IFQA: Interpretable Face Quality Assessment

Byungho Jo[1] Donghyeon Cho[2] In Kyu Park[1] Sungeun Hong[1]
[1]Inha University [2]Chungnam National University
byunghojo12@gmail.com cdh12242@cnu.ac.kr {pik, csehong}@inha.ac.kr

Abstract

Existing face restoration models have relied on general assessment metrics that do not consider the characteristics of facial regions. Recent works have therefore assessed their methods using human studies, which is not scalable and involves significant effort. This paper proposes a novel face-centric metric based on an adversarial framework where a generator simulates face restoration and a discriminator assesses image quality. Specifically, our per-pixel discriminator enables interpretable evaluation that cannot be provided by traditional metrics. Moreover, our metric emphasizes facial primary regions considering that even minor changes to the eyes, nose, and mouth significantly affect human cognition. Our face-oriented metric consistently surpasses existing general or facial image quality assessment metrics by impressive margins. We demonstrate the generalizability of the proposed strategy in various architectural designs and challenging scenarios. Interestingly, we find that our IFQA can lead to performance improvement as an objective function. The code and models are available at `https://github.com/VCLLab/IFQA`.

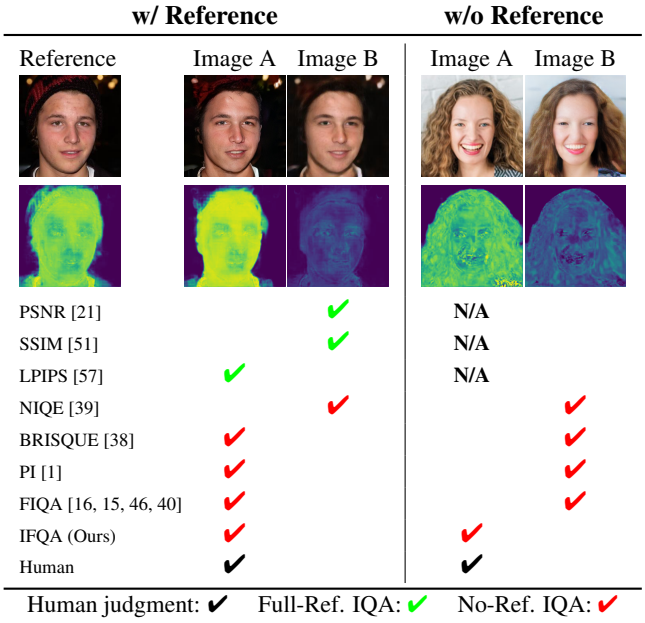

Figure 1. Which of 'Image A' or 'Image B' is closer to the given reference image or looks high-quality? General full-reference metrics (*e.g.* PSNR/SSIM), no-reference metrics (*e.g.* NIQE, BRISQUE, PI), and FIQA methods are inconsistent with human judgment. LPIPS agrees with human judgment but cannot be applied to the blind face restoration scenario. Our IFQA is consistent with human judgment and can provide interpretability maps where the brighter the area, the higher the quality.

1. Introduction

Considerable efforts have been devoted to restoring facial images from degraded images [37, 53, 49]. Conventional face restoration studies adopt full-reference metrics widely used in general image restoration, *e.g.* PSNR [21], SSIM [51], LPIPS [57], to evaluate the similarity between reference and restored images. Blind face restoration (BFR) studies that can handle multiple unknown degradations adopt no-reference metrics such as NIQE [39] and BRISQUE [38]. However, because existing general metrics do not consider facial characteristics, their judgments could differ from human perceptions as shown in Figure 1.

Recent face restoration studies [48, 37, 53] have evaluated their methods using human study rather than evaluation metrics. However, human-oriented assessments widely used in the face restoration field have fatal limitations: *first*, they are unscalable, *second*, a number of assessors and their (large) variances between each other and, *third*, cost of conducting the assessment. The absence of appropriate face-oriented metrics results in significant expense and time for evaluation, which is becoming one of the major bottlenecks for the emerging face restoration field. A question naturally arises in this context is whether we need a face-specific evaluation metric. Crucially, the face domain is different from conventional object categories in ImageNet [43] or COCO [34] because of its unique properties (*e.g.* geometry and textures) and a variety of downstream tasks [22, 7, 35]. We argue that face images should be evaluated differently

from general image domains in terms of image quality assessment (IQA). The findings of early psychological studies [24, 10, 47], in which human brains use different areas (*i.e*. fusiform face area) to recognize common objects and faces, also support our claim.

This paper introduces a novel face-oriented metric called interpretable face quality assessment (IFQA) based on an adversarial network [11]. The generator, which is a plain face restoration model, attempts to restore high-quality images from low-quality images. Sub-regions from high-quality images provide 'real' supervision to the discriminator, whereas low-quality images and regions from restored face images by the restoration model provide 'fake' supervision. Inspired by human face perception [47] in which facial primary regions (*e.g*. eyes, nose, and mouth) have a great effect on human face perception, we propose facial primary regions swap (FPRS) that places a greater emphasis on facial primary regions. Unlike existing mix-based augmentations [8, 56, 54] that randomly extract local patches from arbitrary positions, FPRS changes regions within a set of the facial primary regions. Additionally, our U-shaped architecture allows us to produce not only single image-level quality scores but also interpretable per-pixel scores. The proposed metric is related to face image quality assessment (FIQA) [16]. In contrast to FIQA approaches mainly rely on face recognition systems, our IFQA can be considered a more generalized face-oriented metric independent of a specific high-level task.

We make it clear that our framework aims to realize unresolved face-oriented metrics despite numerous demands raised by existing face restoration studies. In our evaluations across various architectures and scenarios, our proposed metric shows higher correlations with human cognition than general IQA metrics and state-of-the-art FIQA metrics. The contributions of this study are as follows:

- We propose a new dedicated framework for the face-specific metric that considers the importance of the face primary regions, such as eyes, nose, mouth.
- Our face-oriented metric matches human judgment significantly more than existing general no-reference and state-of-the-art FIQA metrics.
- Pixel-level evaluation scores enable interpretable image quality analysis that cannot be provided by traditional single-score-based metrics.

2. Related Work

2.1. Face Image Restoration

A series of face image restoration methods have been proposed for addressing certain types of facial image degradation, such as low-resolution, noise, and blur [60, 55, 19, 59]. Although previous studies have shown promising results, they exhibit poor performance in real-world images with unknown and complex degradation. Some blind face restoration (BFR) approaches have been proposed to address this issue [33]. Prior BFR studies utilized face-specific priors, such as facial component dictionaries [32], facial parsing maps [4], high-quality guided images [33]. Among them, GAN inversion approaches based on StyleGAN [26, 27] have shown promising results [37, 53, 49]. Despite significant advances in face restoration methodologies, evaluation metrics, which cannot reflect facial characteristics, are borrowed from general image restoration. Therefore, state-of-the-art studies have conducted costly human studies to demonstrate the superiority [37, 53, 48] of their models, which highly motivates this study.

2.2. General Image Quality Assessment

Image quality assessment (IQA) aims to measure the perceptual quality of images and existing approaches can be categorized into two groups: full-reference (FR-IQA) and no-reference (NR-IQA). FR-IQA evaluates the statistical or perceptual similarity between restored images and reference images. PSNR [21] and SSIM [51] are widely used to evaluate face restoration models [20]. Perceptual metrics have been introduced to alleviate the large semantic gap between traditional FR-IQA and human cognition [57, 30, 9]. Although the aforementioned metrics are reasonable choices for measuring restoration results, they cannot be applied to real-world scenarios without reference images.

A series of NR-IQA approaches—BRISQUE [38], NIQE [39], and PI [1]—have been devised to measure the naturalness of images in the blind image quality assessment. Along with the successful application of NR-IQA in natural scenes, there have been attempts to apply NR-IQA to the BFR problem. However, numerous BFR studies [53, 48, 37] discovered that general NR-IQA metrics have limitations for assessing restored facial images; and therefore, they conduct human studies for evaluation. Consequently, the absence of appropriate evaluation metrics for face restoration requires substantial expenses and becomes a major bottleneck in the face image restoration field. To address this critical issue, we propose an evaluation metric specifically designed to focus on facial primary regions.

2.3. Face Image Quality Assessment

Face image quality assessment (FIQA) supports face recognition systems by deciding whether or not to discard LQ face images as a preprocessing step. This process builds stable and reliable face recognition systems in real-world scenarios (*e.g*. surveillance cameras and outdoor scenes). Early FIQA studies exploited analytics-based methods while recent FIQA studies have concentrated their efforts on a learning-based strategy that gen-

Figure 2. Comparison of PSNR/SSIM and human assessment on restored face images. PSNR/SSIM provides higher scores to 'Image A' than 'Image B' while human subjects vote 'Image B' as higher quality face images than 'Image A'.

erates image quality scores directly from face recognition models [16, 46, 40]. Although previous FIQA studies have shown remarkable results compared with general no-reference IQA [38, 39, 1], they solely focused on the face recognition task. Crucially, FIQA aims to assess the quality of a face image from the point of view of its use in face recognition tasks, involving objective functions derived from face recognition. Unlike FIQA, our metric does not focus on specific face-related tasks, and thus can be considered a more generalized face image evaluation assessment.

3. Proposed Metric

3.1. Pilot Study

In the preliminary experiments for the assessment of face restoration results, we discovered that facial primary regions play a critical role in human visual perception. Most of the participants in the preliminary human study answered that images from 'Image B' (third row) in Figure 2 are more realistic than the ones shown in 'Image A' (second row). However, PSNR and SSIM metrics score 'Image A' higher than 'Image B'. Notably, human visual perception is significantly affected by the overall structure and distortions in facial primary regions such as the eyes and nose as shown in the first and second columns.

3.2. Proposed Framework

Motivated by the observation in the pilot study, we introduce an evaluation metric considering facial characteristics. The overall framework is illustrated in Figure 3.

Generator for image restoration: The generator consisting of a simple encoder-decoder architecture can be considered a plain face restoration model that outputs restored face images. The generator is trained to restore LQ images to 256×256 HQ images. During the training phase, we deliberately corrupt HQ images in the FFHQ dataset to make input LQ images as the following BFR formulation:

$$I_{LQ} = ((I_{HQ} \otimes k) \downarrow_r + n_\sigma)_{JPEG_q}, \tag{1}$$

where k is a kernel randomly selected between Gaussian and motion-blur kernel. Factors of downsampling, Gaussian noise, and JPEG compression are denoted as r, n_σ and q. Following previous BFR studies [32, 53, 49], the range of each factor is set as r: $[0.4, 0.9)$, n_σ: $[50, 250)$, q: $[5, 50)$.

Discriminator for quality assessment: The discriminator aims to evaluate the quality of query images trained in an adversarial manner with the generator. We design our discriminator to output per-pixel scores using U-Net-based architecture [44] to enhance the generalization ability. This architecture enables us to classify the regions from HQ images as 'real' while the regions from LQ or RF images as 'fake'. Quality score as a single value (*i.e.* an image-level score) can be obtained via aggregating pixel-level scores. Notably, unlike traditional discriminators with adversarial training, we provide 'fake' supervision to the generator's input (*i.e.* LQ images) as well as the generator's output (*i.e.* RF images). We only consider the face region in HQ images as 'real' labels instead of entire HQ images. This simple trick helps our metric to focus more on the face region than the entire image. In the ablation study, we experimentally show that the 'real' supervision setting using only the face region has higher correlations with human judgment than the global region. We exploit the off-the-shelf face segmentation model [5] pre-trained on CelebAMask-HQ [31] to obtain binary facial masks from the images.

Furthermore, we propose a novel augmentation technique called facial primary regions swap (FPRS) to reflect facial characteristics to the proposed metric, as shown in Figure 4. Firstly, we apply an off-the-shelf landmark detector [2] to HQ images to obtain facial primary regions. Unlike augmentation techniques such as original CutMix for general purpose, we utilize RoIAlign [13] to crop the primary region of facial components. Subsequently, the regions extracted from the LQ or RF images are arbitrarily swapped with the facial primary regions from the HQ images. Let $I_{LQ/RF}$ denote an LQ or RF image and I_{HQ} denote an HQ image. Through FPRS operation, we can generate a new image pair to be used for discriminator training as follows:

$$\begin{aligned} I_{HQ \to LQ/RF} &= \mathbf{M_{FPRS}} \odot I_{HQ} \\ &\quad + (\mathbf{1} - \mathbf{M_{FPRS}}) \odot I_{LQ/RF} \\ I_{LQ/RF \to HQ} &= \mathbf{M_{FPRS}} \odot I_{LQ/RF} \\ &\quad + (\mathbf{1} - \mathbf{M_{FPRS}}) \odot I_{HQ}, \end{aligned} \tag{2}$$

where $\mathbf{M_{FPRS}} \in \{0, 1\}^{H \times W}$ is a binary mask for randomly selected facial primary regions. $\mathbf{1}$ is a binary mask filled with ones. $\odot$ indicates element-wise multiplication.

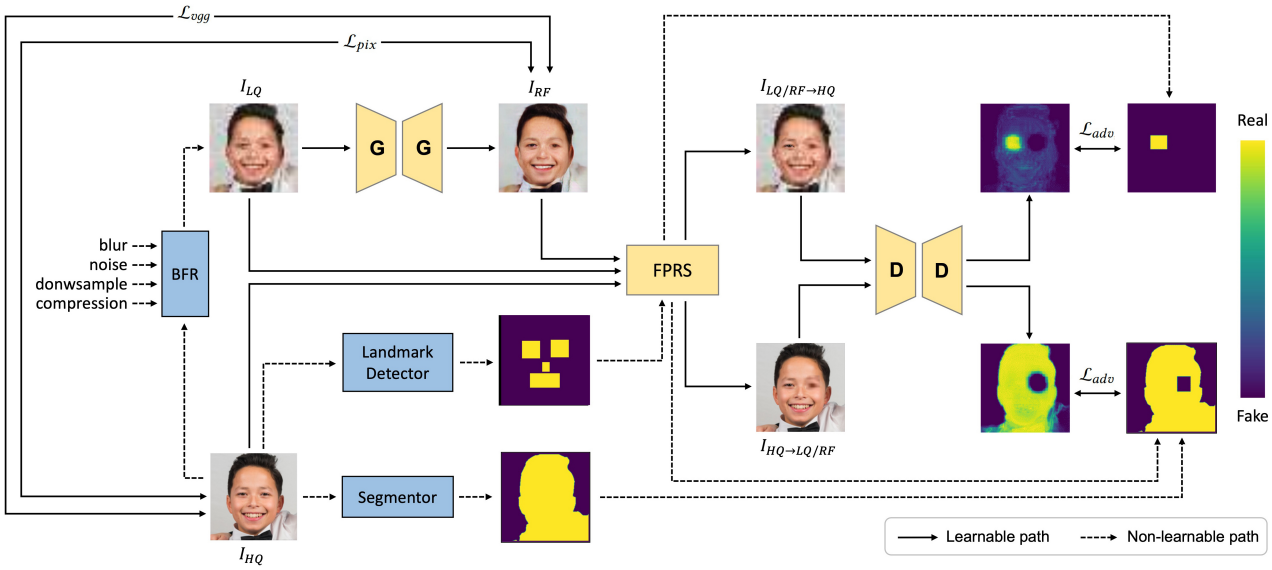

Figure 3. IFQA framework outline. Given HQ images, we obtain LQ images via BFR formulation. The generator (G) mimics face restoration models, while the discriminator (D) is used to evaluate image quality by determining high-quality regions as 'real' and low-quality or restored regions as 'fake'. Through its U-Net architecture, the discriminator is able to evaluate the image pixel-by-pixel. FPRS allows the proposed metric to give more weight to facial primary regions that have a significant impact on human visual perception.

Objective function: The IFQA framework is trained by least-square-based adversarial learning [36] between the generator and discriminator. The generator is trained to fool the discriminator, and the objective function for the generator is defined as follows:

$$\mathcal{L}_{adv,G} = \mathbb{E}_{I_{RF}}[(D^U(I_{RF}) - \mathbf{1}))^2], \tag{3}$$

where $D^U(\cdot)$ refers to U-Net-based discriminator that outputs per-pixel scores. Also, we adopt pixel loss to enforce the generator to make I_{RF} to be similar to the corresponding HQ image. The pixel loss compares all of the pixel values between the I_{RF} and HQ images as follows:

$$\mathcal{L}_{pix} = \mathbb{E}_{I_{RF},I_{HQ}}[||I_{RF} - I_{HQ}||_2]. \tag{4}$$

To produce photo-realistic facial images, we leverage perceptual loss [23] using the weights of the pre-trained VGG-19 as follows:

$$\mathcal{L}_{vgg} = \sum_i ||f_i(I_{RF}) - f_i(I_{HQ})||_1, \tag{5}$$

where $f_i(\cdot)$ is the i-th feature extracted from the pre-trained VGG-19 network. Specifically, we use pooling layers in five convolutional blocks for perceptual loss.

Meanwhile, the objective function of the discriminator is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv,D} = \mathbb{E}_{I_{HQ}}[(D^U(I_{HQ}) - \mathbf{M_{FACE}})^2] \\ + \mathbb{E}_{I_{LQ},I_{RF}}[D^U(I_{LQ/RF})^2],\end{aligned} \tag{6}$$

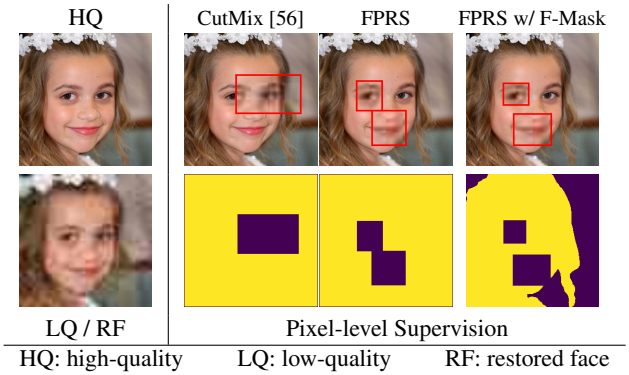

Figure 4. Supervision for IFQA metric. Regions from high-quality images provide 'real' labels (yellow), while regions from low-quality or restored face images give 'fake' labels (purple). The red box indicates the randomly selected swapped region.

where I_{HQ} and $\mathbf{M_{FACE}}$ are HQ images and facial binary masks filled with 1s only for the face area, respectively. $I_{LQ/RF}$ is LQ or RF images in which $I_{RF} = G(I_{LQ})$. The full objective function can be summarized as follows:

$$\min_G \max_D \quad \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{pix} + \lambda_2 \mathcal{L}_{vgg}, \tag{7}$$

where $\mathcal{L}_{adv} = \mathcal{L}_{adv,G} + \mathcal{L}_{adv,D}$, and λ_1 and λ_2 are scaling parameters. We set λ_1 and λ_2 as 50 and 5, respectively.

Assessment protocol: Once the IFQA framework is trained, we only use the per-pixel discriminator for image quality assessment. The pixel-level assessment score of the discriminator enables us to perform an interpretable in-depth analysis. Given an input image I, we can obtain an

image-level quality score (QS) by simply averaging every pixel-level score from the per-pixel discriminator as:

$$QS = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} D_{i,j}^{U}(I) \tag{8}$$

4. Experiments

4.1. Implementation Details

Datasets: Randomly selected 20,000 images from FFHQ [26] were used for training the IFQA framework. Meanwhile, we constructed three types of benchmark test datasets. First, we construct a test set by combining CelebA-HQ [25] and FFHQ images with high-quality images and considerable variations, widely used in face restoration tasks. Given high-quality images of the test set, we obtained low-resolution query images using the BFR formulation Eq. 1. Second, considering real-world scenarios, we constructed a test set using in the wild face (IWF) [53], which is widely used for BFR problems. Notably, the IWF dataset only provides low-quality images without high-quality reference images. Third, CelebA-HQ, FFHQ, and IWF were combined and used as a test set for the ablation study to demonstrate the generalization ability of the proposed metric.

Image restoration models: We used various image restoration models including general image restoration models (*e.g*. RCAN [58], DBPN [12]) and face restoration models (*e.g*. HiFaceGAN [52], DFDNet [32], GPEN [53]) to evaluate the proposed metric quantitatively and qualitatively.

4.2. Quantitative Analysis

Human study protocol: For quantitative comparison with the proposed IFQA metric and the existing IQA metrics, we conducted a human study of ranking the realistic facial images from given images. We carefully designed survey questions and prepared face images to use to estimate human visual judgments. One sample consisted of six images, including an LQ image and one image each restored from RCAN, DBPN, DFDNet, HiFaceGAN, and GPEN. All 200 sample images were randomly selected from the FFHQ, CelebA-HQ, and IWF datasets. The total number of images for the human study was 1,200.

We asked participants to rank given samples from closest to furthest to a realistic human face. We used Amazon Mechanical Turk crowdsourcing [41] to gather participant's responses systematically. We also received responses from researchers who are majors in various AI-related fields and are not directly related to this study. We assigned 30 subjects per sample, and the total number of responses across all samples was 6,000. The final rank of each sample was calculated by the weighted average rank. Figure 5 present the box plot of our human study results.

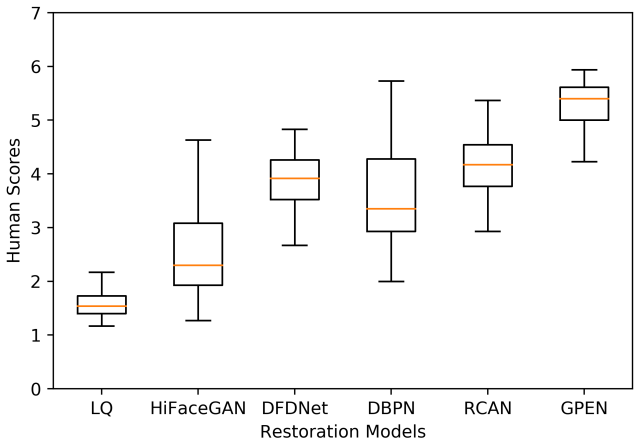

Figure 5. Box plot of restoration models through human study.

IQA comparative analysis: Once human ranking responses were obtained for each sample, we measured Spearman's rank order correlation coefficient (SRCC) [18] and Kendall rank order correlation coefficients (KRCC) [28] for quantitative analysis. These approaches are widely used to measure the correlation between human judgments and other metrics across various fields.

Aanalysis on FFHQ & CelebA-HQ: First, we performed a comparative evaluation on FFHQ and CelebA-HQ with the existing no-reference IQA (NR-IQA) metrics, including state-of-the-art FIQA metrics. Table 1 shows that IFQA has the highest correlations with human preferences in both SRCC and KRCC metrics. Interestingly, NIQE, the widely used NR-IQA metric, shows the lowest correlation. We compared with the recent FIQA methods, which are specially designed for face recognition. IFQA is superior to other FIQA metrics. Evidently, our face-oriented metric is steadily more consistent with human judgment compared with the existing NR-IQA metrics.

Although the proposed IFQA is an NR-IQA metric, we conducted a comparative analysis with full-reference metrics (*i.e*. FR-IQA) to prove the general applicability of the proposed metric. The traditional but widely used metrics, PSNR [21] and SSIM [51], show values less than 0.2 in both SRCC and KRCC. Perceptual metric, LPIPS [57], shows 0.6685 for SRCC and 0.5560 for KRCC, showing better performance than the existing FR-IQA metrics. Crucially, all FR-IQA metrics cannot be applied to practical scenarios in the wild in which there is no reference image. The proposed IFQA shows the highest correlations with the human perception among NR-IQA metrics and is comparable to the state-of-the-art FR-IQA metric despite not requiring any reference image.

Analysis on a real-world dataset: We measure the correlation value using real-world face images [53] to prove the generalization ability of the assessment. We exclude

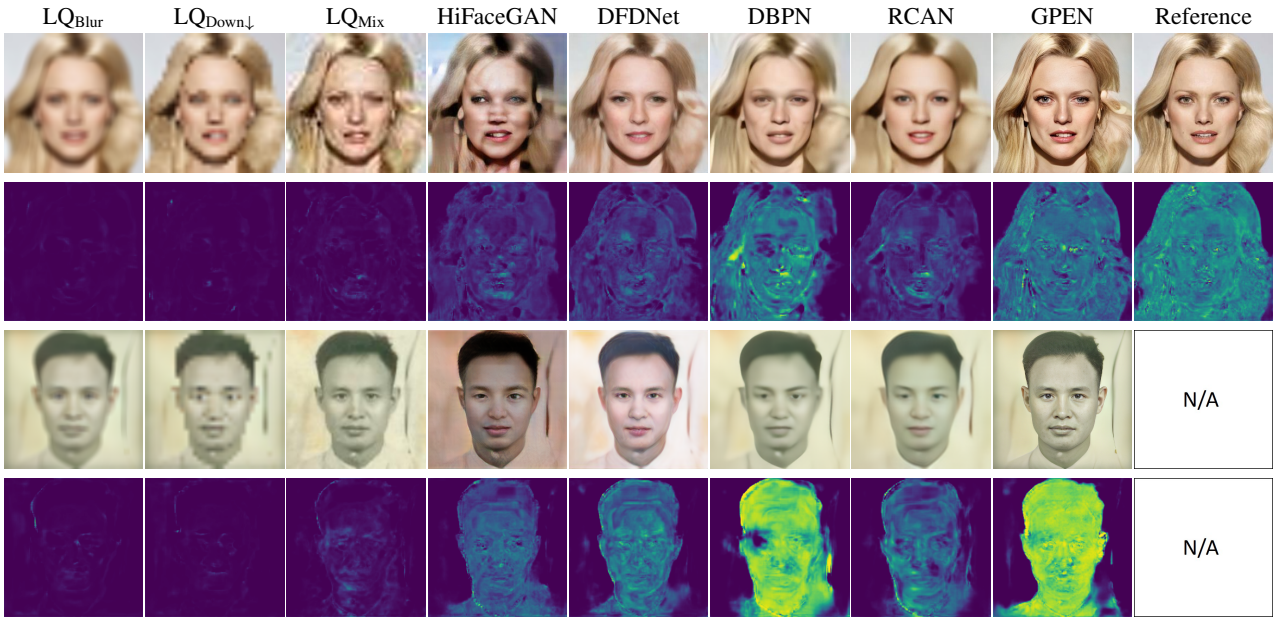

Figure 6. Interpretable visualization of the proposed metric on various types of LQ images, HQ images (*i.e.*, reference), and RF images from the restoration models. The first and second rows show images from FFHQ and their corresponding interpretability maps, respectively. The third and fourth rows present pairs from IWF that does not provide reference images. Brighter area indicates the higher quality.

Table 1. Comparative analysis on FFHQ and CelebA-HQ.

Metric	Type	SRCC ↑	KRCC ↑
NIQE [39]	NR-IQA (General)	0.2668	0.2039
PI [1]		0.4125	0.3173
BRISQUE [38]		0.4405	0.3373
IFQA **(Ours)**		**0.6400**	**0.5186**
SER-FIQ [46]	NR-IQA (FIQA)	0.3554	0.2706
FaceQnet-V1 [15]		0.4560	0.3453
FaceQnet-V0 [16]		0.5491	0.434
SDD-FIQA [40]		0.5920	0.4840

Table 2. Comparative analysis on IWF.

Metric	Type	SRCC ↑	KRCC ↑
NIQE [39]	NR-IQA (General)	0.5005	0.4053
PI [1]		0.6382	0.5320
BRISQUE [38]		0.6451	0.5573
IFQA **(Ours)**		**0.6988**	**0.6013**
SER-FIQ [46]	NR-IQA (FIQA)	0.1657	0.1386
FaceQnet-V1 [15]		0.2725	0.2106
FaceQnet-V0 [16]		0.4474	0.3813
SDD-FIQA [40]		0.5131	0.4120

the FR-IQA metrics and use NR-IQA and FIQA metrics for comparison because real-world face images have no HQ reference images. Table 2 shows that existing FIQA metrics show low correlation values except for the recently proposed state-of-the-art SDD-FIQA. SDD-FIQA metric shows a reasonable correlation value; general NR-IQA metrics have similar or higher values than FIQA metrics. Our proposed metric shows the highest correlation with human visual perception than general NR-IQA and FIQA metrics on the real-world dataset.

4.3. Qualitative Analysis

Interpretability evaluation: To prove the effectiveness and utilization of the interpretability of the proposed metric, we generated a variety of LQ images (*e.g.* blur, downsampling, and mix) from the HQ reference images based on the BFR protocol [53, 49]. Subsequently, we restored images by applying widely used general image restoration models (*e.g.* RCAN, DBPN) and face restoration models (*e.g.* DFDNet, HiFaceGAN, GPEN) to the LQ_{Mix} image, which is a combination of various degradation factors.

The interpretability maps of IFQA are shown in Figure 6. HiFaceGAN produces plausible eyes, mouth, and teeth arrangements, but it contains several artifacts in the no-reference IWF dataset. DFDNet generates reasonable facial structures but fails to restore the details compared with other recent face restoration models. DBPN usually incurs an unnatural shape of the facial primary regions or facial contour, causing IFQA to attain low scores in those regions. RCAN results in over-smoothed restored images, which leads to an overall lower score than face restoration model results. GPEN generates the most realistic facial details compared with other models, which results in high scores in facial regions. We can confirm that the overall results are consistent with the human judgment in Figure 5. Even though our metric shows the highest con-

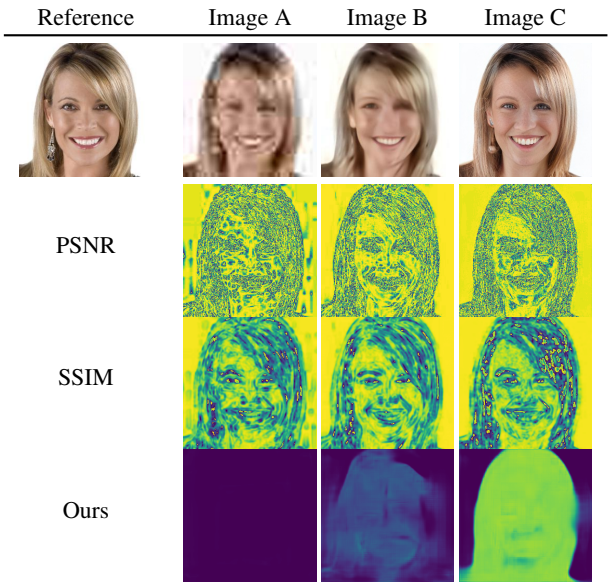

Figure 7. Comparison of the proposed metric with PSNR/SSIM with respect to pixel-level score. Bright areas indicate higher similarity to the reference image.

sistency with human judgment compared to other metrics, there are still limitations. Since our metric is learning-based on a synthetic dataset, inconsistent results could be produced on real-world data. Incorporating real-world performance degradation into the learning process remains as a future work.

Comparison with PSNR/SSIM: The proposed face-oriented metric enables pixel-level visualization, whereas general NR-IQA metrics cannot provide pixel-level scores. Traditional FR-IQA metrics, such as PSNR and SSIM, can provide pixel-level scores; however, the results are not interpretable. We compare PSNR and SSIM maps with the proposed IFQA in terms of pixel-level scores in Figure 7. The map of PSNR is obtained by $L2$ distance between the reference image and the restored images. For a clear qualitative comparison with IFQA, we reversed the distance map of PSNR and SSIM. The brighter the area, the more similar it is to the reference image. In the figure, for the severely degraded 'Image A', IFQA attains an overall low score, whereas PSNR and SSIM attain a sparse low score. For 'Image B', which is of low quality except for a tiny part of the face, IFQA scores high in these undamaged regions.

4.4. In-depth Analysis

Ablation study for main modules: We present an ablation study considering the following variants: (i) a baseline model consisting of a learnable generator and an encoder-based discriminator that outputs a single value, (ii) a model that differs only in the discriminator from the first baseline, which outputs the pixel-level score, (iii) a model with original CutMix added to the second baseline model, (iv) a model with the proposed FPRS added to the second baseline model, and (v) a model with the facial mask information added to the fourth baseline model (*i.e.* our final model).

Table 3. Ablation study of IFQA framework on FFHQ, CelebA-HQ, and IWF datasets. We report the average correlation for the entire test datasets.

Discriminator	SRCC ↑	KRCC ↑
Baseline (single-output)	0.5885	0.4840
Baseline (per-pixel)	0.4674	0.3820
Baseline (per-pixel) + CutMix [56]	0.5437	0.4420
Baseline (per-pixel) + FPRS	0.6265	0.5213
Baseline (per-pixel) + FPRS + F-Mask	**0.6694**	**0.5600**

F-Mask: facial masks using a segmentation model

Table 4. Performance comparison with respect to generator models on FFHQ, CelebA-HQ, and IWF.

Generator	Task	Parameters	SRCC ↑	KRCC ↑
GPEN [53]	FIR	pre-trained	0.4997	0.4166
DFDNet [32]	FIR		0.5391	0.4366
DBPN [12]	GIR		0.5582	0.4586
RCAN [58]	GIR		0.5711	0.4680
RCAN	GIR	learnable	0.6454	0.5480
Plain model	FIR	learnable	**0.6694**	**0.5600**

FIR: face image restoration GIR: general image restoration

Table 3 reports the quantitative results on FFHQ, CelebA-HQ, and IWF datasets. From the table, we can see the following observations. Without the proposed modules, a single-output discriminator composed of only an encoder is more similar to human judgment than a per-pixel discriminator. Although CutMix results in performance improvement, it is significantly inconsistent with human judgment compared to our IFQA model with FPRS. Moreover, adding face mask information to the baseline model results in a metric that is even closer to human perception.

Generator change analysis: We hypothesize that a trainable naive model as a generator is more suitable for learning the discriminator than pre-trained general or face restoration models. To prove this hypothesis, we compare our plain model with the four conventional approaches. In Table 4, our plain U-Net-based generator shows the highest correlation with human perception, whereas cutting-edge GPEN shows the lowest value.

Discriminator backbone analysis: Because the proposed IFQA metric depends on the trainable discriminator, we performed a comparison considering the following backbone architectures of the discriminator. Table 5 shows that all the variants of our metric are superior to most of the existing IQA metrics and our VGG-19-based model produces the best performance.

Table 5. Performance comparison with respect to the backbone of discriminators on FFHQ, CelebA-HQ, and IWF.

Discriminator	Parameters	SRCC ↑	KRCC ↑
U-Net [42]	learnable	0.6100	0.5006
U-Net + SFT [50]		0.6314	0.5253
ResNet-50 [14]		0.6311	0.5346
VGG-16 [45]		0.6365	0.5286
VGG-19 [45]		**0.6694**	**0.5600**

Figure 8. IFQA results in more challenging scenarios. Brighter area indicates the higher quality.

4.5. Further Use Case Analysis

IFQA under realistic conditions: We applied various degradations to the images from VGGFace2 [3] to validate the generalization ability of the proposed metric in more challenging scenarios. In Figure 8, IFQA provides acceptable results for most scenarios. Unlike the high-quality reference image, low-quality 'Image A' shows low pixel-level scores in entire regions. Even Black or white box occlusion as seen in ('Image B') and ('Image C') is not considered as facial degradation factors, but interestingly, the proposed metric shows reasonable results.

IFQA in face manipulation tasks: There are increasing attempts to use face images to demonstrate the superiority and effectiveness of the methodology in general image generation or image-to-image translation tasks, *e.g.* StarGANv2 [6] and U-GAT-IT [29]. These tasks commonly use conventional metrics such as FID [17] and LPIPS [57] to evaluate results. Figure 9 shows the image-to-image translation results using StarGANv2 and their visualization maps. Overall results especially for 'Output C' show the possibility that our metric designed to assess face restoration models can also be used to evaluate the results of various face-related image generation tasks.

IFQA as objective function: To evaluate the generalization ability of IFQA, we adopted it as an additional objective function in StarGAN v2, which is the state-of-the-art method for face manipulation. The proposed IFQA metric evaluates the per-pixel realness of generated images and gives feedback to the generator during the training phase.

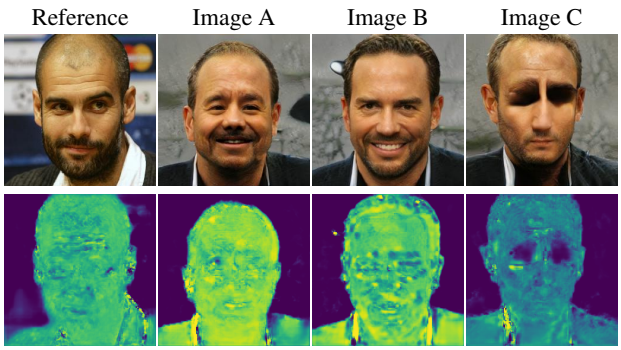

Figure 9. IFQA results in face manipulation scenarios. Brighter area indicates the higher quality.

Table 6. Quantitative StarGAN v2 performance comparison with and without our metric as an additional objective function.

Method	FID↓	
	Latent-guided	Reference-guided
StarGAN v2 w/o IFQA	14.6657	23.7138
StarGAN v2 w IFQA	**13.8008**	**22.6457**

As a result, the generator is not only trained to generate diverse styles of images but also trained to generate image realness. Table 6 clearly shows that our IFQA strategies enhance the performance of StarGAN v2 in terms of FID.

5. Conclusion

The face domain cannot be seen merely as a common object category because of facial geometry, variety of applications, and psychological evidence. Nevertheless, existing face image restoration studies have used general IQA metrics. The main finding of our study is tiny distortions in facial primary regions have a significant impact on human perception. Considering this, our framework arbitrarily swaps primary regions among low-quality, high-quality, and restored images and utilizes them as supervision for the discriminator. As a result, IFQA metric shows higher correlations with human visual perception than traditional general metrics across various architectures and scenarios.

Acknowledgement

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1054569, No. 2022R1A4A1033549). This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)). We thank Eunkyung Jo for helpful feedback on human study design and Jaejun Yoo for constructive comments on various experimental protocols.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, pages 6228–6237. Computer Vision Foundation / IEEE, 2018.

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). In *ICCV*, pages 1021–1030. Computer Vision Foundation / IEEE, 2017.

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

[4] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *CVPR*, pages 11896–11905. Computer Vision Foundation / IEEE, 2021.

[5] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8185–8194. Computer Vision Foundation / IEEE, 2020.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019.

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[9] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020.

[10] M. J. Farah, K. L. Levinson, and K. L. Klein. Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33(6):661–674, 1995.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.

[12] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, pages 1664–1673. Computer Vision Foundation / IEEE, 2018.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988. Computer Vision Foundation / IEEE, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. Computer Vision Foundation / IEEE, 2016.

[15] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *arXiv preprint arXiv:2006.03298*, 2020.

[16] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *ICB*, pages 1–8. IEEE, 2019.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637. Curran Associates Inc., 2017.

[18] R. Hogg, J. Mckean, and A. Craig. *Introduction to Mathematical Statistics*. Pearson, 2005.

[19] Sungeun Hong and Jongbin Ryu. Unsupervised face domain transfer for low-resolution face recognition. *IEEE Sign. Process. Letters*, 27:156–160, 2019.

[20] Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *ICPR*, pages 2366–2369. IEEE Computer Society, 2010.

[21] Q. Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[22] Woobin Im, Sungeun Hong, Sung-Eui Yoon, and Hyun S. Yang. Scale-varying triplet ranking with classification loss for facial age estimation. In *ACCV*, pages 247–259. Springer, 2018.

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[24] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997.

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*. OpenReview.net, 2018.

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019.

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116. Computer Vision Foundation / IEEE, 2020.

[28] M. G. Kendall, A. Stuart, and J. K. Ord. *Kendall's Advanced Theory of Statistics*. Oxford University Press, Inc., 1987.

[29] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*. OpenReview.net, 2020.

[30] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P. Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. In *HVEI*, pages 1–6. Ingenta, 2016.

[31] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pages 5548–5557. Computer Vision Foundation / IEEE, 2020.

[32] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, pages 399–415. Springer, 2020.

[33] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, pages 272–289. Springer, 2018.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[35] Farkhod Makhmudkhujaev, Sungeun Hong, and In Kyu Park. Re-aging gan: Toward personalized face age transformation. In *ICCV*, pages 3908–3917, 2021.

[36] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2813–2821. Computer Vision Foundation / IEEE, 2017.

[37] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, pages 2434–2442. Computer Vision Foundation / IEEE, 2020.

[38] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. Blind/referenceless image spatial quality evaluator. In *ACSCC*, pages 723–727. IEEE, 2011.

[39] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2013.

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

[41] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351, pages 234–241. Springer, 2015.

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[44] Edgar Schönfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *CVPR*, pages 8204–8213. Computer Vision Foundation / IEEE, 2020.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*. MIT Press, 2015.

[46] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *CVPR*, pages 5650–5659. Computer Vision Foundation / IEEE, 2020.

[47] D. Y. Tsao and M. S. Livingstone. Mechanisms of face perception. *Annu Rev Neurosci*, 31(1):411–437, 2008.

[48] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, pages 2744–2754. Computer Vision Foundation / IEEE, 2020.

[49] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, pages 9168–9178. Computer Vision Foundation / IEEE, 2021.

[50] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615. Computer Vision Foundation / IEEE, 2018.

[51] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[52] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *ACM MM*, pages 1551–1560. ACM, 2020.

[53] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. GAN prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681. Computer Vision Foundation / IEEE, 2021.

[54] Jaejun Yoo, Namhyuk Ahn, and Kyung-Ah Sohn. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *CVPR*, pages 8372–8381. Computer Vision Foundation / IEEE, 2020.

[55] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, pages 318–333. Springer, 2016.

[56] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6022–6031. Computer Vision Foundation / IEEE, 2019.

[57] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595. Computer Vision Foundation / IEEE, 2018.

[58] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 294–310. Springer, 2018.

[59] Yang Zhang, Ivor W. Tsang, Yawei Luo, Chang-Hui Hu, Xiaobo Lu, and Xin Yu. Copy and paste GAN: face hallucination from shaded thumbnails. In *CVPR*, pages 7353–7362. Computer Vision Foundation / IEEE, 2020.

[60] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, pages 614–630. Springer, 2016.