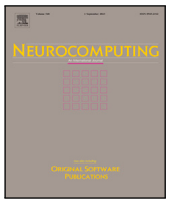
Neurocomputing 614 (2025) 128780

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Perceptual metric for face image quality with pixel-level interpretability

Byungho Jo [a], In Kyu Park [a,b], Sungeun Hong [c,*]

[a] Inha AI Convergence Research Center, Inha University, Incheon, South Korea
[b] Department of Electrical and Computer Engineering, Inha University, Incheon, South Korea
[c] Department of Immersive Media Engineering, Sungkyunkwan University, Seoul, South Korea

ARTICLE INFO

Communicated by L. Celona

Keywords:
Face image quality assessment
Face restoration
Evaluation metric
Vision transformer
Facial primary regions

ABSTRACT

This paper tackles the shortcomings of image evaluation metrics in evaluating facial image quality. Conventional metrics do neither accurately reflect the unique attributes of facial images nor correspond with human visual perception. To address these issues, we introduce a novel metric designed specifically for faces, utilizing a learning-based adversarial framework. This framework comprises a generator for simulating face restoration and a discriminator for quality evaluation. Drawing inspiration from facial neuroscience studies, our metric emphasizes the importance of primary facial features, acknowledging that minor changes in the eyes, nose, and mouth can significantly impact perception. Another key limitation of existing image evaluation metrics is their focus on numerical values at the image level, without providing insight into how different areas of the image contribute to the overall assessment. Our proposed metric offers interpretability regarding how each region of the image is evaluated. Comprehensive experimental results confirm that our face-specific metric surpasses traditional general image quality assessment metrics for facial images, including both full-reference and no-reference methods. The code and models are available at https://github.com/AIM-SKKU/IFQA.

1. Introduction

Over the years, numerous face-related studies have been conducted due to their wide-ranging applications and significant impact [1–3]. Face image restoration (FIR) aims to recover high-quality face images from low-quality versions that have been degraded by factors such as low-resolution, compression, and blur. FIR approaches have advanced significantly, addressing the critical need for enhancing visual quality in applications ranging from security surveillance to social media. This is particularly crucial in the context of blind face restoration (BFR), where the specific nature of degradation — be it due to noise, resolution loss, or compression artifacts — is unknown, making the restoration process inherently more complex. In evaluating the performance of FIR methods, most existing studies rely on general full-reference image quality assessment (FR-IQA) metrics such as PSNR [4], SSIM [5], and LPIPS [6], which measure the similarity between reference and restored images. For BFR, where the type of degradation is unknown, no-reference image quality assessment (NR-IQA) metrics like NIQE [7] and BRISQUE [8] are used.

Critically, existing general metrics widely used in face-related tasks [9–11] neglect the specific characteristics of faces, which can lead to differences between metric scores and human visual perception, as shown in Fig. 1. All thirty participants ranked Image A as more realistic than Image B. The details of the protocol for human evaluation are provided in the human study procedure described in Section 4. Consequently, recent face restoration studies have utilized human evaluations rather than evaluation metrics to evaluate the proposed method. Unfortunately, human study-based assessments in the FIR field suffers from scalability issues, variances among human assessors, and high evaluation costs. The absence of appropriate metrics for evaluating FIR results in significant expenses and time-consuming evaluations, presenting a major obstacle to the advancement of the FIR field.

For these reasons, we have become interested in designing face-oriented metrics. In general, designing specific image quality metrics for various object categories is unfeasible due to their extensive diversity. However, the distinct characteristics of the face domain, along with the requirements of several downstream tasks, necessitate a targeted approach for evaluating the quality of face images. The need for face-specific metrics is further supported by the particular attributes of faces, such as their geometry and textures, highlighted in studies. This approach aligns with an early psychological finding [12], which identified brain areas exclusively involved in face recognition, emphasizing the specialized processing that faces receive in the human brain.

In this study, we introduce Interpretable Face Quality Assessment (IFQA), a novel face-centered metric. Our work is inspired by learning-based perceptual metrics, which are notably more aligned with human

* Corresponding author.
E-mail address: csehong@skku.edu (S. Hong).

https://doi.org/10.1016/j.neucom.2024.128780
Received 17 May 2024; Received in revised form 28 August 2024; Accepted 19 October 2024
Available online 28 October 2024

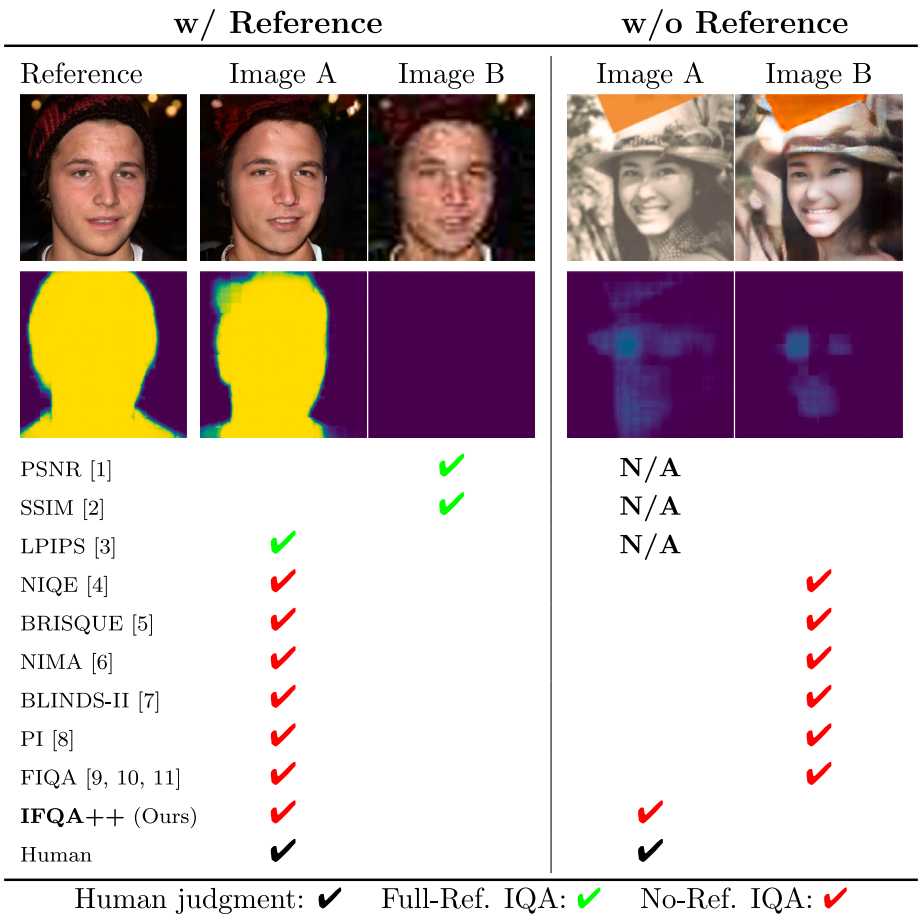


Fig. 1. Which of 'Image A' or 'Image B' is closer to the given reference or looks high-quality? Conventional full-reference metrics and no-reference metrics are mostly inconsistent with human judgment. LPIPS agrees with human judgment but cannot be applied to the scenario without references. Our IFQA++ is consistent with human judgment and offers interpretability maps, where brighter regions indicate higher quality as in the second-row images.

Fig. 2. The first and third columns present the original facial images, while the second and fourth columns show faces with primary regions rearranged randomly. This demonstrates the brain's focused response to these key facial areas in recognition, as emphasized by neuroscience research. The purpose of rearranging these features was to explore how changes in their spatial arrangement affect face perception. Research [13] has highlighted the importance of these facial features in perception, supporting our face-oriented metric.

judgment in evaluating image quality compared to conventional metrics such as PSNR and SSIM. General perceptual metrics are a type of full-reference image quality assessment (FR-IQA) that is learned from general images. To measure the image quality, this approach utilizes the distance between features derived from a pre-trained model for both the reference and query images. The assumption that FR-IQA requires a high-quality image as a reference image may be impractical in real-world conditions. In many cases 'in the wild,' one might only encounter low-quality images and not have the opportunity to compare them with high-quality counterparts that would serve as a ground truth. Considering this, our metric introduces a novel no-reference image quality assessment (NR-IQA) method, focusing on the distinct features of the face.

Specifically, our metric leverages insights from neuroscience, particularly Liu et al. [13]'s findings as described in Fig. 2. Our approach emphasizes the brain's unique reactions to facial elements. These studies underscore the human brain's selective attention to key facial regions — eyes, nose, and mouth — during recognition tasks. Thus, our metric is crafted to concentrate on these essential areas of the face. By doing so, it offers a detailed image quality assessment that mirrors human perception of faces, enhancing both accuracy and a focus on human-like understanding. Our method stands out by providing comprehensive IQA scores and pixel-specific, interpretable ratings, thanks to its U-shaped architecture. This goes beyond traditional IQA scores like those from PSNR [4], SSIM [5], or NIQE [7], offering deeper insights into image quality. This feature not only enriches the metric's informativeness but also its readability and utility in grasping image quality nuances.

Our face-oriented metric also differs from existing face image quality assessment (FIQA) methods [14–16] that heavily rely on face recognition systems. Our metric is more general and independent of specific

high-level tasks. The key idea of our approach assigning significant weight to facial primary regions has not been addressed in existing FIQA methods, and we demonstrate its effectiveness through extensive ablation studies. We also address the need for a face-oriented metric that previous face restoration studies have yet to resolve, despite the demands raised. Through evaluations across different architectures and scenarios, our metric exhibits higher correlations with human perception compared to both IQA (full and no-reference) and state-of-the-art FIQA metrics.

To further improve the evaluation performance, we present IFQA++, which serves as an extended version of our previous work [17]. This extended version stands out by incorporating the vision transformer architecture [18], renowned for its strong performance in diverse visual tasks. The self-attention mechanism [19–21] of the vision transformer captures the full context of images, enabling the learning of more nuanced representations. Utilizing pre-trained weights from masked-autoencoding-based self-supervised learning [22] allows our model to capture meaningful visual representations with less need for large training datasets. We have expanded our evaluation to include a wider range of NR-IQA methods, providing a comprehensive analysis supported by experiments and a human study.

With IFQA++, we observe significant improvements compared to our earlier work, which are further detailed and explained in the experiment section. Additional experiments have been conducted to explore different use cases, each with a thorough analysis of the results. In summary, our main contributions are as follows:

- We introduce IFQA++, a face-specific image quality metric. This metric particularly focuses on the significance of key facial features such as the eyes, nose, and mouth, underscoring their importance in the evaluation process.
- Our no-reference metric not only outperforms existing no-reference and full-reference metrics but also surpasses state-of-the-art FIQA metrics. It demonstrates a higher correlation with human judgment.
- By offering pixel-level evaluation scores, our approach provides a more interpretable analysis of image quality. This method moves beyond the constraints associated with metrics that rely solely on a single score.
- Our metric is designed for easy application, and we commit to fully disclosing all aspects related to the evaluation metric. This openness ensures that our metric can be readily used across various face-related tasks.

2. Related work

2.1. Face image restoration

Unlike general image restoration techniques, FIR methods have been developed by leveraging facial prior knowledge. Recent advancements in deep convolutional neural networks (CNNs) and generative adversarial networks (GANs) have led to the proposal of deep models for tasks such as super-resolution (SR) [11,23] and denoising [24].

For instance, Zhang et al. [23] proposed a GAN-based small face restoration framework designed to incorporate contextual information around the face regions, specifically for use in face detection scenarios. Wei et al. [11] proposed a multi-order head attention framework that models the relationship of facial components by incorporating facial landmarks and component heatmaps as prior information. However, despite their promising results on artificially degraded images, these methods often exhibit poor performance when applied to real-world low-quality (LQ) images.

Extensive efforts have been dedicated to tackling the challenges posed by BFR scenarios, a task focused on restoring low-quality (LQ) face images suffering from complex and unknown degradation [25]. These methods heavily rely on accurate facial prior knowledge to guide the restoration process. Therefore, despite their initial promise, these approaches often fall short when it comes to recovering real-world degraded details, particularly in scenarios where prior information is not available. In recent years, GAN inversion techniques based on the popular StyleGAN architecture [26] have gained significant attention within the field of face restoration. These techniques leverage the power of generative adversarial networks to invert the image synthesis process, enabling the restoration of high-quality facial images. Similarly, diffusion models [27] have also emerged as a powerful approach in image generation and restoration tasks. By iteratively refining noise into coherent images, diffusion models have demonstrated remarkable performance in producing high-quality results, making them an increasingly popular choice in recent advancements in face restoration. [28,29]

Despite significant advances in FIR methodologies, the evaluation metrics used are typically borrowed from general image restoration, which do not adequately capture facial characteristics. Consequently, state-of-the-art methods often resort to costly human studies to demonstrate the superiority of their model, as observed in a study [30]. This study is motivated by the limitations identified in previous face restoration studies and aims to introduce a novel metric specifically designed for faces. This face-oriented metric is developed to overcome the challenges that have not been adequately addressed in the facial image restoration (FIR) research area.

2.2. General image quality assessment

Image quality assessment (IQA) is a field dedicated to measuring the perceptual quality of images. Existing IQA approaches can be broadly classified into two categories: full-reference (FR-IQA) and no-reference (NR-IQA). FR-IQA evaluates the similarity between restored and reference images based on statistical or perceptual measures. Traditionally, metrics such as PSNR [4] and SSIM [5] have been used to quantitatively evaluate face image restoration (FIR) models when reference images are available. PSNR computes the mean squared error between pixels, while SSIM measures structural similarity in terms of luminance, contrast, and structure information. In an attempt to bridge the semantic gap between FR-IQA and human judgment, perceptual metric like LPIPS [6] have been introduced.

LPIPS calculates the distance between restored and ground truth images in a feature space using a pre-trained network. Moreover, recent FR-IQA works have proposed a hybrid manner which combines CNNs and transformer-based methods [31]. Combined features interact in a mutually complementary manner, enhancing the overall effectiveness of the assessment. While these metrics are reasonable choices for evaluating restoration results, they are not applicable in real-world scenarios where reference images for comparison are unavailable

To address such real-world scenarios, several NR-IQA methods have been proposed [7,32]. Although NR-IQA has been successfully applied to natural scenes, attempts to apply it to BFR problems without reference facial images have revealed limitations in general NR-IQA metrics for evaluating restored facial images. Consequently, to overcome these limitations, several BFR studies have resorted to human evaluations to assess the success of restoration efforts. The absence of suitable evaluation metrics for face restoration has significant cost implications and poses a major obstacle in the field of face image restoration. To address this critical issue, we propose an evaluation metric specifically designed to focus on facial primary regions. The proposed metric is not only more consistent with human visual perception than conventional general metrics but also provides interpretable visualization.

2.3. Face image quality assessment

Face image quality assessment (FIQA) plays a crucial role in supporting face recognition systems by determining whether low-quality face images should be discarded as a preprocessing step especially

Fig. 3. 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'. For the images within row 'Image A', PSNR/SSIM scores are (27.30/0.8359), (26.10/0.8042), (28.53/0.8019), (21.78/0.5546), respectively. For the images within row 'Image B', PSNR/SSIM scores are (24.36/0.7439), (25.10/0.7353), (26.54/0.7011), (20.76/0.5164), respectively.

in challenging real-world scenarios like surveillance and outdoor environments. Additionally, Fu et al. [33] demonstrated the necessity of employing face-specific quality metrics, as general image quality measures may not adequately address the unique requirements of face recognition systems. Traditional FIQA approaches, ranging from analytics-based to modern learning-based methods [14,34], have primarily been developed with a focus on face recognition tasks. While FIQA has shown remarkable results in the context of face recognition, it is crucial to recognize its limitations when applied to Face Image Restoration (FIR). FIR presents unique challenges that are not adequately addressed by existing FIQA models, as these models are tailored for face recognition rather than restoration. Chen et al. [35] also presents biometric FIQA methods produces unsatisfied assessment results on perceptual quality. Consequently, FIR requires a distinct approach that considers the particularities of restoring primary facial regions, not just evaluating the overall face quality.

To the best of our knowledge, the proposed IFQA is the first face-centric evaluation metric that focuses on the primary facial regions rather than the entire face area. Through empirical evaluations, we demonstrate that the proposed metric significantly outperforms all baseline FIQA approaches, underscoring its effectiveness in face image quality assessment.

3. Proposed metric

3.1. Pilot study

In our preliminary experiments on evaluating face restoration results, we found that the primary facial regions (*e.g.*, eyes, nose, mouth, etc.) are crucial for human visual perception. As shown in Fig. 3, the majority of participants in our study stated that the images from 'Image B' (third row) are more realistic than those in the second row 'Image A'. However, PSNR and SSIM metrics assess that images in 'Image A' are more similar than in 'Image B' with the reference images. Human visual perception is notably influenced by the overall structure and distortions in primary facial regions as demonstrated by neuroscience research [13]. We argue that this finding is critical in FIR, emphasizing the importance of these key facial areas.

3.2. Proposed framework

Motivated by the observation in the pilot study, we introduce an evaluation metric that considers primary facial regions. We utilize the GAN architecture, which consists of two models: generator and discriminator. These models are trained in an adversarial manner, where the generator aims to produce realistic images to fool the discriminator, and the discriminator is trained to distinguish whether images are real or generated by the generator. After training process, the discriminator is leveraged as a metric function, assessing the realism of given facial images. Fig. 4 presents the overall framework.

Generator for image restoration: The generator has a simple encoder–decoder architecture that can be considered a plain face restoration model that outputs restored face images. Specifically, the layers of both the encoder and decoder consist of CNN-based residual blocks. We use the average pooling in down-sampling blocks and nearest-neighbor interpolation in up-sampling blocks. As the final output activation function, we use the hyperbolic tangent.

We train our generator to restore LQ images to 256 × 256 HQ images. During the training phase, we deliberately corrupt HQ images in the FFHQ dataset [26] 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 [30, 36], 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 is designed to evaluate the quality of query images by training in an adversarial manner with the generator. We design our discriminator to output per-pixel scores and to enhance the generalization ability using encoder and decoder architecture. This architecture allows us to distinguish the regions from HQ images as 'real' while the regions from LQ and restored images as 'fake'.

Our IFQA++ framework adopts the vision-based transformer architecture, which is composed of both encoder and decoder components. This architecture first processes an input image X by dividing it into N patches $\{x_1, x_2, \ldots, x_N\}$, each of size is 16 × 16. Then the patches are linearly projected to 768-dimensional embeddings by w. With the position embedding tensor (e_{pos}) which is designed to retain the spatial information of patches, the combined embeddings (z_0) are fed to the transformer encoder as follows:

$$\begin{aligned} z_0 &= xW + e_{pos}, \\ z_l &= \text{Encoder}_l(z_{l-1}) \end{aligned} \tag{2}$$

for $l = 1, \ldots, L(= 32)$, where $\text{encoder}_l(\cdot)$ in each layer encapsulates the operations of self-attention mechanisms. After processing by the encoder, the latent representation z is conveyed to the decoder as follows:

$$\text{ScoreMap} = \text{Sigmoid}(\text{Decoder}(z)), \tag{3}$$

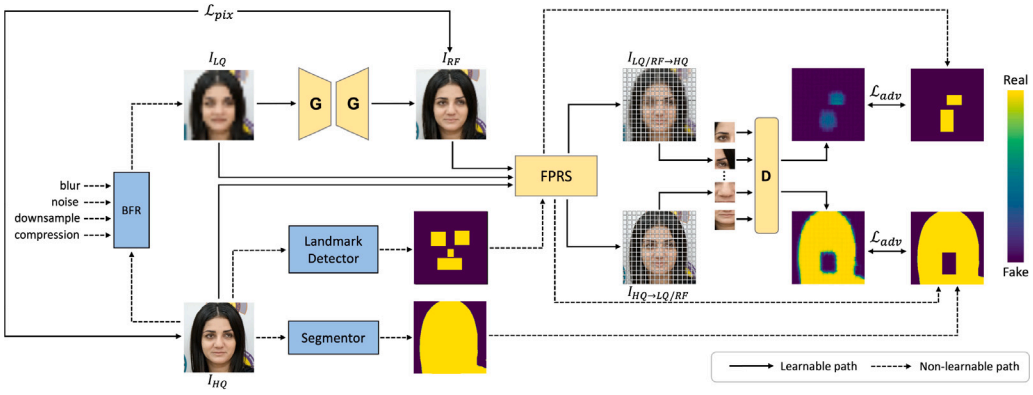


Fig. 4. IFQA++ framework outline. Given HQ images, we obtain LQ images via BFR formulation. The generator (G) mimics face restoration models, while the discriminator (D) is used to evaluate image quality by determining high-quality regions as 'real' and low-quality or restored regions as 'fake'. Through vision transformer-based U-Net architecture, the discriminator is able to evaluate the image pixel-by-pixel from the patched input. FPRS allows the proposed metric to give more weight to facial primary regions that have a significant impact on human visual perception. Once the IFQA++ framework is trained, we leave only the per-pixel discriminator and use it as a metric function.

Utilizing a transformer architecture similar to the encoder, the decoder consists of eight attention block layers. The decoder aims to produce a score map with the same spatial resolution as the original image, but with a single channel. The outcome of the decoder is a score map for pixel-level evaluation, achieved by applying a Sigmoid function to express pixel-wise scores effectively. The final evaluation is determined by averaging the score map values, resulting in a scalar value that represents the overall assessment.

To enhance the performance of the encoder and decoder within the IFQA++ framework, we utilize the pre-trained weights derived from a masked-auto-encoder (MAE) based on self-supervised learning approach [22]. The MAE utilizes ViT [18]-based encoder–decoder architecture, where the encoder processes only a subset of patches from the masked image, and produces latent features. Subsequently, the decoder reconstructs the original image from the encoded features. The self-supervised learning approach of the MAE significantly refines the encoder by training it to predict missing parts of an image without human labeled information. This method enhances the model's capability of extracting feature representations and robustness. The enhanced encoder demonstrates significant versatility in a wide range of tasks *e.g.*, image classification and object detection. Furthermore, they demonstrate efficiency in generalizing from learned dataset, thus mitigating the risk of overfitting. By leveraging model's generalized representations, the encoder requires minimal labeled data for fine-tuning across various tasks.

We adopt pretrained ViT huge model in our IFQA++ framework. In the decoder, we modify the last prediction layer by replacing the three output channels with a single channel and applying the Sigmoid function.

During the training process, the learnable weights of the encoder and the position embedding weights remain fixed. This setup allows us to derive a single image-level quality score by aggregating the pixel-level scores. In contrast to traditional discriminators in adversarial training, we provide 'fake' supervision to both the input (LQ images) and output (RF images) of the generator. Instead of considering the entire high-quality image, we focus solely on the face region as the 'real' label. Our ablation study results demonstrate that this face region-focused approach yields higher correlations with human perception compared to using the global region as the 'real' label. To obtain binary facial masks from the images, we utilize an off-the-shelf face segmentation model, which has been pretrained on the CelebAMask-HQ dataset [37].

We also design a novel augmentation technique referred to as facial primary regions swap (FPRS) to reflect facial characteristics to the proposed metric, as described in Fig. 5. We first utilize a pretrained landmark detector [38] to obtain facial primary regions from HQ images. Unlike other augmentation techniques, such as original CutMix, we utilize RoIAlign to obtain the regions of facial primary components. Therefore, selected facial primary regions from the LQ or RF images are swapped with the HQ images. Let $I_{LQ/RF}$ represent an LQ or RF image, and I_{HQ} represent an HQ image. Through the FPRS operation, we generate a new image pair for the discriminator during training as follows:

$$\begin{aligned} I_{HQ\to LQ/RF} &= \mathbf{M_{FPRS}} \odot I_{HQ} \\ &\quad + (\mathbf{1} - \mathbf{M_{FPRS}}) \odot I_{LQ/RF} \\ I_{LQ/RF\to HQ} &= \mathbf{M_{FPRS}} \odot I_{LQ/RF} \\ &\quad + (\mathbf{1} - \mathbf{M_{FPRS}}) \odot I_{HQ}, \end{aligned} \tag{4}$$

where $\mathbf{M_{FPRS}} \in \{0,1\}^{H\times W}$ is a randomly selected binary mask from facial primary regions. $\mathbf{1}$ is a binary mask filled with ones. $\odot$ indicates element-wise multiplication. When employing the facial primary regions swap (FPRS) method, we anticipate that the model will exhibit heightened sensitivity towards unrealistic shapes in the essential facial regions of the input image.

Objective function: We adopt least-square-based adversarial learning [39] between the generator and discriminator for training the IFQA++ framework. The generator is trained to fool the discriminator, and the objective function of the generator is defined as follows:

$$\mathcal{L}_{adv,G} = \mathbb{E}_{I_{RF}}[(D^U(I_{RF}) - \mathbf{1})^2], \tag{5}$$

where $D^U(\cdot)$ refers to encoder–decoder discriminator which has U-Net-shaped that produces per-pixel scores maps. We also adopt pixel-wise distance loss to enforce the generator to recover I_{RF} to be similar to the corresponding HQ image. The pixel-wise distance loss compares all of the pixel values between the I_{RF} and HQ images as follows:

$$\mathcal{L}_{pix} = \mathbb{E}_{I_{RF},I_{HQ}}[\|I_{RF} - I_{HQ}\|_2]. \tag{6}$$

The objective function for the discriminator is defined as follows:

$$\begin{aligned} \mathcal{L}_{adv,D} &= \mathbb{E}_{I_{HQ}}[(D^U(I_{HQ}) - \mathbf{M_{FACE}})^2] \\ &\quad + \mathbb{E}_{I_{LQ},I_{RF}}[D^U(I_{LQ/RF})^2], \end{aligned} \tag{7}$$

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:

$$\min_G \max_D \quad \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{pix}, \tag{8}$$

where $\mathcal{L}_{adv} = \mathcal{L}_{adv,G} + \mathcal{L}_{adv,D}$, and λ_1 is scaling parameters. We empirically set λ_1 as 5.

Assessment protocol: We only use the per-pixel discriminator for image quality assessment after the IFQA++ framework training is done.

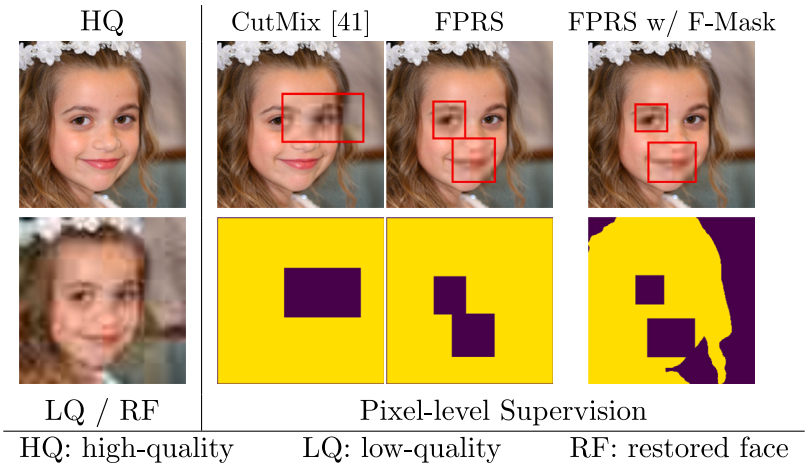


Fig. 5. Supervision for IFQA++ metric. Regions from high-quality images provide 'real' labels (yellow), while regions from low-quality or restored face images give 'fake' labels (purple). The red box indicates the randomly selected swapped region. Note that our approach utilizes estimated facial region masks and swaps only between facial primary regions, *e.g.*, eyes, nose, and mouth, via the proposed FPRS.

The pixel-level assessment score by the discriminator allows us to perform an interpretable in-depth analysis. To obtain an image-level quality score from the given input image I, we average every pixel-level score from the per-pixel discriminator as:

$$QS = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} D^{U}_{i,j}(I) \tag{9}$$

4. Experiments

4.1. Implementation details

Datasets: The IFQA++ metric intrinsically should not be highly dependent on a specific dataset. We separate the training dataset for the learning-based metric and the test dataset to prove the generalization of our metric. For training the IFQA++ framework, we randomly selected 20,000 images from the FFHQ [26] dataset. Additionally, we constructed three types of benchmark test datasets. We first construct a test set by combining high-quality images with considerable variations, including those from FFHQ and CelebA-HQ [40], which are mainly used in face restoration tasks. Based on the data split, the images used for training and those used for testing do not overlap in the FFHQ dataset. 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) [30], which is widely used for BFR problems. Notably, the IWF dataset solely provides low-quality images without high-quality reference images. Fifty images each from the FFHQ and CelebA-HQ datasets, along with 100 images from the IWF dataset, were randomly sampled to create a test dataset, which is utilized for the human study. The selected low-quality facial images, whether generated or naturally occurring, were restored using five different Image Restoration models and subsequently utilized in the human study. The detailed procedure is explained in Section 4.2, under Human Study Procedure. Additionally, we combined the FFHQ, CelebA-HQ, and IWF datasets for an ablation study to demonstrate the robustness and generalization ability of our proposed metric.

Image Restoration models: We used various image restoration models including general image restoration models (*e.g.*, RCAN [41], DBPN [42]) and face restoration models (*e.g.*, HiFaceGAN [25], DFDNet [36], GPEN [30]) to evaluate the proposed metric quantitatively and qualitatively. For a fair comparison, we used each model's official training code and hyperparameters. Because DFDNet did not release the training code, we trained DFDNet using the same training protocol as much as possible based on the descriptions of the paper. Additionally, we used the state-of-the-art GPEN model pre-trained on FFHQ. The GPEN provides the pre-trained model, which restores LQ images to 512 × 512 HQ images. To handle this issue, we restore LQ images using the pre-trained model and then down-sample the restored images to 256 × 256.

Comparative IQA metrics: We used a series of assessment metrics commonly used in face restoration tasks: three traditional FR-IQA metrics (*e.g.*, PSNR [4], SSIM [5], MS-SSIM [43]), perceptual FR-IQA metric such as LPIPS [6]), three conventional NR-IQA metrics (*e.g.*, BRISQUE [8], BLINDS-II, NIQE [7]), and perceptual NR-IQA metrics such as PI [32] and NIMA [44]). Furthermore, while FIQA is not explicitly tailored for the FIR task, we employ it as a comparative benchmark. It emphasizes the critical need for a face-specific metric dedicated to the FIR field. We compared with four state-of-the-arts FIQA methods (*e.g.*, FaceQnet-V0 [45], FaceQnet-V1 [14], SER-FIQ [34], and SDD-FIQA [15]). FaceQnets are trained with VGGFace2 dataset [46] and the input resolution is set to 224 × 224. SER-FIQ and SDD-FIQA take 112 × 112. We use the implemented codes provided by the original authors.

4.2. Quantitative analysis

Human study procedure: To compare the proposed metric to existing IQA metrics quantitatively, we conducted a human study of ranking the realistic facial images from given images. Firstly, we prepared 200 face image samples. Each sample involves six images, consisting of LQ images and restored results from SISR models and FIR models. We constructed two types of test sets using CelebA-HQ with FFHQ for the full-reference (FR) scenario and used IWF for the no-reference (NR) scenario. As a result, 200 samples were randomly selected from the FR and NR sets for our survey. Note that the number of images from the FR set is 100 and that of the NR set is 100. To prevent participants from being biased toward ground truth (GT) images (i.e. high-quality reference images), we excluded GT images from the FR set in our survey. We also shuffled the display order of the images for each sample in the survey question. Overall, the total number of images used for human study is 1,200.

To gather the human responses systematically, we first created a survey using the survey software tool and integrated the survey with Amazon Mechanical Turk (AMT) for crowdsourcing. We asked participants from AMT to rank given samples from closest to furthest to a realistic human face. 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

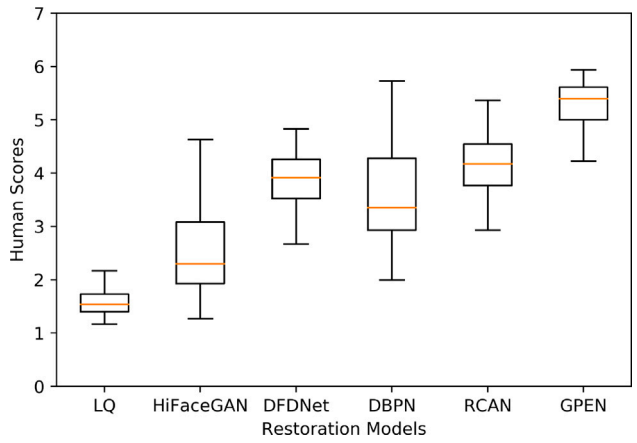


Fig. 6. Box plot of restoration models through our human study.

Table 1
Comparative analysis with NR-IQA on FFHQ and CelebA-HQ.

Metric	Type	SRCC ↑	KRCC ↑
NIQE [7]	NR-IQA (General)	0.2668	0.2039
NIMA [44]		0.3891	0.3026
PI [32]		0.4125	0.3173
BRISQUE [8]		0.4405	0.3373
BLINDS-II [47]		0.4634	0.3480
IFQA [17]		0.6400	0.5186
IFQA++ (Ours)		**0.7822**	**0.6813**
SER-FIQ [34]	NR-IQA (FIQA)	0.3554	0.2706
FaceQnet-V1 [14]		0.4560	0.3453
FaceQnet-V0 [45]		0.5491	0.434
SDD-FIQA [15]		0.5920	0.4840

of responses across whole samples was 6,000. The final rank of each sample is calculated by the weighted average scores as follows:

$$score = \frac{1}{N}\sum_{i=1}^{6} w_i r_i, \tag{10}$$

where N refers to the number of total raters, r_i represents the rank for ith image {1, 2, 3, 4, 5, 6}, and w_i is corresponding weights {6, 5, 4, 3, 2, 1}. As a result, each image is ranked in descending order of the calculated scores. Fig. 6 presents the box plot of our human study results.

In the overall rank sequence, the models are arranged as follows: GPEN, RCAN, DFDNet, DBPN, HiFaceGAN, and LQ. GPEN shows a narrow box while holding the highest values among the models. Both DBPN and HiFaceGAN present extended whisker lengths in the box plots, with HiFaceGAN showing significantly longer upper whiskers compared to the lower ones. In contrast, LQ's box is narrow and indicates the lowest values.

Analysis on FFHQ and CelebA-HQ: To quantitatively analyze the human ranking responses for each sample, we measured Spearman's rank order correlation coefficient (SRCC) and Kendall's rank order correlation coefficients (KRCC). These approaches are widely used to assess the correlation between human judgments and other metrics in various fields.

We conducted a comparative study of FFHQ and CelebA-HQ using existing no-reference image quality Assessment (NR-IQA) metrics and the state-of-the-art face image quality assessment (FIQA) metrics as shown in Table 1. Notably, our IFQA++ performance shows significant improvement over our previous work in both SRCC and KRCC. Interestingly, the widely utilized NR-IQA metric NIQE demonstrated the weakest correlation. We also compare with recent FIQA methods specifically designed for face recognition. Our IFQA family outperforms existing FIQA metrics. The results find that our face-oriented metric is consistently more aligned with human perception compared to existing NR-IQA metrics.

Table 2
Quantitative comparative analysis with FR-IQA metrics on FFHQ and CelebA-HQ. Note that our metric does not require reference image for image quality assessment.

Metric	Reference	SRCC ↑	KRCC ↑
IFQA [17]	NR-IQA	0.6400	0.5186
IFQA++ (Ours)		**0.7822**	**0.6813**
PSNR [4]	FR-IQA	0.1011	0.0893
SSIM [5]		0.1885	0.1999
MS-SSIM [43]		0.3571	0.2799
LPIPS [6]		0.6685	0.5560

Table 3
Comparative analysis on IWF (In the Wild Face).

Metric	Type	SRCC ↑	KRCC ↑
BLINDS-II [47]	NR-IQA (General)	0.4931	0.4147
NIQE [7]		0.5005	0.4053
NIMA [44]		0.6114	0.5066
PI [32]		0.6382	0.5320
BRISQUE [8]		0.6451	0.5573
IFQA [17]		0.6988	0.6013
IFQA++ (Ours)		**0.7062**	**0.6066**
SER-FIQ [34]	NR-IQA (FIQA)	0.1657	0.1386
FaceQnet-V1 [14]		0.2725	0.2106
FaceQnet-V0 [45]		0.4474	0.3813
SDD-FIQA [15]		0.5131	0.4120

To demonstrate the general applicability of our proposed NR-IQA metric, we conducted a comparative analysis with FR-IQA. Critically, all FR-IQA metrics are not suitable for real-world scenarios where reference images are not available. We select five frequently utilized FR-IQA metrics (PSNR, SSIM, MS-SSIM, and LPIPS) in FIR for comparison. The comparison results are shown in Table 2.

The traditional FR-IQA metrics, PSNR and SSIM, exhibit low correlations with human perception in both SRCC and KRCC. The MS-SSIM metric demonstrates relatively better performance than PSNR and SSIM, yet still shows a low correlation compared to the IFQA++. The LPIPS perceptual metric shows better performance with a large margin for both SRCC and KRCC than PSNR, SSIM, and previous IFQA metrics. Notably, our proposed IFQA++ achieved the highest correlation with human perception among NR-IQA metrics and also FR-IQA metrics, despite not requiring reference images. This significant improvement underscores the effectiveness of our approach, particularly when applied to widely used datasets.

Analysis on a real-world dataset: To demonstrate the generalization capability of the assessment, we measure the correlation value on real-world IWF dataset [30] that only provides low-quality images without high-quality reference images. Because of the absence of HQ reference images, we leverage NR-IQA and FIQA metrics for comparison while excluding FR-IQA metrics.

Table 3 shows that existing FIQA metrics show low correlation values. SDD-FIQA metric shows a reasonable correlation value; general NR-IQA metrics have similar or higher values than FIQA metrics. Our IFQA++ demonstrated the highest correlation with human visual perception among general NR-IQA and FIQA metrics on the real-world dataset. Although the improvements in SRCC and KRCC might appear modest, it is important to note that these enhancements were achieved on challenging 'in the wild' face images.

Analysis of Resource Usage and Complexity: ViTs architecture utilizes more power and computational resources than traditional CNNs. We conducted a detailed analysis of increased resource usage and complexity of the ViT-based IFQA++ model compared to the CNN-based IFQA model, providing insights into the trade-offs and performance benefits of using ViTs. For the experiments, we utilize a NVIDIA A6000 GPU (48 GBs).

Table 4
Comparative analysis of computational complexity between IFQA++ (ViT-based) and IFQA (CNN-based) models.

Metric	IFQA++	IFQA
FLOPs (GFLOPs)	337.25	53.23
MACs (GMACs)	168.54	26.53
Params (Million)	656.41	105.17

Table 5
Comparative analysis of IFQA++ (ViT-based) and IFQA (CNN-based) models in inference time, throughput, and VRAM usage.

Metric	IFQA++	IFQA
Inference time (s)	0.0345	0.0112
Throughput (samples/s)	41.20	359.94
VRAM (GBs)	32.8	12.2

Table 4 shows that the IFQA++ has significantly higher FLOPs, MACs, and parameter counts compared to the IFQA. These findings highlight the increased computational complexity and resource demands of the ViT-based model. However, despite the increased complexity and resource usage, the inference time for the ViT-based IFQA++ model remains reasonable (see Table 5).

4.3. Qualitative analysis

Interpretability evaluation: One of the main characteristics of IFQA++ is that it can be visualized by providing interpretable pixel-level scores. To demonstrate the effectiveness and utilization of the proposed interpretable metric, we produced a variety of LQ images (*e.g.*, blur, downsampling, and mix) from the HQ reference images based on the BFR protocol [30]. The restored images were generated 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 comprised of multiple degradation factors. The interpretability maps of IFQA++ are shown in Fig. 7. In our qualitative analysis, we observed that IFQA++ effectively assessed areas of severe degradation. Moreover, our model demonstrated its precision by assigning lower scores to areas with unrealistic textures and facial features. Lastly, we noted that IFQA++ consistently assessed higher scores for regions with realistic facial characteristics.

Comparison with PSNR/SSIM: The proposed face-oriented metric provides pixel-level visualization that is not attainable from general NR-IQA metrics. While FR-IQA metrics such as PSNR and SSIM can provide pixel-level scores, however, they are not easily interpretable. In Fig. 8, we compare the pixel-level scores of PSNR and SSIM maps with the proposed IFQA++. The PSNR map is generated by reversing the $L2$ distance between the reference image and the restored image. For a clearer comparison with IFQA++, the PSNR and SSIM distance maps have been inverted, where brighter areas indicate greater similarity to the reference image. Otherwise, IFQA++ produces an overall low score, while PSNR and SSIM produce sporadic low scores, for the severely degraded 'Image A'. For 'Image B', which is of low quality except for a tiny part of the face, IFQA++ scores high in these undamaged regions.

Ablation study for main modules: We performed an ablation experiment to evaluate different variants, including: (i) a baseline model consisting of a CNN-based per-pixel discriminator, (ii) a baseline model consisting of a learnable generator and an encoder-based discriminator (IFQA++) that provides a single value as output, (iii) a variant of the baseline model where the discriminator generates pixel-level scores, (iv) a variant that integrates the proposed facial primary regions swap (FPRS) into the third baseline model, (v) a variant that combines the original CutMix with the second baseline model, and (vi) a variant that incorporates facial mask information into the fourth baseline model, yielding our final model, IFQA++.

Table 6
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 (CNN) (per-pixel)	0.4674	0.3820
Baseline (single-output)	0.6614	0.5640
Baseline (per-pixel)	0.6814	0.5767
Baseline (per-pixel) + FPRS	0.6823	0.5793
Baseline (per-pixel) + CutMix [48]	0.7177	0.6100
Baseline (per-pixel) + FPRS + F-Mask	**0.7822**	**0.6813**

F-Mask: facial masks using a segmentation model.

Table 7
Performance comparison with respect to generator models on FFHQ, CelebA-HQ, and IWF.

Generator	Task	Parameters	SRCC ↑	KRCC ↑
GPEN [30]	FIR	pre-trained	0.4997	0.4166
DFDNet [36]	FIR		0.5391	0.4366
DBPN [42]	GIR		0.5582	0.4586
RCAN [41]	GIR		0.5711	0.4680
RCAN	GIR	learnable	0.6454	0.5480
Plain model	FIR	learnable	**0.7105**	**0.6059**

FIR: face image restoration GIR: general image restoration

4.4. In-depth analysis

We report the quantitative results on the FFHQ, CelebA-HQ, and IWF datasets in Table 6. We can make the following observations: Per-pixel (ViT-based) baseline consistently outperforms the CNN-based baseline across all metrics, indicating the superior ability of Vision Transformers in capturing detailed facial features. Although CutMix does improve performance, it demonstrates significant inconsistencies with human judgment compared to our IFQA++ model. Finally, the combination of face masks with our FPRS method leads to a substantial improvement, indicating that leveraging face-specific knowledge is essential.

Generator change analysis:

Our framework consists of two sub-networks where one is a generator used as a face restoration model while the other one is a discriminator for assessment. Our hypothesis is that a naive trainable model as a generator would be better suited for learning the discriminator, compared to pre-trained general or face restoration models. To prove this hypothesis, we compare our plain model with the four conventional approaches. (i) GPEN, a state-of-the-art face image restoration model that uses generative prior, (ii) DFDNet, a face image restoration model based on facial components dictionaries, (iii) DBPN, a general image restoration model with backprojection mechanism, and (iv) RCAN, a general image restoration model with residual channel-attention networks.

Table 7 shows that our plain U-Net-based generator shows the highest correlation with human perception, whereas cutting-edge GPEN shows the lowest value.

Discriminator backbone analysis: As proposed IFQA metric relies on the trainable discriminator, a comparison is conducted using various backbone architectures for the discriminator.

U-Net variant using spatial feature transform (SFT) [49], (iii) U-Net variant based on ResNet-50 [50] encoder, (iv) U-Net variant based on VGG-16 [51] encoder, (v) U-Net variant based on VGG-19 [51] encoder, and (vi) U-Net variant based on VIT [18,22] encoder,

Table 8 indicates that all variants of the IFQA++ metric outperform existing IQA metrics, with the best performance achieved by the VIT-based model.

Qualitative results for various degradations: We assess the performance of our proposed IFQA++ metric across different types of degradations. As shown in Fig. 9, we apply motion blur, downsampling,

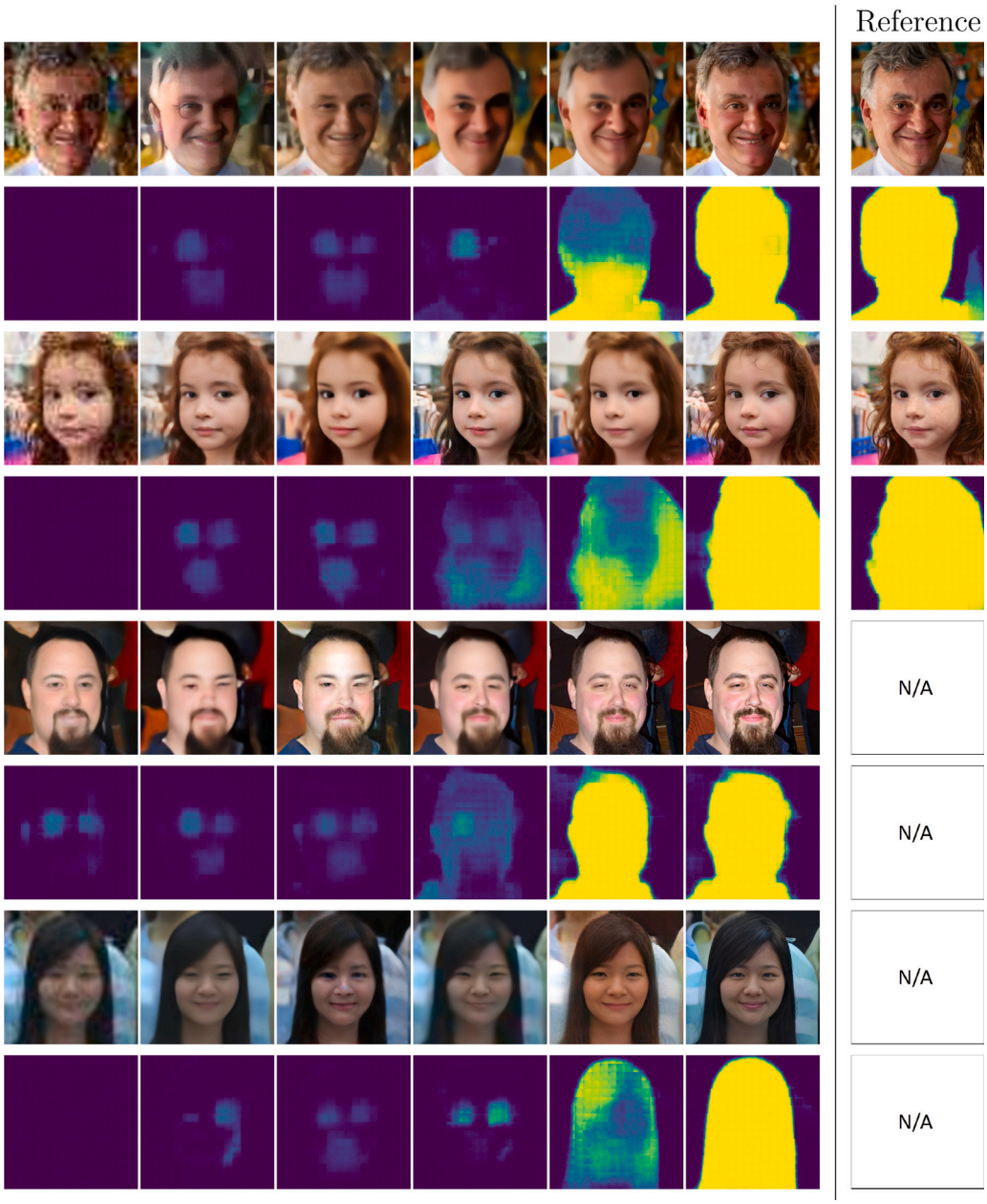


Fig. 7. Interpretable visualization of the proposed metric on various types of LQ images, HQ images (*i.e.*, reference), and RF images from the restoration models. The first to fourth rows show images from FFHQ and their corresponding interpretability maps, respectively. From fifth to eighth rows present pairs from IWF that do not have reference images due to the its protocol. Brighter area indicates the higher quality.

Table 8
Performance comparison with respect to the backbone of discriminators on FFHQ, CelebA-HQ, and IWF.

Discriminator	Parameters	SRCC ↑	KRCC ↑
U-Net [52]	learnable	0.6100	0.5006
U-Net + SFT [49]		0.6314	0.5253
ResNet-50 [50]		0.6311	0.5346
VGG-16 [51]		0.6365	0.5286
VGG-19 [51]		0.6694	0.5600
MAE (VIT) [18,22]		**0.7105**	**0.6059**

and JPEG compression to the face images, and assess the resulting degradation using our metric. As the degradation becomes more severe, our model produces progressively lower assessment scores, demonstrating a direct correlation between the level of degradation and the metric's evaluation. This observation illustrates the assessment performance of the IFQA++ metric in reflecting the quality of images.

4.5. Further use case analysis

Unstable realistic conditions: The proposed metrics have primarily been evaluated on commonly used face image datasets that exhibit relatively stable conditions, typically involving cropped frontal faces. However, real-world face images encompass a wide range of conditions, including variations in lighting and different viewpoints. To assess the generalization ability of our proposed metric in more challenging scenarios, we utilize the VGGFace2 [46] and the CelebA dataset [53].

Although our learning-based metric primarily considers well-aligned scenarios, it still yields reasonable results in more complex settings, as illustrated in Fig. 10. Specifically, IFQA++ generates high-score maps for high-quality 'Image A' and 'Image C,' while producing low-score

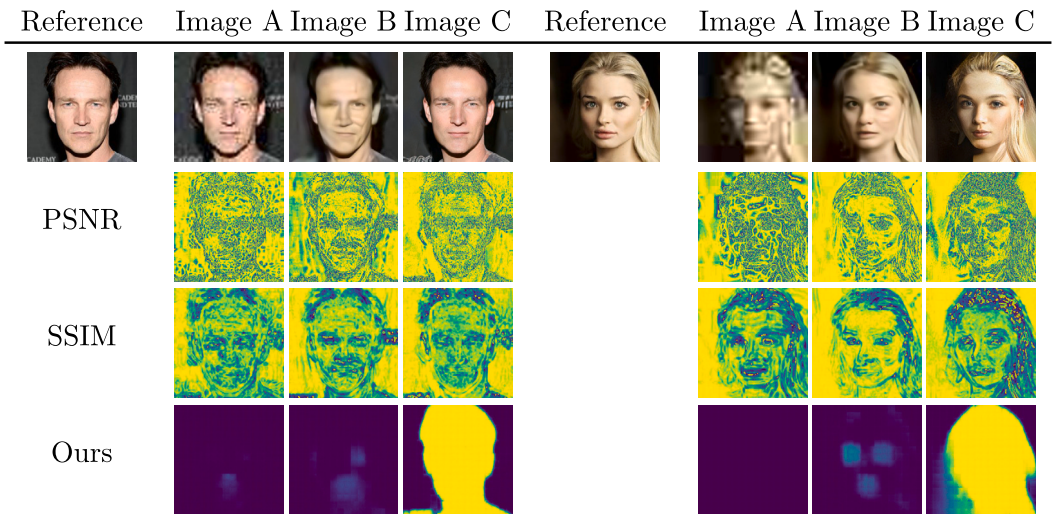


Fig. 8. Comparison of the proposed metric with PSNR/SSIM with respect to pixel-level score. Bright areas indicate higher similarity to the reference image.

Fig. 9. Assessment results on different types and degrees of degradation, including JPEG compression, motion blur ratios, and downsampling. The severity of the degradation gradually increases from left to right. As the degradation becomes more severe, our model produces lower assessment results.

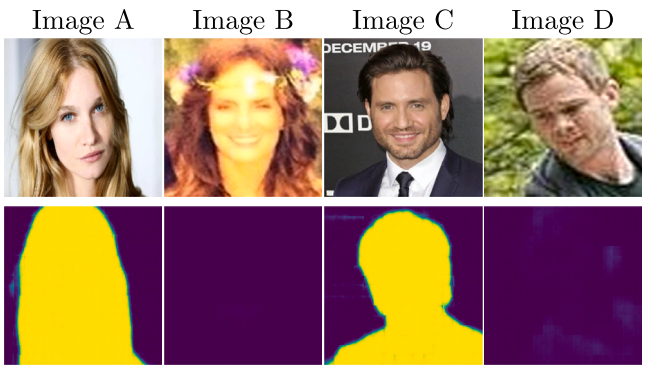


Fig. 10. Results on realistic scenarios. Note that occlusion is not considered as a facial degradation factor for our framework.

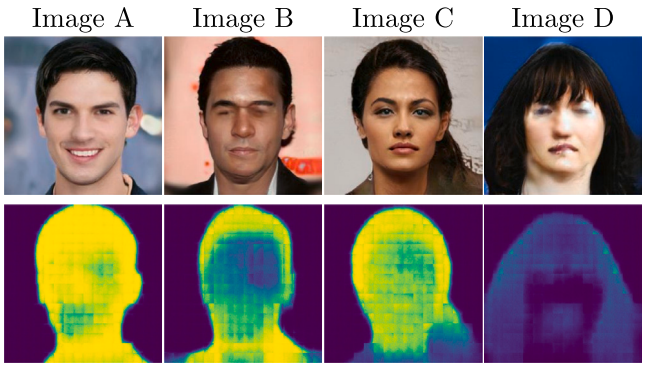


Fig. 11. Results on generated images from StarGANv2 results. Brighter area indicates the higher quality.

maps for the low-quality 'Image B' and 'Image D.' This indicates that our metric successfully discriminates between different image qualities in challenging and less controlled conditions.

Face manipulation tasks: The face domain has received significant attention due to its distinct geometric characteristics, ease of acquisition, and the availability of various established techniques. In recent years, there has been a growing number of efforts to utilize face images as a means to showcase the superiority and effectiveness of methodologies in tasks such as image generation and image-to-image translation. A prominent example of such work is StarGANv2 [54], which employs conventional metrics like FID and LPIPS to evaluate its results. The image-to-image translation results using StarGANv2 and their corresponding visualization maps are depicted in Fig. 11. Overall, the assessment results, particularly for unnatural or degraded regions, suggest that our metric, designed for evaluating face restoration models, has the potential to be utilized for assessing a wide range of face-related image generation tasks.

IFQA as objective function: To investigate the generalization ability of our IFQA metric, we incorporate it as an additional objective function for generative tasks. Due to limitations in GPU memory, we utilize IFQA instead of IFQA++ metrics. Specifically, we apply IFQA to two state-of-the-art face manipulation models, StarGANv2 and StyleMapGAN [55], during the training process. The objective function of IFQA ($\mathcal{L}_{IFQA}$) is defined as follows:

$$\mathcal{L}_{IFQA} = \mathbb{E}_x\,[(D^U(G(x)) - 1)^2], \tag{11}$$

where D^U indicates IFQA model, and G represent the generator in StarGANv2 and StyleMapGAN, respectively.

The proposed IFQA metric evaluates the per-pixel realness of generated images and gives feedback to the generator during the training phase. As a result, the generator is not only trained to generate diverse styles of images but also trained to generate the image's realness. Table 9 clearly shows that our IFQA strategies enhance the performance of StarGANv2 in terms of FID. We also provide a performance comparison in Table 10. StyleMapGAN with our IFQA demonstrates improved performance in MSE and LPIPS metrics, which estimate the quality of reconstructed images at both pixel-level and perceptual-level.

Table 9
Quantitative StarGAN v2 performance comparison with and without our metric as an additional objective function.

Model/Metric	FID-R↓	LPIPS-R↑	FID-L↓	LPIPS-L↑
StarGAN v2	23.7138	**0.3922**	14.6657	**0.4546**
StarGAN v2 w/ IFQA	**22.6457**	0.3914	**13.8008**	0.4516

Table 10
Quantitative StyleMapGAN performance comparison with and without our metric as an additional objective function.

Model/Metric	MSE-M↓	LPIPS-M↓	FID-W↓	FID↓
StyleMapGAN	0.0239	0.2413	**10.2188**	**4.9403**
StyleMapGAN w/ IFQA	**0.0201**	**0.2179**	10.2214	4.9954

5. Conclusion

This study highlights the limitations of general image quality assessment (IQA) by demonstrating their low correlation with human perception. To address this issue, we introduce IFQA++, a novel and specialized face-centric metric specifically designed for evaluating face images. IFQA++ incorporates the Vision Transformer (ViT) architecture, which processes images patch-wise and effectively captures long-range dependencies and global context through self-attention mechanisms. Furthermore, to reduce the dependency on large-scale training data and enhance the generalization ability, we leverage pre-trained weights obtained through masked-autoencoding-based self-supervised learning. By capitalizing on the unique characteristics of facial images, IFQA++ delivers more reliable assessment results compared to existing full-reference IQA (FR-IQA) and no-reference IQA (NR-IQA) metrics. Its specialized design ensures a focus on the specific attributes and challenges associated with face images, enabling more accurate and meaningful evaluations. However, while IFQA++ demonstrates compatible assessment performance compared to widely used existing metrics, there are still limitations to applying it as a universal face image quality assessment metric. Additionally, although our method shows promise in face image quality assessment, it has not yet been fully validated for biometric Face Image Quality Assessment (FIQA) tasks. In the future, we aim to enhance the model's generalization ability by leveraging various face datasets. Expanding the robustness of our metric by incorporating diverse conditions (*e.g.*, different facial demographics, angles of face, and real-world degraded faces) into the training dataset is expected to significantly improve the capability of the proposed method, including its potential application in FIQA.

CRediT authorship contribution statement

Byungho Jo: Writing – original draft, Methodology, Data curation, Conceptualization. **In Kyu Park:** Writing – review & editing, Supervision, Formal analysis. **Sungeun Hong:** Writing – review & editing, Supervision, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sungeun Hong reports administrative support was provided by Sungkyunkwan University. Sungeun Hong reports a relationship with Sungkyunkwan University that includes: employment. If there are other

authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A4A1033549, No. RS-2023-00211348).

Data availability

Data will be made available on request.

References

[1] W. Im, S. Hong, S. Yoon, H.S. Yang, Scale-varying triplet ranking with classification loss for facial age estimation, in: ACCV, Springer, 2018, pp. 247–259.
[2] S. Hong, J. Ryu, Unsupervised face domain transfer for low-resolution face recognition, IEEE Sign. Process. Lett. 27 (2019) 156–160.
[3] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, W. Liu, Face anti-spoofing: Model matters, so does data, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 3507–3516.
[4] Q. Huynh-Thu, M. Ghanbari, Scope of validity of PSNR in image/video quality assessment, Electron. Lett. 44 (13) (2008) 800–801.
[5] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[6] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2018, pp. 586–595.
[7] A. Mittal, R. Soundararajan, A.C. Bovik, Making a "completely blind" image quality analyzer, IEEE Sign. Process. Lett. 20 (3) (2013) 209–212.
[8] A. Mittal, A.K. Moorthy, A.C. Bovik, Blind/referenceless image spatial quality evaluator, in: ACSCC, IEEE, 2011, pp. 723–727.
[9] F. Makhmudkhujaev, S. Hong, I.K. Park, Re-aging gan: Toward personalized face age transformation, in: Int. Conf. Comput. Vis., 2021, pp. 3908–3917.
[10] S. Hong, J. Ryu, Attention-guided adaptation factors for unsupervised facial domain adaptation, Electron. Lett. 56 (16) (2020) 816–818.
[11] F. Wei, S. Wang, J. Yang, X. Sun, Y. Wang, Y. Chen, A composite network model for face super-resolution with multi-order head attention facial priors, Pattern Recognit. 139 (2023) 109503.
[12] M.J. Farah, K.L. Levinson, K.L. Klein, Face perception and within-category discrimination in prosopagnosia, Neuropsychologia 33 (6) (1995) 661–674.
[13] J. Liu, A. Harris, N. Kanwisher, Perception of face parts and face configurations: an fMRI study, J. Cogn. Neurosci. 22 (1) (2010) 203–211.
[14] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, L. Beslay, Biometric quality: Review and application to face recognition with FaceQnet, 2020, arXiv preprint arXiv: 2006.03298.
[15] F. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, Y. Wang, SDD-FIQA: unsupervised face image quality assessment with similarity distribution distance, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2021, pp. 7670–7679.
[16] F. Boutros, M. Fang, M. Klemt, B. Fu, N. Damer, CR-FIQA: Face image quality assessment by learning sample relative classifiability, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 5836–5845.
[17] B. Jo, D. Cho, I.K. Park, S. Hong, IFQA: Interpretable face quality assessment, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2023, pp. 3444–3453.
[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[19] S. Choi, Y. Zhang, S. Hong, Intra-inter modal attention blocks for RGB-d semantic segmentation, in: Int. Conf. Multimedia Retrieval, 2023, pp. 217–225.
[20] Y. Zhang, S. Choi, S. Hong, Spatio-channel attention blocks for cross-modal crowd counting, in: Asian Conf. on Computer Vis., 2022, pp. 90–107.
[21] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 16000–16009.
[23] Y. Zhang, M. Ding, Y. Bai, B. Ghanem, Detecting small faces in the wild based on generative adversarial network and contextual information, Pattern Recognit. 94 (2019) 74–86.
[24] R. Xu, K. Wang, C. Deng, M. Wang, X. Chen, W. Huang, J. Feng, W. Deng, Depth map denoising network and lightweight fusion network for enhanced 3D face recognition, Pattern Recognit. 145 (2024) 109936.
[25] L. Yang, S. Wang, S. Ma, W. Gao, C. Liu, P. Wang, P. Ren, Hifacegan: Face renovation via collaborative suppression and replenishment, in: ACM Int. Conf. Multimedia, ACM, 2020, pp. 1551–1560.
[26] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2019, pp. 4401–4410.
[27] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: International Conference on Learning Representations, 2021.
[28] Z. Yue, J. Wang, C.C. Loy, ResShift: Efficient diffusion model for image super-resolution by residual shifting, in: Thirty-Seventh Conference on Neural Information Processing Systems, 2023.
[29] Z. Yue, C.C. Loy, DifFace: Blind face restoration with diffused error contraction, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–15.
[30] T. Yang, P. Ren, X. Xie, L. Zhang, GAN prior embedded network for blind face restoration in the wild, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2021, pp. 672–681.
[31] J. Ke, Q. Wang, Y. Wang, P. Milanfar, F. Yang, MUSIQ: Multi-scale image quality transformer, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 5128–5137.
[32] Y. Blau, T. Michaeli, The perception-distortion tradeoff, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2018, pp. 6228–6237.
[33] B. Fu, C. Chen, O. Henniger, N. Damer, A deep insight into measuring face image utility with general and face-specific image quality metrics, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, IEEE Computer Society, 2022, pp. 1121–1130.
[34] P. Terhörst, J.N. Kolf, N. Damer, F. Kirchbuchner, A. Kuijper, SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2020, pp. 5650–5659.
[35] W.-T. Chen, G. Krishnan, Q. Gao, S.-Y. Kuo, S. Ma, J. Wang, DSL-FIQA: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 2931–2941.
[36] X. Li, C. Chen, S. Zhou, X. Lin, W. Zuo, L. Zhang, Blind face restoration via deep multi-scale component dictionaries, in: Eur. Conf. Comput. Vis., Springer, 2020, pp. 399–415.
[37] C. Lee, Z. Liu, L. Wu, P. Luo, MaskGAN: Towards diverse and interactive facial image manipulation, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2020, pp. 5548–5557.
[38] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230, 000 3D facial landmarks), in: Int. Conf. Comput. Vis., Computer Vision Foundation / IEEE, 2017, pp. 1021–1030.
[39] X. Mao, Q. Li, H. Xie, R.Y.K. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, in: Int. Conf. Comput. Vis., Computer Vision Foundation / IEEE, 2017, pp. 2813–2821.
[40] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: Int. Conf. Learn. Represent., OpenReview.net, 2018.
[41] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Eur. Conf. Comput. Vis., Springer, 2018, pp. 294–310.
[42] M. Haris, G. Shakhnarovich, N. Ukita, Deep back-projection networks for super-resolution, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2018, pp. 1664–1673.
[43] Z. Wang, E. Simoncelli, A. Bovik, Multiscale structural similarity for image quality assessment, in: ACSCC, 2003, pp. 1398–1402.
[44] H. Talebi, P. Milanfar, NIMA: Neural image assessment, IEEE Trans. Image Process. 27 (8) (2018) 3998–4011.
[45] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, R. Haraksim, L. Beslay, FaceQnet: Quality assessment for face recognition based on deep learning, in: ICB, IEEE, 2019, pp. 1–8.
[46] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, VGGFace2: A dataset for recognising faces across pose and age, in: IEEE International Conference on Automatic Face & Gesture Recognition, FG, IEEE Computer Society, 2018, pp. 67–74.
[47] M.A. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (8) (2012) 3339–3352.
[48] S. Yun, D. Han, S. Chun, S.J. Oh, Y. Yoo, J. Choe, CutMix: Regularization strategy to train strong classifiers with localizable features, in: Int. Conf. Comput. Vis., Computer Vision Foundation / IEEE, 2019, pp. 6022–6031.

[49] X. Wang, K. Yu, C. Dong, C.C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2018, pp. 606–615.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2016, pp. 770–778.
[51] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Int. Conf. Learn. Represent., MIT Press, 2015.
[52] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, Vol. 9351, Springer, 2015, pp. 234–241.
[53] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision, ICCV, 2015.
[54] Y. Choi, Y. Uh, J. Yoo, J. Ha, Stargan v2: Diverse image synthesis for multiple domains, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2020, pp. 8185–8194.
[55] H. Kim, Y. Choi, J. Kim, S. Yoo, Y. Uh, Exploiting spatial dimensions of latent in GAN for real-time image editing, in: IEEE Conf. Comput. Vis. Pattern Recog., Computer Vision Foundation / IEEE, 2021, pp. 852–861.

Byungho Jo received the B.S. degree in computer science and engineering from Incheon National University, South Korea, in 2020, and the M.S. degree in electrical engineering and computer science from Inha University, South Korea, in 2022. He is currently a researcher with AI Convergence Research Center at Inha University. From April 2022 to April 2023, he was a visiting researcher in the Department of Artificial Intelligence and Informatics at Mayo Clinic Florida. His current research interests include multimodal AI, neurosymbolic AI, and computer vision.

In Kyu Park received the B.S., M.S., and Ph.D. degrees from Seoul National University in 1995, 1997, and 2001, respectively, all in electrical engineering and computer science. From September 2001 to March 2004, he was a Member of Technical Staff at Samsung Advanced Institute of Technology. Since March 2004, he has been with the Department of Information and Communication Engineering, Inha University, where he is a full professor. From January 2007 to February 2008, he was an exchange scholar at Mitsubishi Electric Research Laboratories. From September 2014 to August 2015, he was a visiting associate professor at MIT Media Lab. From July 2018 to June 2019, he was a visiting scholar at the Center for Visual Computing in University of California, San Diego. Currently he is serving as an Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence. Dr. Park's research interests include the joint area of computer vision and graphics, including 3D shape reconstruction from multiple views, 3D human modeling, computational photography, and deep learning. He is a senior member of IEEE and a member of ACM.

Sungeun Hong received the B.S. degree in computer engineering from Hanyang University, South Korea, in 2010, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 2012 and 2018, respectively. He is currently an Associate Professor with Sungkyunkwan University, South Korea. Prior to his position at Sungkyunkwan University, he was an Assistant Professor with Inha University and a Research Scientist with T-Brain, AI Center, SK Telecom, South Korea. His current research interests include multimodal learning, vision–language models, face understanding, and 3D perception.