming due to the large number of iterations needed for each instance image.

Learning-based approaches train an encoder to map an image to the latent code. By modifying encoder architectures and the objective function, various image manipulation and reconstruction results can be achieved. pSp [33] trained a multi-stage encoder to generate a series of style codes for StyleGAN2 [24], in order to handle various facial image translation tasks including conditional image synthesis, facial frontalization, and inpainting, etc. GLEAN [34] proposed an encoder-generator-decoder design to fulfill the large-factor image super-resolution task. GFP-GAN [29], and GPEN [35] fused target image features with generative prior features to restore real-world degraded face images. [30] warps and modulates generative prior features to achieve controllable and diverse image colorization results. Unlike optimization-based approaches, learning-based approaches obtain image

restoration results by only one feed-forward pass. However, extra data are usually needed to train these task-specific models.

In this paper, we in the first time explore the potential usage of generative priors to the task of IQA. Specifically, we employ rich statistics encoded by StyleGAN2 as latent references from the pristine face image manifold to facilitate the blind IQA problem. In order to utilize generative priors efficiently and effectively, we propose to train a multi-stage encoder and take advantage of multi-level attributes controlled by the style codes to obtain the generative statistics. The proposed approach avoids both the expensive optimization procedures and the extra data collection, and shows its superiority in the objective face IQA problem.

III. THE GFIQA-20K DATABASE

A. Face image dataset creation

Face images were collected from YFCC100M [14], a massive public multimedia database to ensure quality diversity. 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 [36] to detect faces and their corresponding key points, where the minimum size parameter of the face to detect was set as 400. 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 mid-point between the left and right eye. Next, the detected face image was cropped such that both the width and height of the crop are equal to four times the distance between the left and right eye. Finally, the crop was rescaled to 512×512 pixels. With this procedure, we collected 86,026 face images in the wild.

The MCTNN model cannot ensure that all faces are detected accurately. Sometimes, false positives (not human faces) and inaccurate key points were detected. In light of this, we manually checked and removed wrongly detected faces. This step reduced the number of samples to 53,058.

In the final step, to ensure the identity diversity of selected face images, we extracted their 512-dimensional deep features using the FaceNet model [37]. We next applied k -means clustering on the deep features to partition the 53,058 images into 20,000 clusters. In each cluster, an image is randomly selected as a representative. With this step, the number of face images decreased to 20,000, which formed the face images of the GFIQA-20k.

B. Subjective face image quality assessment

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 contained 40 images initially. To better monitor and analyze participants' performance, two reliability mechanisms were used. One is adding gold-standard or test data for which the correct answers are already known [38]. The other is utilizing a consistency test by posing the same question multiple times [39]. 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.

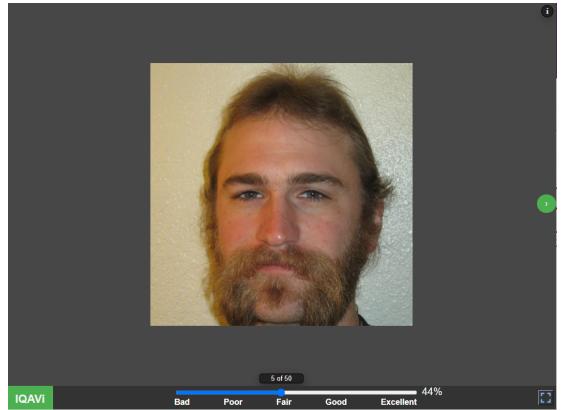


Fig. 1. UI for the subjective generic face IQA study. Each time participants are presented an image within a batch, they dragged a slider below the face to rate its visual quality from Bad (1%) to Excellent (100%).

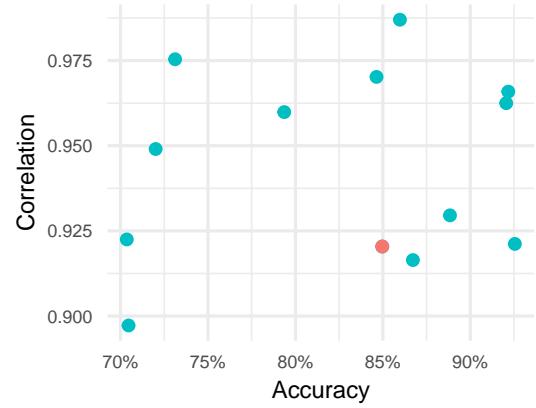


Fig. 2. Statistics of reliability analysis for freelancers. The self-consistency on repeatedly presented images, expressed as the SRCC (y-axis) between the two scores provided for each image vs the accuracy (x-axis) achieved on test images. All participants achieve an excellent self-consistency while maintaining a high level of accuracy relative to the gold-standard ratings. All participants submitted ratings for all images, except for the one shown in red, who submitted 153 batches (6,120 study images).

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 basic hardware requirements and detailed study steps were explained in the second and third sections, respectively. In the final section, apart from showing examples with different quality scales, we also gave some examples to differentiate technical face image quality and face attractiveness.

The user interface (UI) for the subjective generic face IQA study is shown in Fig. 1. We used the standard 5-point absolute category rating (ACR) scale, i.e., Bad, Poor, Fair, Good, and Excellent. To be more specific, participants are presented with a batch of face images, one at a time. Each

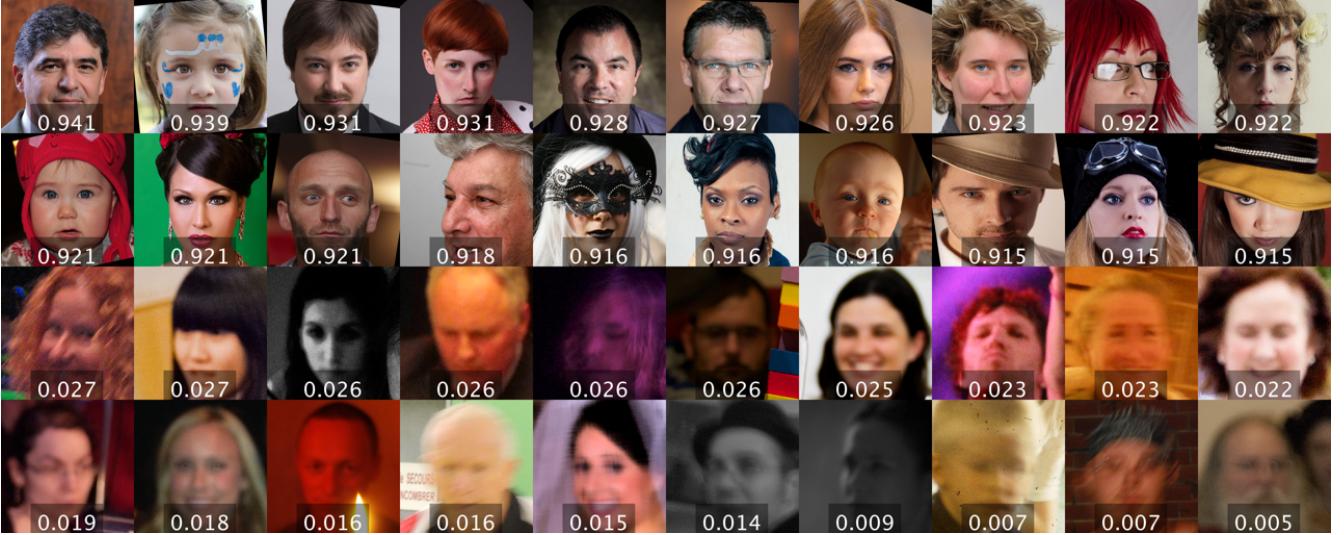


Fig. 3. Sampled face images with their corresponding MOS (white digits) in the GFIQA-20k dataset. Top two rows correspond to the 20 faces with highest MOS, whereas bottom two rows correspond to the 20 faces with lowest MOS.

time participants dragged a slider below the face image to rate its visual quality from Bad (1%) to Excellent (100%). As participant was required to drag a slider, we linear mapped the scale to the range of [0.01, 1] (rather than [0, 1]). To be more specific, let x be the original 5-point ACR, 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%, 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 [1]. 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 about that, and a range for the slider position was suggested. Freelancers could only proceed after having moved the slider into the correct range.

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 had achieved an 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 and the rest completed the entire study.

C. Reliability analysis of freelancers

To analyze the reliability of freelancers, we used accuracy to measure the performance of freelancers on gold-standard images. For a test image, a freelancer's answer would be counted as correct if his answer falls in the range of 1% to 35% when the image is labeled as low quality or of 65% to 100% when the image is labeled as high quality. For the consistency test, we used Spearman's rank correlation coefficient (SRCC).

The statistics of reliability analysis for freelancers is shown in Fig. 2. As can be observed, the accuracy (x-axis) on test images varies from 70% to 92%, whereas the SRCC (y-axis)

on repeated images various from 0.90 to 0.99. It shows that although some freelancers might not agree with the ground truth answers of gold-standard images we provided, they still keep a very high self-consistency on repeated images.

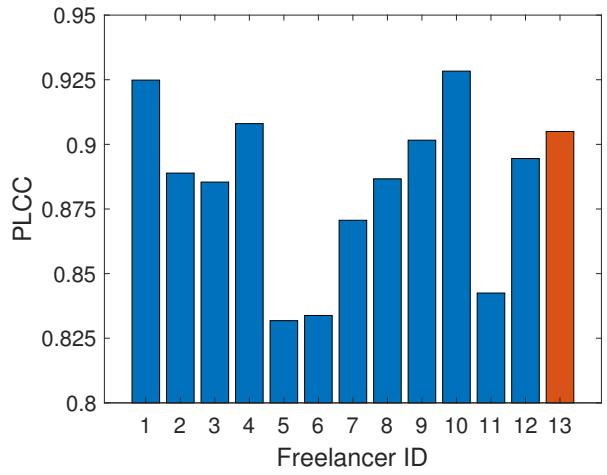


Fig. 4. Correlation between ratings of each individual freelancer and MOS. The PLCC various from 0.832 to 0.928. All participants submitted ratings for all images, except for the one shown in red, who submitted 153 batches (6,120 study images).

Apart from the reliability analysis, we report the Pearson linear correlation coefficient (PLCC) between individual ratings and MOS in Fig. 4. As can be observed, the ratings of each freelancer are highly correlated with MOS, which also demonstrate their reliability.

Although our analysis demonstrates the reliability of freelancers to some extent, freelancers might have paid insufficient attention during their work. Therefore, we screen ratings based on the assumption that ratings provided by reliable ones lie in an interval around the mean of all ratings in an image. To be more specific, the length of the interval is two times the standard deviation of all ratings from an image, ratings outside

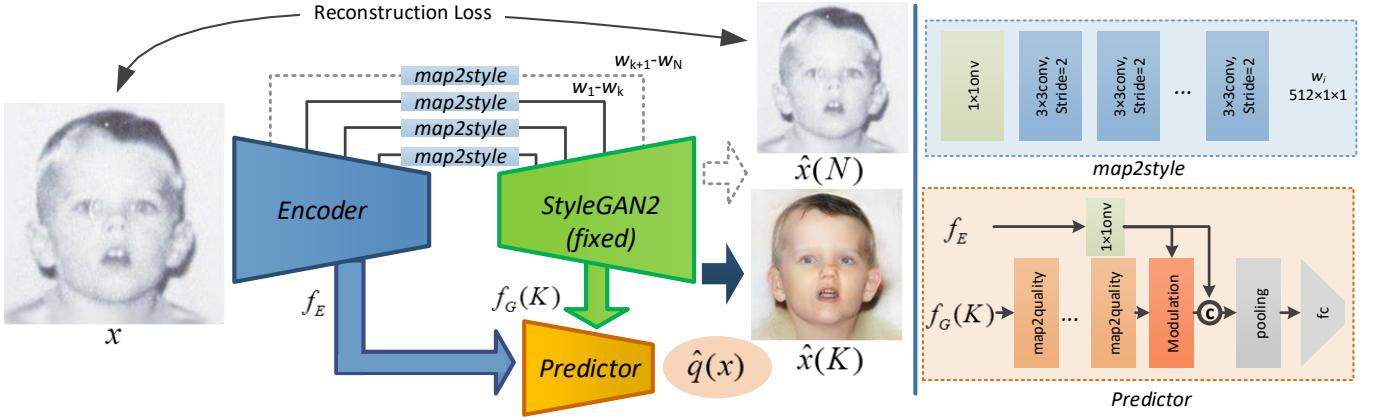


Fig. 5. The proposed face IQA model utilizing generative priors. Our framework consists of three parts: an encoder to both invert target image and extract distortion features, a generator to produce latent reference features in a pretrained GAN space, and a predictor to make quality estimations by refining and fusing target image features and latent reference features.

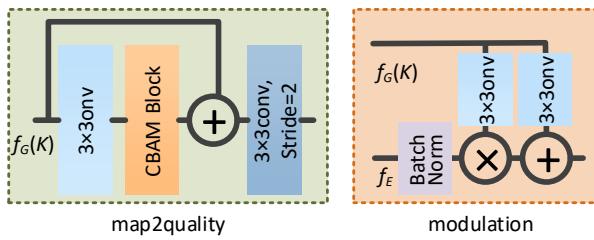


Fig. 6. The detailed architecture of the proposed map2quality and modulation blocks.

the interval will be removed and the rest yield mean opinion scores (MOS). Some face examples with their corresponding MOS are shown in Fig. 3.

IV. FACE IMAGE QUALITY ASSESSMENT WITH GENERATIVE PRIORS

In this section, we give a detailed description of the proposed objective face IQA model utilizing generative priors. As shown in Fig. 5, the overall framework consists of three parts: a multi-stage style code encoder to map the target image into the latent GAN space, a pretrained GAN model to generate intermediate reference features, and a quality predictor to make 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, semantic meaningful and attribute aware representations can be encoded. Second, by feeding forward the latent codes, intermediate GAN encoded statistics can be acquired and served as latent references to facilitate the challenging no-reference quality prediction task.

A. Obtaining GAN Encoded Statistics

Due to the lack of inference ability of GAN model, we first invert the target image x into latent codes $\{w_i \mid i =$

$1, 2, \dots, N\}$ in the GAN input space. Specifically, we choose to train an encoder E to map target images into the \mathcal{W}^+ space of StyleGAN2 [24], a state-of-the-art GAN model being capable of generating diverse facial images with high resolution and visual quality. Similar to [33], we encode N style codes with 512 dimensions from multi-stages of a ResNet50 [40] backbone network,

$$\{w_i\}, f_E = E(x; \theta_E), \quad i = 1, 2, \dots, N \quad (1)$$

where f_E are intermediate features and θ_E are parameters of E . The N latent codes $\{w_i\}$ are then fed to the different scales of a fixed StyleGAN2 generator G to produce a reconstruct result $\hat{x}(N)$. During generation, we add $\{\bar{w}\}$ to the average latent code \bar{w} in the pretrained generator space to achieve a good initialization,

$$\hat{x}(N) = G(\{w_i\} + \bar{w}), \quad i = 1, 2, \dots, N \quad (2)$$

where $\hat{x}(N)$ denotes the reconstructed result from x .

To train the encoder, we minimize θ_E over the reconstruction error between $\hat{x}(N)$ and x ,

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

where \mathcal{L} denotes loss functions.

In this way, we train encoder to map a distorted target image x into the GAN latent code space, and obtain the intermediate distorted features f_E for further quality prediction. However, in order to utilize rich generative priors, it is not enough to simply reconstruct the target image and extract the corresponding generative features. Since under the scenario where target images are contaminated with distortions, the reconstructed results also contain degradation patterns and thus harm the GAN encoded statistics. In order to obtain facial statistics in the original GAN space, we take advantage of the interpretable and controllable attributes of the multi-scale latent codes $\{w_i\}$. As latent codes at different scales are responsible for controlling level specific facial attributes [23], [41], we observed that the low level distortion attributes are

inherently encoded in latent codes at finer scales, and GAN statistics in early stages are still preserved. Therefore, during feed forward, we propose to inject the first $K, K < N$ codes to G , and discard 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.

$$\hat{x}(K), f_G(K) = G(\{w_i\} + \bar{\mathbf{w}}), \quad i = 1, 2, \dots, K \quad (4)$$

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

Under the framework, the first K codes are 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 encoded in low-level codes are discarded. It is worth noting that 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, the results already obtain sufficient statistical priors to be served as latent references for facilitating the IQA problem. By further refining reference features to target image features, we combine them together to make objective quality predictions.

B. Quality Assessment with Generative References

After obtaining target image features f_E and generative reference features $f_G(K)$, we refine and fuse them for quality prediction. Specifically, we extract high level representations $f_E \in \mathbb{R}^{2048 \times 16 \times 16}$ from the last stage of E to avoid another encoding process, and $f_G(K) \in \mathbb{R}^{32 \times 256 \times 256}$ from the last scale of G since it contains most generative information. We then apply an 1×1 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 reconstruct structures do not perfectly match target image, in order to refine the reference features, we use a CBAM [42] with residual connection inside each block to gradually adjust the features, and a 3×3 convolution with stride 2 and doubled channel numbers to resize features, as shown in Figure 6. We then modulate f_E by $f_G(K)$, following the spatially-adaptive denormalization (SPADE) operation proposed in [43]:

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

where $\gamma^{n,c,y,x}(f_G(K))$ and $\beta^{n,c,y,x}(f_G(K))$ are element-wise modulation parameters after convolving $f_G(K)$ with 3×3 kernels, n, c, y, x are batch, channel and spatial indexes, respectively. μ_E^c and σ_E^c denotes channel-wise 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 serves as refined reference features to the target image. Finally, we concatenate f_E with f_{mod} , and apply global average pooling followed by 3 fully connection

layers to regress the features to the quality prediction score $\hat{q}(x)$.

C. Objective Functions

We use three parts of 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, containing an \mathcal{L}_2 loss, a perceptual loss \mathcal{L}_{percep} and a face identity loss \mathcal{L}_{ID} , represented as,

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

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

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

where $f_{percep}(\cdot)$ extracts perceptual features from a pretrained VGG [44] model, and $R(\cdot)$ extracts identity vectors from a pretrained ArcFace [45] model.

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

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

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

$$\mathcal{L}_q(x) = \|q(x) - \hat{q}(x)\|_1. \quad (10)$$

Finally, we sum up the above loss functions with weighted factors $\lambda_i, i = 1, 2, \dots, 5$ and jointly train the proposed model,

$$\mathcal{L}(x) = \lambda_1 \mathcal{L}_2(x) + \lambda_2 \mathcal{L}_{percep}(x) + \lambda_3 \mathcal{L}_{ID}(x) + \lambda_4 \mathcal{L}_{reg}(x) + \lambda_5 \mathcal{L}_q(x). \quad (11)$$

D. Implementation Details

In Table I, we show detailed architecture of the proposed model. We show each module operation with its source input and output settings in the table. Output size is shown in the order of *Channels* \times *Height* \times *Width*. It is worth noting that for a pretrained StyleGAN2 generating a 512×512 resolution image, total 8 stages (1 stage without and 7 stages with upsampling) are included, and we combined every two stages in Table I for simplicity.

We implemented our model by Pytorch, and StyleGAN2 is implemented based on its Pytorch version re-implementation. Pretrained StyleGAN2 model is taken from GFPGAN [29], where they provided the parameters for a 512×512 generator model. We selected $K = 12$ for our model. During training, batch size is set to 16, and learning rate is set to 5×10^{-5} and then decayed by a factor of 10 at every 10 epochs. We trained the model with Adam optimizer [46] for total 25 epochs to report the final results. The whole model is trained using eight NVIDIA 1080Ti GPUs.

TABLE I
DETAILED ARCHITECTURE OF OUR PROPOSED MODEL.

Module	Operation	Input	Output Size
Encoder	ResNet Stage1	3 × 512 × 512 target image	256 × 128 × 128
	ResNet Stage2	ResNet Stage1	512 × 64 × 64
	ResNet Stage3	ResNet Stage2	1024 × 32 × 32
	ResNet Stage4	ResNet Stage3	2048 × 16 × 16
	map2style1	ResNet Stage1	512 × 1 × 1
	map2style2	ResNet Stage2	512 × 1 × 1
	map2style3	ResNet Stage3	512 × 1 × 1
	map2style4	ResNet Stage4	512 × 1 × 1
Generator	StyleGAN2 Stage1	map2style4, 4 × 4 constant	512 × 8 × 8
	StyleGAN2 Stage2	map2style3, StyleGAN2 Stage1	512 × 32 × 32
	StyleGAN2 Stage3	map2style2, StyleGAN2 Stage2	128 × 128 × 128
	StyleGAN2 Stage4	map2style1, StyleGAN2 Stage3	32 × 512 × 512
Predictor	map2quality × 5 modulation	StyleGAN2 Stage4	1024 × 16 × 16
	concat	ResNet Stage4, map2style	1024 × 16 × 16
	global average pool	ResNet Stage4, modulation	2048 × 16 × 16
	fully connection1	global average pool	2048
	fully connection2	fully connection1	1024
	fully connection3	fully connection2	512
			1

TABLE II
PERFORMANCE COMPARISONS BY TRAINING ON PREVIOUS GENERALIZED 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

V. EXPERIMENTS

A. Setup

We first 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). During testing, we selected the best performing model with the highest SRCC on the validation set for performance comparisons. We use SRCC, Pearson Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE) to evaluate model prediction accuracy and monotonicity.

B. How Do Generalized IQA Datasets Perform on Face Images?

In order to reveal quality properties of face data, we first conducted cross database tests to evaluate how previous generalized IQA datasets and models perform on the face IQA task. Specifically, we selected one synthetic IQA dataset LIVE [47] and four in-the-wild IQA datasets including LIVE Challenge (LIVEC) [48], KonIQ-10k [1], SPAQ [2] and KonIQ++ [49] for cross testing. We trained IQA models MEON [50], HyperIQA [51], Koncept512 [1], MT-A [2], and BIQA model [49] on the five datasets respectively, and tested them on the GFIQA-20k test subset. Among testing models, MEON [50] and HyperIQA [51] are state-of-the-art (SOTA) IQA methods performing well on synthetic and authentic distortions respectively, and the other models are proposed along with their training datasets. We report 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 images of smartphone photography collected in SPAQ. Surprisingly, we found that KonIQ-10k and its extension KonIQ++ both performed relatively well (around 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 [14], and content overlapping might exist. Despite this, there still left a space for further performance improvement, and we discuss the development of face IQA models in the following.

C. Performance Evaluation with Competing Models

In this subsection, we conduct performance comparisons with models trained with GFIQA-20k data. Due to the lack of baselines in the generalized face IQA task, we first came up with diverse baseline models from different upstream tasks. Specifically, we selected ArcFace [45] pretrained on a refined face recognition dataset MS1M [52], Koncept512 [1] pretrained on a general IQA dataset KonIQ-10k [1], and ResNet50 [40] pretrained on the image classification dataset ImageNet [53]. We finetuned these models on the GFIQA-20k training subset and report the results in Table III. As can be seen, by simply transfer learning, all the baseline models achieved high performance (over 0.95 SRCC). The result demonstrates the effectiveness of the collected data in handling face IQA task.

TABLE III
PERFORMANCE COMPARISONS OF TRANSFER LEARNING OF BASELINE MODELS, GENERALIZED 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
PQR	0.9517	0.9530	0.0557
HyperIQA	0.9626	0.9628	0.0502
MUSIQ	0.9628	0.9634	0.0504
Proposed	0.9639	0.9644	0.0489

We further compared the proposed model with five general IQA models BRISQUE [54], CORNIA [55], PQR [56], HyperIQA [51] and MUSIQ [57]. Here BRISQUE represents a natural scene statistic based IQA method, CORNIA is a traditional learning based method and the rest are SOTA in-the-wild IQA models. Among all the competing models, the proposed model outperformed the others by all three criteria. We also observe that the transfer learned model ResNet50 performed even slightly better than the competing SOTA IQA models. This is probably because the cropping strategy applied in HyperIQA and MUSIQ damages facial structures which are, however, crucial to the task.

To further validate the effectiveness of the proposed model, in Table IV, we evaluated how the model performed when the number of training samples varies. We compared the proposed

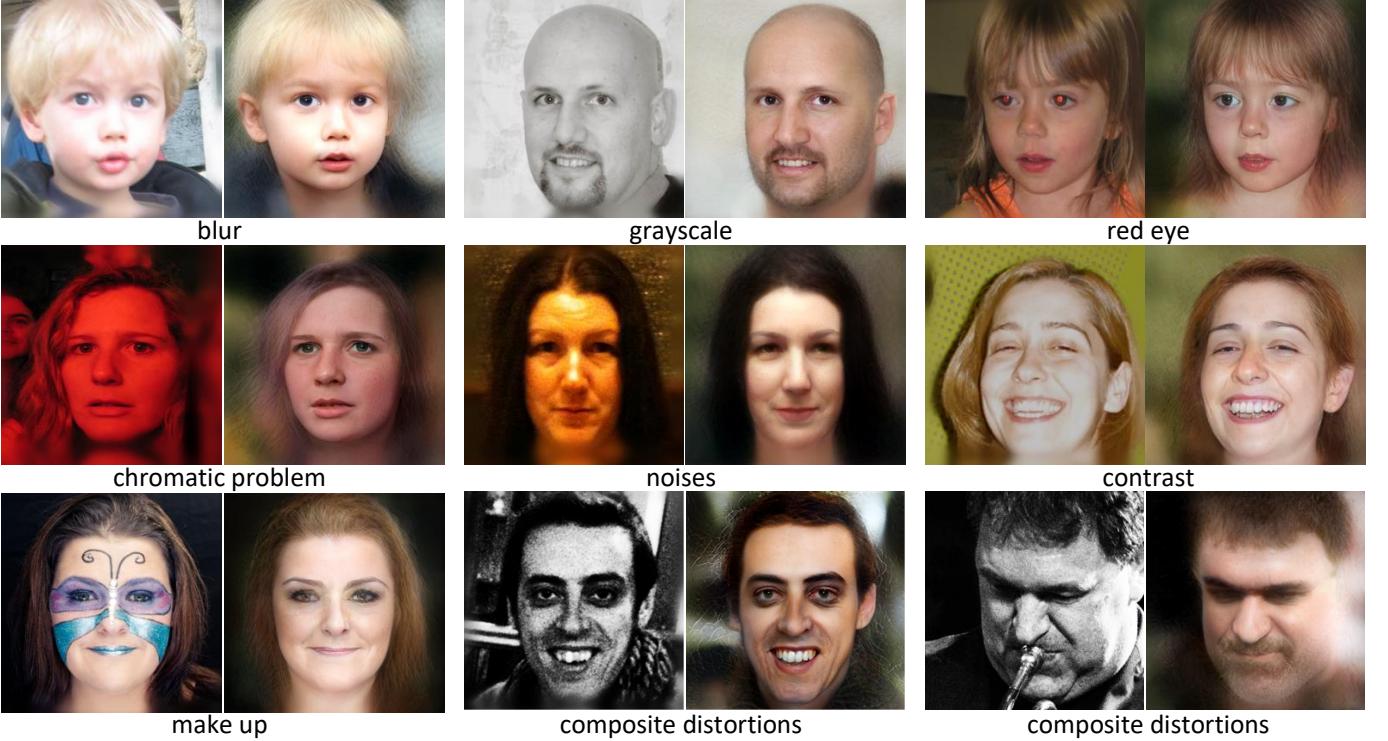


Fig. 7. We visualize some generated latent reference images (right) with respect to their distorted images (left). Thanks to the powerful generative prior information, latent reference images are constructed despite various distortions contained in inputs, being able to facilitate our quality prediction task.

TABLE IV
PERFORMANCE COMPARISONS WHEN THE TRAINING SAMPLE SIZE VARIES.

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

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 for testing, except for the validation subset. Similarly, the proposed model showed consistently superior prediction accuracy for variable training sample sizes.

D. 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. 7, we show pairs of distorted images x and the reconstructed latent references $\hat{x}(K)$. 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 and thus well serve as latent references to the degraded input images, which further facilitates the blind face IQA task. It is worth noting that since we impose loss constraints mainly on face regions, the generated latent reference images might not

precisely match target images in background. However, since HVS is extremely sensitive to human faces, the difference in background regions does not contribute critically to the perceptual face image quality. Therefore, in the IQA task, we do not need perfect reconstruct images as references, but instead we extract the intermediate features and modulate them to promote the model performance.

E. Ablation Study

In this subsection, we conducted several ablation experiments to evaluate the effectiveness of the model design. We first compared with the baseline encoder ResNet50 trained by only \mathcal{L}_q loss. We then added StyleGAN2 and constraints of reconstruction error 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 $K = 4, 8, 12, 16$ respectively, while keeping other components fixed. Last, we validated the designation of the quality predictor. We substituted the map2quality module to ordinary convolution blocks (w/o map) while keeping other



Fig. 8. We show some flawed generative prior references, the images include rotating faces, turning out to be challenging samples for a pretrained StyleGAN2 model to reconstruct. Involving images with more diverse rotation angles to train more powerful generative priors might be a solution.

parts 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 \uparrow	PLCC \uparrow	RMSE \downarrow
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	0.9639	0.9644	0.0489

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

VI. DISCUSSION

Though the proposed model showed its superiority in face quality prediction, we find that generative priors occasionally lead to unsatisfactory reconstruction results, as shown in Fig. 8. We assume this is because the generative prior model StyleGAN2 was mostly trained with frontally viewed samples but few rotated faces. Thus, the generator can produce frontally viewed faces but underperforms otherwise. To address this challenging issue, training the generative model with diverse view angles to provide more powerful priors might be a solution, and we leave the task for future work.

VII. CONCLUSION

In this paper, we created GFIQA-20k which is the largest in-the-wild database for human face quality prediction. It may fa-

cilitate several applications such as digital album optimization, face image restoration/enhancement, and generalized IQA. We further proposed a novel face IQA model, the first to exploit generative priors for this task. Benefiting from referencing rich statistics encoded in pre-trained deep generative models, the model made accurate objective predictions, and experiments validated its effectiveness.

REFERENCES

- [1] 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.
- [2] 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 (CVPR)*, 2020, pp. 3677–3686.
- [3] T. Kanade, “Picture processing by computer complex and recognition of human faces,” *Ph. D. Thesis, Kyoto University*, 1973.
- [4] 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.
- [5] P. Grother, A. Horn, M. Ngan, and K. Hanaoka, “Ongoing face recognition vendor test (FRVT) part 5: Face image quality assessment,” *NIST Interagency Report*, 2020.
- [6] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, “Face image quality assessment: A literature survey,” *ACM Computing Surveys*, 2021.
- [7] 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.
- [8] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, “Graphiq: Learning distortion graph representations for blind image quality assessment,” *IEEE Transactions on Multimedia*, 2022.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] S. Bharadwaj, M. Vatsa, and R. Singh, “Can holistic representations be used for face biometric quality assessment?” in *IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 2792–2796.

- [17] 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.
- [18] X. Zhao, Y. Li, and S. Wang, "Face quality assessment via semi-supervised learning," in *International Conference on Computing and Pattern Recognition (ICCP)*, 2019, pp. 288–293.
- [19] 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.
- [20] 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.
- [21] 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 (ICLR)*, 2018.
- [22] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations (ICLR)*, 2018.
- [23] 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 (CVPR)*, 2019, pp. 4401–4410.
- [24] 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 (CVPR)*, 2020, pp. 8110–8119.
- [25] 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 (CVPR)*, 2018, pp. 8798–8807.
- [26] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the stylegan latent space?" in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 4432–4441.
- [27] ———, "Image2StyleGAN++: How to edit the embedded images?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8296–8305.
- [28] 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 (ECCV)*. Springer, 2020, pp. 262–277.
- [29] 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 (CVPR)*, 2021, pp. 9168–9178.
- [30] 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 (ICCV)*, 2021, pp. 14377–14386.
- [31] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "GAN inversion: A survey," *arXiv preprint arXiv:2101.05278*, 2021.
- [32] J. Gu, Y. Shen, and B. Zhou, "Image processing using multi-code GAN prior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3012–3021.
- [33] 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 (CVPR)*, 2021, pp. 2287–2296.
- [34] 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 (CVPR)*, 2021, pp. 14245–14254.
- [35] 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 (CVPR)*, 2021, pp. 672–681.
- [36] 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.
- [37] 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 (CVPR)*, 2015, pp. 815–823.
- [38] 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.
- [39] 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 (ISM)*, 2011, pp. 494–499.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [41] Y. Alharbi and P. Wonka, "Disentangled image generation through structured noise injection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5134–5142.
- [42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 3–19.
- [43] 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 (CVPR)*, 2019, pp. 2337–2346.
- [44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [45] 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 (CVPR)*, 2019, pp. 4690–4699.
- [46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.
- [47] 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.
- [48] 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.
- [49] 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 (BMVC)*, 2021.
- [50] 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.
- [51] 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 (CVPR)*, 2020, pp. 3667–3676.
- [52] 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 (ECCV)*. Springer, 2016, pp. 87–102.
- [53] 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 (CVPR)*, 2009, pp. 248–255.
- [54] 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.
- [55] 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 (CVPR)*, 2012, pp. 1098–1105.
- [56] 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.
- [57] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 5148–5157.