Generalized Visual Quality Assessment of GAN-Generated Face Images

Yu Tian, Zhangkai Ni, *Member, IEEE*, Baoliang Chen, Shiqi Wang, *Senior Member, IEEE*, Hanli Wang, *Senior Member, IEEE,* and Sam Kwong, *Fellow, IEEE*

***Abstract*—Recent years have witnessed the dramatically increased interest in face generation with generative adversarial networks (GANs). A number of successful GAN algorithms have been developed to produce vivid face images towards different application scenarios. However, little work has been dedicated to automatic quality assessment of such GAN-generated face images (GFIs), even less have been devoted to generalized and robust quality assessment of GFIs generated with unseen GAN model. Herein, we make the first attempt to study the subjective and objective quality towards generalized quality assessment of GFIs. More specifically, we establish a large-scale database consisting of GFIs from four GAN algorithms, the pseudo labels from image quality assessment (IQA) measures, as well as the human opinion scores via subjective testing. Subsequently, we develop a quality assessment model that is able to deliver accurate quality predictions for GFIs from both available and unseen GAN algorithms based on meta-learning. In particular, to learn shared knowledge from GFIs pairs that are born of limited GAN algorithms, we develop the convolutional block attention (CBA) and facial attributes-based analysis (ABA) modules, ensuring that the learned knowledge tends to be consistent with human visual perception. Extensive experiments exhibit that the proposed model achieves better performance compared with the state-of-the-art IQA models, and is capable of retaining the effectiveness when evaluating GFIs from the unseen GAN algorithms.**

***Index Terms*—Generative adversarial network, domain generalization, meta-learning, face image quality assessment.**

I. Introduction

GENERATIVE adversarial network (GAN) [1], [2] has achieved remarkable success in various areas such as image generation [3], [4], [5], image restoration [6], [7], image quality enhancement [8], as well as the face images generation [9], [10], [11]. In recent years, facial image synthesis and editing have been applied in different areas, such as film, medical aesthetics institutions, and photography technologies. For those specific applications in which the ultimate receiver is human visual system (HVS), face images synthesized by the generative models are expected to be consistent with the perception of HVS. Therefore, the image quality assessment (IQA) models that can predict the perceptual quality of GAN-generated face images (GFIs) are highly desirable.

In the literature, numerous IQA models have been proposed. According to the availability of the reference image, existing IQA models can be divided into three categories: full-reference IQA (FR-IQA) [12], [13], [14], reduced-reference IQA (RR-IQA) [15], [16], and no-reference IQA (NR-IQA) [17], [18], [19], [20]. However, most existing methods are dedicated to natural images with traditional distortion types such as compression, noise and blur. As shown in Fig. 1, GFIs may exhibit substantially different types of distortions, and those distortions on critical face attributes even make GFI look unreal. Therefore, the specific IQA model which predicts the quality of GFIs is in high demand.

Evaluating the performance of GANs and the quality of GAN-generated images have attracted increasing attention during the past years. The most commonly adopted metrics are Inception score (IS) [21] and Frechet Inception distance (FID) [22]. IS [21] exploits the pre-trained Inception [23] classifier to extract the class probability of each generated image. Then the Kullback–Leibler (KL) divergence between the probability of this image and the marginal distribution is regarded as the performance of the GAN model. FID [22] compares activations between real data and generated data in an intermediate pooling layer of Inception [23]. It is worth mentioning that IS and FID can only measure the overall performance of GAN model instead of each individual image. Furthermore, Gu *et al.* [24] proposed three GAN-generated images quality assessment (GIQA) methods from different perspectives, and the GIQA model with the best performance is based on Gaussian mixture model (GMM), termed as GMM-GIQA. Unfortunately, GMM-GIQA mainly focuses on the probability of the generated data in the distribution of real data (i.e., training data) while ignoring the perceptual characteristics relevant to HVS.

In this paper, we make one of the first attempts to conduct subjective and objective studies on the quality assessment of GFIs. Regarding the dataset, we collect a large number of GFIs from four different GAN algorithms to establish metric-labeled training data and human-labeled testing data. Eventually, our database contains 200,000 training image pairs and 2,000 testing images. Regarding the objective model, given the fact that there exists large domain gaps among GFIs from different GAN algorithms, which can be observed in the form of the average spectra as shown in Fig. 2, we aim to develop an objective model that is robust to different GAN algorithms

Yu Tian, Baoliang Chen and Shiqi Wang are with the Department of Computer Science, City University of Hong Kong, Hong Kong 999077 (e-mail:ytian73-c@my.cityu.edu.hk; blchen6-c@my.cityu.edu.hk, shiqwang@cityu.edu.hk).

Zhangkai Ni is with the Department of Computer Science & Technology, Tongji University, Shanghai 200092, P. R. China (e-mail: eezkni@gmail.com).

Hanli Wang is with the Department of Computer Science & Technology, Key Laboratory of Embedded System and Service Computing (Ministry of Education), and Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, P. R. China (e-mail: hanliwang@tongji.edu.cn).

Sam Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong 999077, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China (e-mail: cssamk@cityu.edu.hk).

Fig. 1. Illustration of the samples in the proposed GFID.

even in the unseen domain. To this end, we develop the GFIs quality assessment measure by associating the meta-learning optimization strategy with quality-relevant features, which enhances the generalization capability and improves the prediction performance. In summary, the main contributions of this paper are summarized as follows,

- We establish the large-scale GFIs database (GFID) for GFIs quality assessment, which contains the training image pairs associated with relative quality information and the testing images with human-annotated mean opinion scores (MOSs).
- We develop a new GFIs quality assessment measure that is robust to different GAN algorithms with high generalization capability. More specifically, we devise the convolutional block attention (CBA) and facial attributes-based analysis (ABA) modules to extract the quality-relevant features over GFIs of each source GAN, and adopt the meta-learning strategy to mine the shared features representation across GFIs from different GAN algorithms.
- We perform extensive experiments, which provide useful evidence that our model is more consistent with the perception of the HVS when assessing GFIs quality compared to state-of-the-art IQA methods. In particular, our model is able to achieve good generalization performance on unknown GANs.

II. Related works

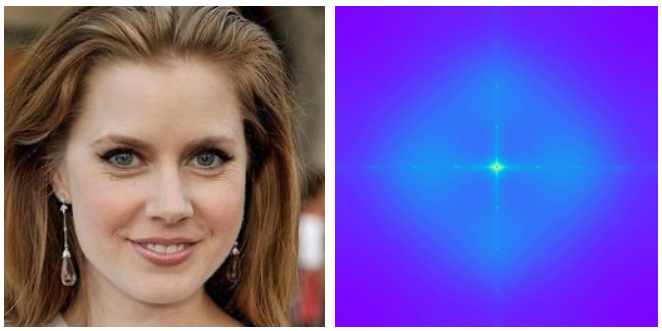
A. Image Quality Assessment

IQA is a long-standing research topic regarding how to provide an objective measurement of the image quality that is consistent with the HVS. Most FR-IQA models have achieved promising performance by directly comparing the distorted image and its corresponding pristine version, such as the structural similarity (SSIM) index [12], the visual saliency-induced (VSI) index [25], the visual information fidelity (VIF) metric [26]. NR-IQA models, however, have attracted increasing attentions in recent years due to the limited available pristine images. Early NR-IQA models are developed based on knowledge-driven strategy [27], [28], aiming at designing a quality-aware feature descriptor. Subsequently, a regression model, e.g., support vector regression (SVR), maps the feature representation into a quality score. Based on the assumption that the natural scene statistics (NSS) characterized from natural images can govern the quality of natural images and the NSS will be destroyed in the presence of distortion, Moorthy *et al.* [29] predicted the image quality by characterizing the NSS of distorted images in the wavelet domain. Mittal *et al.* [30] exploited the normalized luminance coefficients to map image features in the spatial domain into the quality score. Instead of only measuring the NSS regularities, Mittal *et al.* [31] further proposed a naturalness image quality evaluator (NIQE) based on quality-aware statistical features.

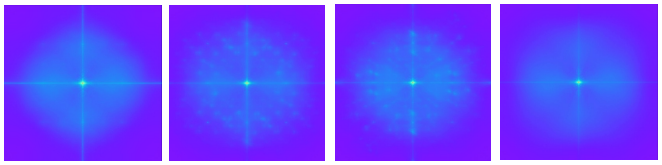
Recently, deep-learning-based NR-IQA methods have been proposed due to the powerful feature representation capability. Ma *et al.* [32] proposed an end-to-end quality assessment model based on multi-task learning, which consists of a quality assessment sub-network and a distortion identification sub-network. Zhang *et al.* [33] proposed a deep bilinear model for synthetically and authentically distorted images. The quality score can be obtained by bilinearly pooling two feature representations extracted from two CNN models. Su *et al.* [34] combined quality prediction with content understanding and proposed a self-adaptive IQA model. To improve the generalization of NR-IQA models, Zhu *et al.* [17] used the deep meta-learning approach to develop a novel NR-IQA model called Meta-IQA. However, the limitation for deep-learning-based NR-IQA methods is that the lack of human-labeled training data makes the model training difficult. To address this problem, Liu *et al.* [18] ranked images according to distortion levels of distorted images and then used the ranked images to train the siamese network. Finally, they fine-tuned the model to transform the task from learning-to-rank to the quality score regression. Ma *et al.* [35] applied RankNet [36] to learn the quality relationship between the quality-discriminable image pairs. Chen *et al.* [20] explored the unified distribution regularization on the feature space and developed an unsupervised domain adaptation based quality assessment model for screen content images.

B. Domain Generalization

Domain generalization (DG) aims to learn a model that can generalize to unseen data distributions by training on one or more data distributions. Previous studies on DG are mainly based on four different strategies. The first is to use the discrepancy measurement, e.g., maximum mean discrepancy (MMD) [38], [39], to learn domain-invariant features. Such domain-invariant features can work well in both source and unseen target domains. The second resorts to the feature disentanglement, which is based on the assumption that each domain can be presented as the combination of the domain-specific component and the domain-agnostic component. Thus,



(a) CelebA-HQ



(b) PENNet (c) ProGAN (d) StarGAN (e) StyleGAN

Fig. 2. The average spectra for the real face images in CelebA-HQ database [3] and face images generated by PENNet [37], ProGAN [3], StarGAN [6] and StyleGAN [4], respectively. Each average spectra is generated by averaging the frequency spectra of over 5,000 face images.

TABLE I
THE DETAILS OF THE GFID TESTING DATASET.

Model	Year	# of Images	Task
StarGAN [6]	2018	469	Image-to-image translation
ProGAN [3]	2018	515	Image generation
PENNet [37]	2019	527	Image inpainting
StyleGAN [4]	2019	489	Image generation

the model learned from the domain-agnostic component [40], [41] can achieve better generalization performance. The third is based on data augmentation, which synthesizes additional training data to enhance the robustness of the models to target domains [42], [43]. The last is to find minima during training, called optimization strategy. The most popular optimization strategy is meta-learning [44], [45], [17], which improves a conventional learning algorithm to improve the robustness of the models on unseen domains. The fundamental idea behind meta-learning is that simulating the real train/test data shift during training improves the generalization ability. In this paper, our approach applies meta-learning to build domain-shift batches at each episode to optimize it for learning more robust face representations.

III. The Proposed Database

The proposed GFID database is established for the purpose of developing a reliable quality assessment model. As such, the database includes a training dataset and a testing dataset.

A. Training Dataset

We first obtain GFIs from different GAN algorithms. To cover various types of distortions in different application scenarios, we select four GAN models, including the pyramid-context encoder network (PENNet) [37] for image inpainting/restoration, the GAN using a progressive training methodology (ProGAN) [3], the style-based GAN (StyleGAN) [4] for unconditional image generation, and the StarGAN [6] for multi-domain image-to-image translation. Each GAN model generates 15,000 face images forming four subdatasets. It is worth mentioning that we use the pre-trained versions of the above-mentioned GAN models which are trained on the large-scale CelebFaces attributes dataset (i.e., CelebA-HQ [3]) including over 30,000 high-quality celebrity images. Furthermore, to acquire the quality levels of the GFIs, we adopt the pseudo-labels based approach, as human annotation of such a large quantity of images is tedious and expensive. Considering that information content weighted SSIM (IW-SSIM) [46] has been widely used to evaluate the quality of GFIs in image inpainting and restoration tasks, and it can reflect image degradations at low-level feature by comparing the restored image and the corresponding pristine version, we select the IW-SSIM values as the pseudo-MOS values of GFIs generated by PENNet. For GFIs from ProGAN, StyleGAN, and StarGAN, the corresponding pristine version of each GFI is unavailable. Herein, the GMM-GIQA model [24] is applied to produce the pseudo-MOS values. More specifically, we first use the Inception-v3 network [23] to extract the feature representations of real face images in CelebA-HQ as suggested in [24]. Subsequently, we apply the method provided by the authors [24] to build the GMM model to capture the data distribution from the feature representations. Finally, the probabilities of GFIs computed from the GMM can be used as the pseudo-MOS values of GFIs generated by ProGAN, StyleGAN, and StarGAN.

Inspired by the promising transferability of pair-wise relationship in IQA, we establish a pair-wise training dataset with the relative quality ranking automatically generated by pseudo-MOS values of two images within each image pair. To further mitigate the over-fitting problem caused by distinguishing image pairs with ambiguous quality, we collect quality-discriminable image pairs from each subdataset. More specifically, we first sort all GFIs within each subdataset according to their corresponding pseudo-MOS values from high to low. Then, we divide the sorted GFIs into three quality levels, of which each quality level contains 5,000 GFIs. Subsequently, we sample the top 50 and the worst 50 GFIs from the first and the third quality levels, respectively. Within the 100 images, each image is paired with 500 GFIs randomly sampled from the second quality level. Finally, our training dataset contains 200,000 image pairs with relative quality rankings in total, of which each subdataset consists of 50,000 quality-discriminable image pairs.

B. Testing Dataset

Testing data contains a total of 2,000 GFIs. More specifically, we initially collect around 3,000 face images generated by PENNet [37], ProGAN [3], StyleGAN [4], and StarGAN [6]. Subsequently, we apply the semantic segmentation network [47] to analyze the generated images and refine the

selection according to the richness of semantic constituents. This step ensures the diversity of generated image content and quality. Finally, 2,000 GFIs are selected, and we resize them to a fixed resolution of 256×256. The details of the testing data can be found in Table I and the sampled images are shown in Fig. 1. The subjective experiment is conducted to collect human opinion scores of the 2,000 testing GFIs. Such scores are treated as the ground-truth labels to examine whether IQA models are well correlated with human perception.

1) Subjective Test: In order to reduce the effect of the viewer fatigue, we randomly and non-overlappingly divide the entire 2,000 GFIs into eight sessions. As specified in ITU-R BT.500-13 [48], each session is evaluated by at least 15 subjects. Before starting the testing session, we provide a sample set of GFIs to ensure each subject has fully understood how to rate the quality scores of GFIs. For the large-scale image database, based on the recommendations of ITU [48], the single stimulus approach with 5-category discrete scales is used for the subjective experiment. From the highest perceptual quality to the lowest perceptual quality, the impairment scales are classified as "Excellent", "Good", "Fair", "Poor", "Bad". During the subjective experiment, subjects are asked to choose the subjective value in the range between 1 to 5 according to the impairment scale of the image. The higher value indicates the better perceptual quality. The graphical user interface is shown in Fig. 4 and we conduct the experiment in a well-controlled laboratory environment.

2) Subjective Score Processing: After obtaining the subjective scores directly rated by subjects, we further process the subjective results to generate the MOS of each GFI. As suggested in [48], [49], we denote S_{ijk} as the opinion score of i-th image evaluated by subject j in k-th testing session, where $i = 1, \ldots, 20, j = 1, \ldots, 250$ and $k = 1, \ldots, 8$. We convert the S_{ijk} to Z-score per session as follows:

$$\begin{aligned} \mu_{jk} &= \frac{1}{N_k}\sum_{i=1}^{N_k}, \\ \sigma_{jk} &= \sqrt{\frac{1}{N_k - 1}\sum_{i=1}^{N_k}(S_{ijk} - \mu_{jk})^2}, \\ Z_{ijk} &= \frac{S_{ijk} - \mu_{jk}}{\sigma_{jk}}, \end{aligned} \tag{1}$$

where N_k is the number of subjects in the test session k.

Assuming that Z-scores assigned by a subject follow a standard Gaussian, 99% values will fall in the range between -3 and +3. Therefore, Z-scores can be re-scaled to the range of [0,100] by the following linear mapping,

$$\tilde{Z}_{ijk} = \frac{100(Z_{ijk} + 3)}{6}. \tag{2}$$

Finally, the MOS of the GFI j in k-th testing session, which is denoted as MOS_{jk}, can be obtained as:

$$MOS_{jk} = \frac{1}{M_k}\sum_{i=1}^{M_k}\tilde{Z}_{ijk}, \tag{3}$$

where M_k is the number of remaining subjects in the session k after the subject rejection.

To analyze the reliability of the proposed testing dataset, we visualize the MOS distribution of all tested GFIs via the histogram and the scatter plots. As shown in Fig. 3, one can easily observe that the MOS values of all GFIs are distributed between 10 and 80, which implies the perceptual quality of GFIs in our dataset has spanned a wide range of visual quality from severely annoying to imperceptible with a good separation.

IV. Proposed Method

A. Overview

Let $\mathcal{I}$ denote an input space consisting of GFIs and $\hat{\mathcal{Y}}$ is the output space representing the pseudo-MOS values of samples in $\mathcal{I}$. Given K training (source) domains generated by K types of GAN methods, $\mathcal{D} = \{\mathcal{D}_k \mid k = 1, \cdots, K\}$, where $\mathcal{D}_k = \left\{\left(I_i^{(k)}, \hat{y}_i^{(k)}\right), \left(I_j^{(k)}, \hat{y}_j^{(k)}\right)\right\}_{i \neq j}$ denotes the k-th training domain containing GFIs pairs and the pseudo-MOS values, we can infer the relative quality rankings by comparing the pseudo-MOS values of two images within image pairs. More specifically, we set the relative quality ranking to 1 if the pseudo-MOS value $\hat{y}_i^{(k)}$ of the image $I_i^{(k)}$ is higher than the pseudo-MOS value $\hat{y}_j^{(k)}$ of the image $I_j^{(k)}$ and 0 otherwise. Our goal is to learn a generalizable predictive function h via the pairwise learning-to-rank approach within K training domains to achieve a minimum prediction error on an unseen testing domain $\mathcal{D}_{\text{unseen}}$:

$$\min_h \mathbb{E}_{(I,MOS)\in\mathcal{D}_{\text{unseen}}}[\ell(h(I), MOS)], \tag{4}$$

where MOS is the human-annotated score of the image I, $\mathbb{E}$ is the expectation, and $\ell(\cdot,\cdot)$ is the loss function. In particular, the $\mathcal{D}_{\text{unseen}}$ is generated by the GAN models which is exactly unseen during the training phase, thus the joint distributions between the training and testing sets are dramatically different. The prediction model is required to perform well on source domains as well as the unseen domains. Herein, we aim to measure the generalization capability of our IQA models with the blind assumption on the specific generative model. Such scenario plays a more practical role in real applications.

To this end, we first propose the CBA and ABA modules to capture the intrinsic and perceptual aware features delicately. In particular, the CBA module aims to mine the visual attention region of the face image and the ABA module attempts to extract perceptual feature representations from different facial attributes in visual attention regions. Subsequently, the meta-learning approach is adopt to learn the prior knowledge of the HVS perception shared by different domains. The overview of the meta-learning strategy and the architecture of the proposed model are shown in Fig. 5 and Fig. 6, respectively, and we will elaborate the design details of each module in the following subsections.

B. Convolutional Block Attention Module

Inspired by human visual attention mechanism, the attention module is designed for high task-relevant feature learning [50], [51]. Considering that channel and spatial attentions have different meanings for the whole image, i.e., "what" is

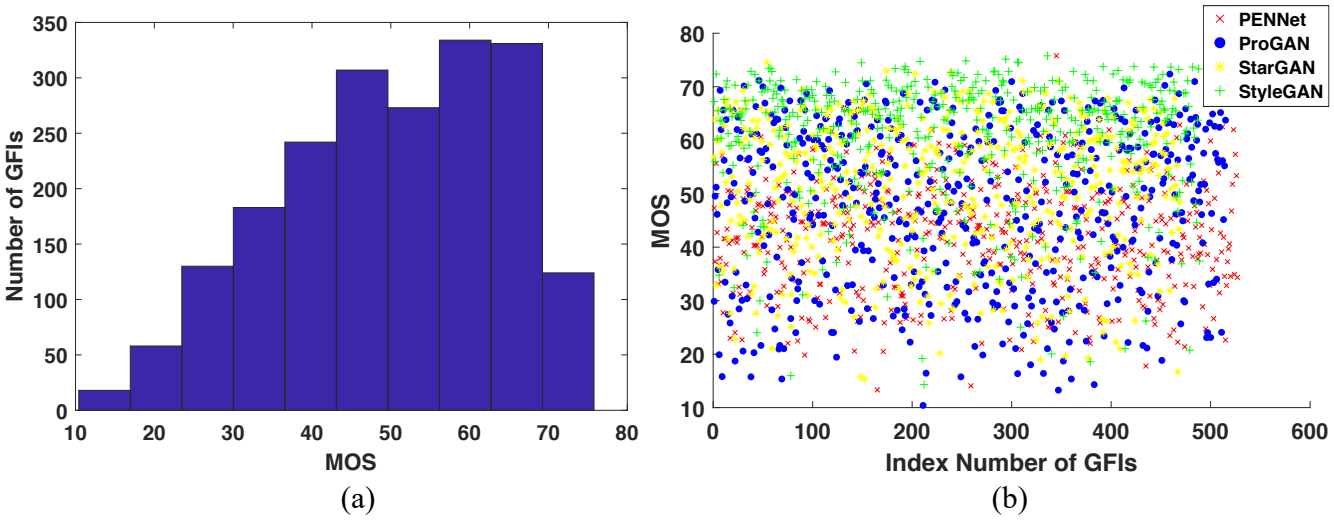


Fig. 3. The subjective scores (i.e., MOS) distribution in the testing dataset of our proposed GFID. (a) Histogram of MOS; (b) Scatter plot for different GAN algorithms.

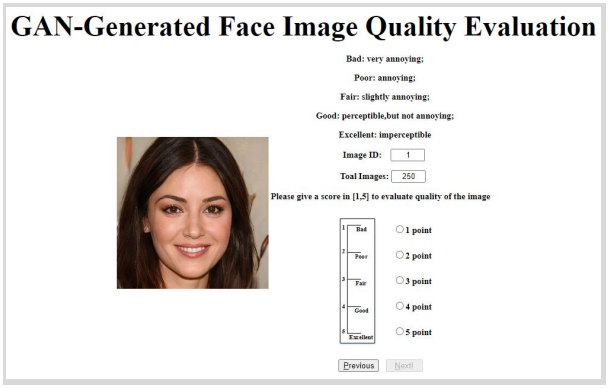


Fig. 4. A screenshot of the user interface in the subjective test.

important and "where" is informative, respectively, we use the CBA module to capture the salient feature by aggregating the channel and spatial attention features. The structure of CBA is shown in Fig. 7(a). To be specific, the input of the t-th CBA module in Fig. 6 is denoted as $F_t \in \mathbb{R}^{C\times H\times W}$. We first take global average pooling operation (denoted as GAP_c) to aggregate spatial features of each channel to obtain the global feature descriptor (denoted as $F_{c,t}^{avg} \in \mathbb{R}^{C\times 1\times 1}$). Then we feed it into two 1×1 convolutions with stride 1 (denoted as $\text{Conv}_{1\times 1}$) and apply a sigmoid activation layer (denoted as σ) to produce the channel weighting map $W_{c,t}$,

$$\begin{aligned} W_{c,t} &= \sigma(\text{Conv}_{1\times 1}(\text{Conv}_{1\times 1}(\text{GAP}_c(F_t)))) \\ &= \sigma(\text{Conv}_{1\times 1}(\text{Conv}_{1\times 1}(F_{c,t}^{avg}))). \end{aligned} \tag{5}$$

It is worth mentioning that in order to reduce the model complexity, the dimension of the feature $F_{c,t}^{avg} \in \mathbb{R}^{C\times 1\times 1}$ should be reduced as $\mathbb{R}^{C/r\times 1\times 1}$ using 1-th $\text{Conv}_{1\times 1}$, where r is the reduction radio. After obtaining the channel weighting map, we multiply it with the input features F_t to generate the channel attention feature $V_{c,t}$,

$$V_{c,t} = W_{c,t} \otimes F_t, \tag{6}$$

where the symbol "$\otimes$" denotes elemental-wise multiplication.

Regarding the spatial attention, the max pooling and average pooling layers (denoted as GMP_s and GAP_s) are further employed along the channel axis of $V_{c,t}$, respectively. Subsequently, we concatenate the outputs to generate the spatial feature descriptor denoted as $F_{s,t} \in \mathbb{R}^{2\times H\times W}$. Finally, the spatial attention map $W_{s,t}$ can be acquired with a convolutional layer (denoted as $\text{Conv}_{5\times 5}$) and a sigmoid activation layer followed, which is given by,

$$\begin{aligned} W_{s,t} &= \sigma(\text{Conv}_{5\times 5}([\text{GMP}_s(V_{c,t}), \text{GAP}_s(V_{c,t})])) \\ &= \sigma(\text{Conv}_{5\times 5}([F_{s,t}^{max}, F_{s,t}^{avg}])) \\ &= \sigma(\text{Conv}_{5\times 5}(F_{s,t})), \end{aligned} \tag{7}$$

where $[\cdot,\cdot]$ represents the channel-wise concatenation, $F_{s,t}^{max}$ and $F_{s,t}^{avg}$ are the outputs of GMP_s and GAP_s.

The final visual attention features $V_t \in \mathbb{R}^{C\times H\times W}$ is given by,

$$V_t = W_{s,t} \otimes V_{c,t} + V_{c,t}. \tag{8}$$

As shown in Fig. 6, the CBA module is incorporated before every two ABA modules, thus the salient information can be exploited at different scales.

C. Attributes-Based Analysis Module

A face image can be described by several meaningful attributes, such as hair, eye, nose, mouth, skin, which play important roles for the quality evaluation of GFIs [52]. Along this vein, we transform the visual attention features extracted by CBA into attributes-based features and predict the quality score by aggregating different attributes. The structure of our ABA module is shown in Fig. 7(b). More specifically, the ABA module requires two inputs, including the output of the previous module and a face segmentation map. For example,

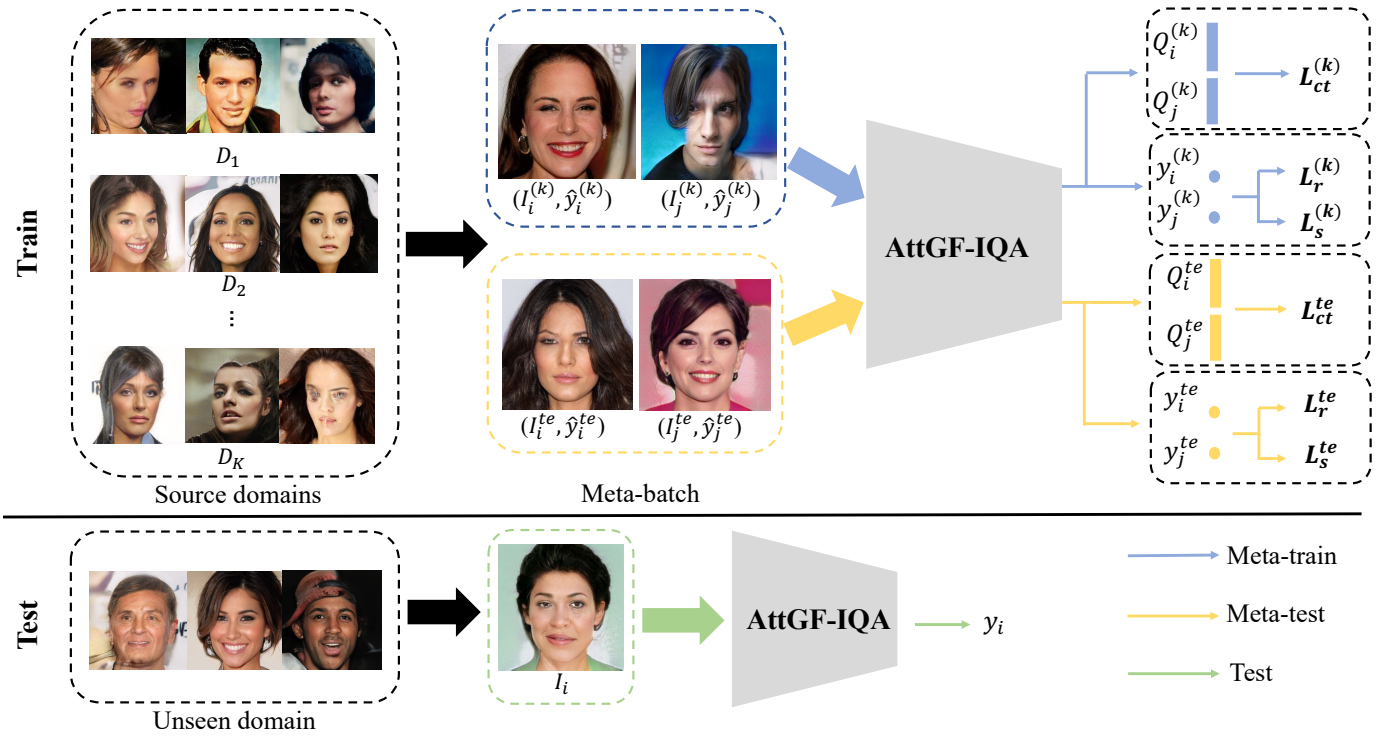


Fig. 5. The overview of the meta-learning strategy. The purpose is to improve the generalization ability of the proposed model. The "Train" part is the training process of our model. We split source domains into a meta-batch to simulate the real train/test data shift at every training iteration. The meta-batch includes meta-train and meta-test sub-batches, and all sub-batches work together for the optimization updates of our model. Finally, the trained model is tested on the unseen domain.

the inputs of the first ABA module after t-th CBA are the visual attention feature V_t and a face segmentation map S generated by a pre-trained face parsing network BiSeNet [47]. Herein, to incorporate the semantic information into the attention features, the conditional normalization is adopted, grounded on the philosophy that different attributes should own different feature statistics [11]. In particular, we first use batch-normalization layer (denoted as BN) to normalize the attention feature V_t as follows,

$$\mu_c = \frac{1}{NHW}\sum_{nyx} V_{n,c,y,x}, \tag{9}$$

$$\sigma_c = \sqrt{\frac{1}{NHW}\left(\sum_{nyx} V_{n,c,y,x}^2 - \mu_c^2\right)}, \tag{10}$$

where H, W and C are the height, width, and the number of channels in V. N is the number of training samples in a batch. The normalized V_t is denoted as $\bar{V}_t$. Following the $\bar{V}_t$, inspired by SPADE [53], we propose an attributes-guided transform (AGT) sub-module to explore the guidance information of the segmentation map and generate affine transformation parameters (γ, β) for $\bar{V}_t$. As shown in Fig. 7(b), our AGT consists of two convolutional layers with their kernels set as 3×3 for the γ, β extraction. Subsequently, we compute the attributes-guided activation feature T by denormalizing the feature $\bar{V}$ according to the affine transformation parameters (γ, β),

$$T = \gamma \otimes \bar{V} + \beta, \tag{11}$$

where the dimensions of γ and β are the same as $\bar{V}$. Then, we use 3×3 convolution after every AGT sub-module to obtain the attribute-based features. For the given input V_t, the output of the ABA module is denoted as Att_{2t-1}. We concatenate attribute-based features at different layers of the network to generate the multi-scale attribute-based feature Q. Finally, the predicted quality score y is defined as:

$$\begin{aligned} y &= \text{FC}(Q) \\ &= \text{FC}([Att_2, Att_4, Att_6, Att_8]), \end{aligned} \tag{12}$$

where Att_i presents the output of i-th attribute-guided transform in Fig. 6. FC is the fully connected layer.

D. Meta-Learning Strategy

As discussed above, the main goal of our IQA model is to achieve a high genralization capability even on the face images generated by unseen GAN models. To this end, the shareable and transferable features should be extracted from the source domains, aiming to mitigate the domain gaps caused by different GAN models. Following this vein, the meta-learning strategy is utilized in our method. To begin with, we first split a meta-batch from source domains at each iteration and a meta-batch contains several sub-batches. Every sub-batch is further sampled as a meta-test dataset (denoted as D_{te}) and the rest as meta-train datasets (denoted as D_{tr}). The training process of the proposed model on a sub-batch is shown in the top of the Fig. 5 that comprises two stages: (1) meta-train on D_{tr}; (2) meta-test on D_{te}. During the meta-train phase, we denote the image pair as $(I_i^{(k)}, \hat{y}_i^{(k)})$ and $(I_j^{(k)}, \hat{y}_j^{(k)})$,

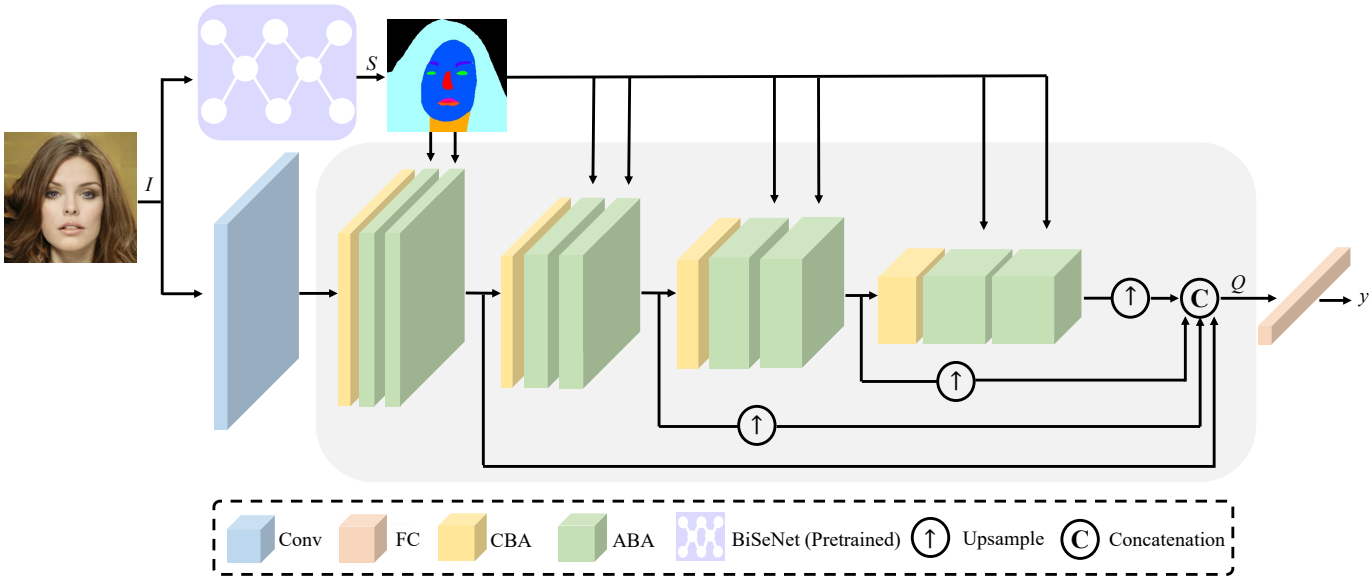


Fig. 6. The architecture of the proposed AttGF-IQA, including the CBA module and the ABA module.

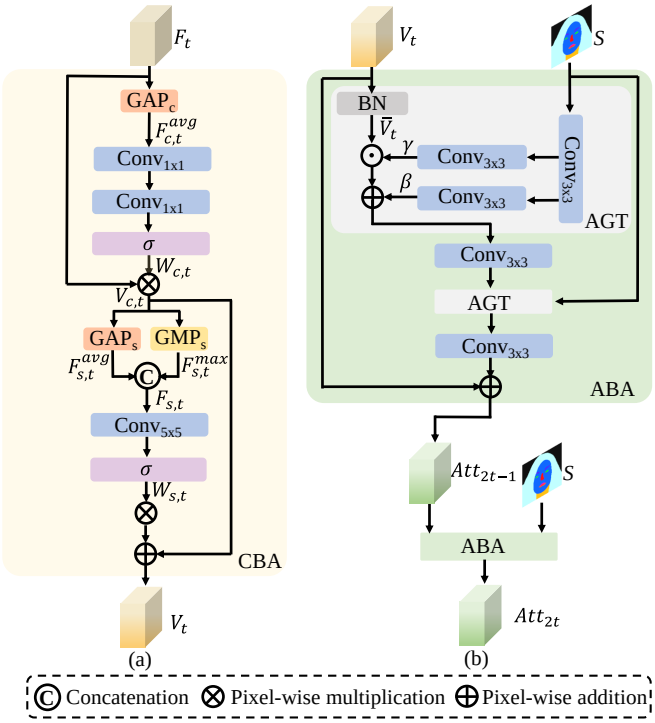


Fig. 7. The detailed structure of (a) CBA and (b) ABA.

where $I_i^{(k)}$ and $I_j^{(k)}$ are the i-th and j-th images from k-th D_{tr}, respectively. $\hat{y}_i^{(k)}$ and $\hat{y}_j^{(k)}$ are the corresponding pseudo-MOS values of $I_i^{(k)}$ and $I_j^{(k)}$, respectively. The quality scores of $I_i^{(k)}$ and $I_j^{(k)}$ predicted by our proposed model are $y_i^{(k)}$ and $y_j^{(k)}$. Then, we can compute the rank loss $L_r^{(k)}$ according to the predicted probability of $I_i^{(k)}$ is better than $I_j^{(k)}$ [54]. The predicted probability $p_{ij}^{(k)}$ is formulated as:

$$p_{ij}^{(k)} = \begin{cases} \frac{\exp(y_i^{(k)} - y_j^{(k)})}{1+\exp(y_i^{(k)} - y_j^{(k)})}, & \text{if } \hat{y}_i^{(k)} \geq \hat{y}_j^{(k)}, \\ 1 - \frac{\exp(y_i^{(k)} - y_j^{(k)})}{1+\exp(y_i^{(k)} - y_j^{(k)})}, & \text{otherwise.} \end{cases} \tag{13}$$

The rank loss $L_r^{(k)}$ aims to minimize the difference between the predicted probability and the ground-truth ranking order.

$$L_r^{(k)} = \frac{1}{N} \sum_{i,j}^{N} -(1 - p_{ij}^{(k)})^{\gamma} \log(p_{ij}^{(k)}), \tag{14}$$

where γ is the focusing parameter for adjusting the importance of easy samples and hard samples. N is the number of training image pairs in k-th D_{tr}.

Although the meta learning strategy adopted here can align the distributions among different domains, the discrimination capability of the learned model is still not ensured [55]. To address this issue, the center loss [56], [20] is further introduced, aiming to align the conditional distributions of different domains. In particular, we learn two renewable prototypes (centers) as the conditions of the ranking features, then the intra-class distances of the ranking features are minimized. More specifically, the ranking feature $R_{ij}^{(k)}$ is the difference between the features $Q_i^{(k)}$ and $Q_j^{(k)}$ before the last fully connected layer:

$$R_{ij}^{(k)} = Q_i^{(k)} - Q_j^{(k)}. \tag{15}$$

Thus, the center loss is computed as

$$\begin{aligned} L_{ct}^{(k)} = & \frac{1}{N} \sum_{i,j}^{N} (\delta(\hat{y}_i^{(k)} < \hat{y}_j^{(k)}) \left\| R_{ij}^{(k)} - c_0 \right\|_2^2 \\ & + \delta(\hat{y}_i^{(k)} \geq \hat{y}_j^{(k)}) \left\| R_{ij}^{(k)} - c_1 \right\|_2^2), \end{aligned} \tag{16}$$

where $\delta(condition) = 1$ if the condition is satisfied, and $\delta(condition) = 0$ otherwise. c_0 and c_1 are the learned class

Algorithm 1: Meta-learning optimization

input: Source domains (GANs) $D = \{D_1, D_2, \ldots, D_K\}$.
Init : The AttGF-IQA model f_θ parametrized by θ, learning rate β.

for *iteration = 1,2, ...* **do**
  // For a meta-batch
  Initialize the gradient g_θ as 0;
  for *each* D_{te} *in* D **do**
    // For a sub-batch
    Sample remaining $K-1$ domains (GAN models) as D_{tr};
    Meta-train:
    Compute the meta-train loss on the parameter θ by:
    $L_{tr} = \sum_{k=1}^{K-1}(\lambda_0 L_r^{(k)} + \lambda_1 L_{ct}^{(k)} + \lambda_2 L_s^{(k)})$
    Update the parameters θ by:
    $\theta^{'} = Adam(L_{tr}, \theta)$ on D_{tr}
    Meta-test:
    Compute the meta-test loss on the updated parameter $\theta^{'}$ by:
    $L_{te} = \lambda_3 L_r^{te} + \lambda_4 L_{ct}^{te} + \lambda_2 L_s^{te}$
    Aggregate gradient:
    $g_\theta \leftarrow g_\theta + \bigtriangledown_\theta L_{tr} + \bigtriangledown_\theta L_{te}$
  end
  update $\theta \leftarrow \theta - \beta \frac{1}{N} g_\theta$
end

centers of the ranking features $R_{ij}^{(k)}$. N is the number of training image pairs.

For the images collected from the face image inpainting model, their corresponding pseudo-MOS values are computed by the FR-IQA objective model IW-SSIM. Those objective scores reflect the similarity between real and generated images in terms of low-level features (e.g., image structure and texture). Thus, we add the score regression loss when training data contains the inpainted face images. Supposing GFIs pairs of α-th source domain is collected from face image inpainting model, the score regression loss is defined as:

$$L_s^{(k)} = \begin{cases} \frac{1}{M}\sum_m^M (\hat{y}_m^{(\alpha)} - y_m^{(\alpha)}), & \text{if } k = \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

where M is the number of training images in the α-th source domain.

The loss function of the meta-train can be summarized as follows,

$$L_{tr} = \sum_{k=1}^{K-1}(\lambda_0 L_r^{(k)} + \lambda_1 L_{ct}^{(k)} + \lambda_2 L_s^{(k)}), \tag{18}$$

where $K-1$ is the number of meta-train databases during the meta-train procedure. λ_0, λ_1 and λ_2 are the weighting factors of $L_r^{(k)}$, $L_{ct}^{(k)}$ and $L_s^{(k)}$, respectively.

After meta-training, the model with the knowledge learned from meta-train datasets D_{tr} needs to remain effective when testing on the meta-test dataset D_{te}. Analogously, we also compute the meta-test loss trained on the meta-test dataset D_{te}. As such,

$$L_{te} = \lambda_3 L_r^{te} + \lambda_4 L_{ct}^{te} + \lambda_2 L_s^{te}, \tag{19}$$

where λ_3 and λ_4 is the weighting factors of L_r^{te} and L_{ct}^{te}, respectively.

We define the AttGF-IQA model represented by a parametrized function f_θ with parameters θ. Following the similar strategies as suggested in [17], we update the parameters θ via Stochastic Gradient Descent (SGD). The whole meta-leaning procedure is summarized in Algorithm 1.

V. Experiments

A. Implementation Details

The proposed GFID database consists of training GFIs pairs and testing GFIs which are collected from four different GAN algorithms, and we treat GFIs generated by each GAN model as an independent domain. Therefore, there are four domains in our proposed GFID database. To simulate the cross-domain scenario, three domains are selected as the source domains for training, and the remaining one domain is used for testing which is strictly unseen in the training phase.

We implement our model by PyTorch [60]. The inputs of our network are resized to $256 \times 256 \times 3$, and the batch size is set as 10. We apply the Adam [61] optimizer with a learning rate of $1e-5$ and a weight decay of $5e-5$. The learning rate is reduced by 5 after every 10 epochs and the maximum epoch is 100. The weighting parameters λ_0, λ_1, λ_2, λ_3, and λ_4 in Eqn. (18) and Eqn. (19) are set as 10, 0.01, 1, 10, and 0.01, respectively.

B. Evaluation Criteria

To compare the performance of our proposed model with other IQA models, we employ two common criteria, i.e., Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC). The higher values of SRCC and PLCC, the better prediction performance the model can achieve. As suggested in the [62], [63], before computing those correlation coefficients, we adopt a nonlinear logistic mapping function to map the dynamic range of the predicted scores from various IQA metrics onto a common scale. The mapped score s_i of i-th face image can be computed as

$$s_i = \beta_1 \left(\frac{1}{2} - \frac{1}{e^{\beta_2 (y_i - \beta_3)}} \right) + \beta_4 y_i + \beta_5, \tag{20}$$

where y_i is the predicted scores of i-th face image from an IQA metric. β_1, β_2, β_3, β_4 and β_5 are to be determined by minimizing the sum of squared errors between the mapped score s_i and the subjective score.

C. Performance Comparisons

To explore the generalization of the proposed model on the unseen domain, we select several NR-IQA metrics for performance comparison, including GMM-GIQA [24], NIQE [31], FRIQUEE [57], MEON [32], NIMA [58], Meta-IQA [17], dipIQ [35], UNIQUE [59], where GMM-GIQA is an IQA

TABLE II
GENERALIZATION PERFORMANCE COMPARISONS OF THE PROPOSED ATTGF-IQA AND STATE-OF-THE-ART NR-IQA MODELS ON THE UNSEEN DOMAIN IN THE GFID TESTING DATASET, WHERE THE BEST, THE SECOND-BEST AND THE THIRD-BEST RESULTS ARE BOLDFACED IN RED, BLUE, AND BLACK, RESPECTIVELY.

Unseen Domain	Measures	NIQE [31]	FRIQUEE [57]	MEON [32]	NIMA [58]	Meta-IQA [17]	dipIQ [35]	UNIQUE [59]	RankIQA [18]	GMM-GIQA [24]	AttGF-IQA (Ours)
StyleGAN	SRCC	0.0919	0.1315	-0.0514	0.0396	0.0955	0.2284	0.0463	**0.4658**	**0.5101**	**0.7326**
	PLCC	0.0396	0.1257	-0.0053	0.0120	0.1762	0.2301	0.0694	**0.6044**	**0.5743**	**0.7538**
StarGAN	SRCC	-0.0483	0.2531	0.0168	0.2711	0.2700	0.0676	0.1055	**0.3275**	**0.3676**	**0.6459**
	PLCC	-0.0640	0.2655	-0.0215	0.2794	0.2814	0.0599	0.1017	**0.3214**	**0.3537**	**0.6495**
ProGAN	SRCC	0.0092	0.1832	0.0834	0.0567	0.1434	0.1640	0.0286	**0.4215**	**0.4726**	**0.7728**
	PLCC	-0.0064	0.1654	0.1215	0.0395	0.1583	0.2162	0.0453	**0.4595**	**0.4914**	**0.7791**
PENNet	SRCC	-0.1160	0.3073	0.1338	0.1757	0.2661	-0.0633	0.3182	**0.3436**	**0.3809**	**0.6952**
	PLCC	-0.1561	0.3283	0.1610	0.1850	0.2323	-0.0479	0.3289	**0.3806**	**0.3906**	**0.6718**

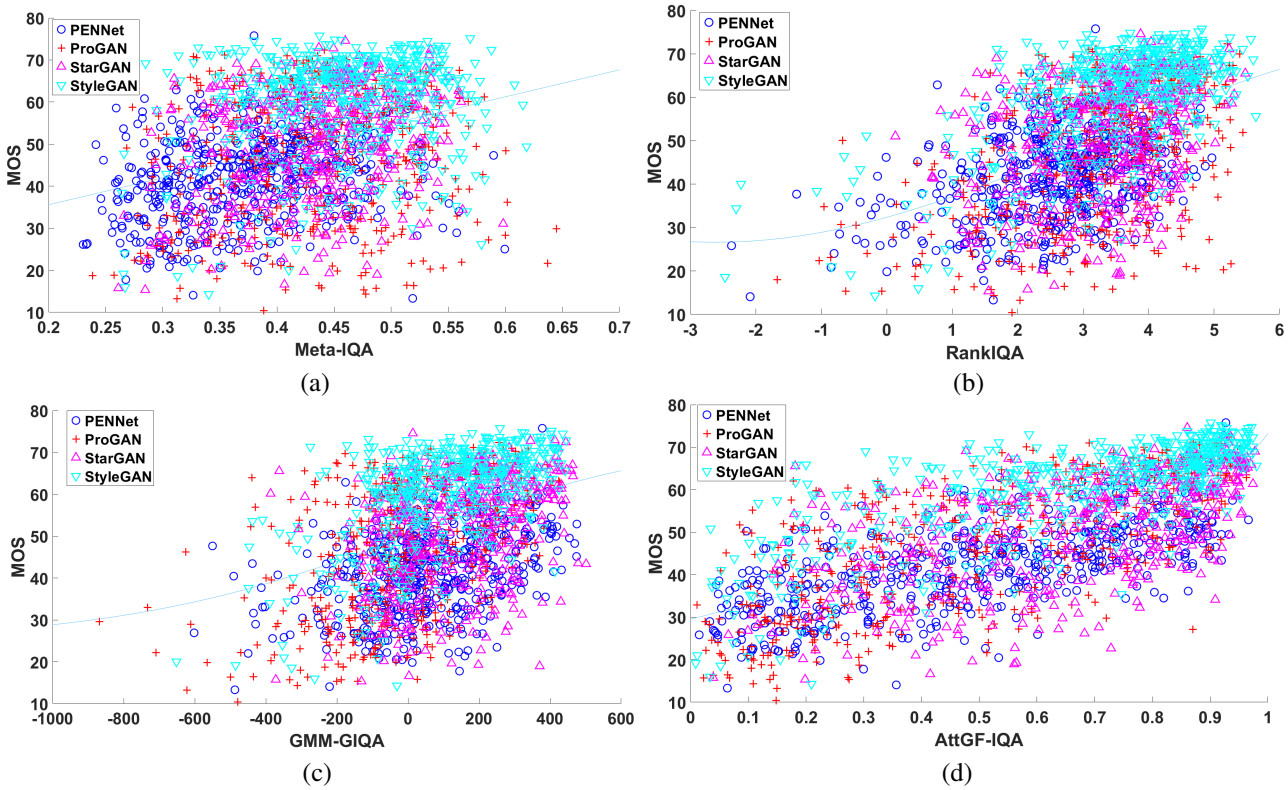


Fig. 8. Scatter plots of the subjective scores (i.e., MOS values) versus the objective scores computed by NR-IQA models on the GFID testing dataset. (a) Meta-IQA; (b) RankIQA; (c) GMM-GIQA; (d) the proposed AttGF-IQA trained on GFIs pairs of StarGAN, StyleGAN, and PENNet in the GFID training dataset.

model specifically for GIQA tasks. It is worth mentioning that the pre-trained versions of the competing models provided by the corresponding authors are used for performance comparison except for GMM-GIQA. For GMM-GIQA, the GMM used in all testing experiments is the same as the GMM built for the generation of pseudo-MOS in our proposed training dataset.

Table II shows the performance of the proposed model and different NR-IQA models when testing on the unseen domain in the GFID testing dataset. It is worth mentioning that the proposed AttGF-IQA is trained on three source domains only in the GFID training dataset. The experimental results show that the proposed AttGF-IQA trained on available source domains can perform well on the unseen domain, and its prediction performance outperforms all competing models. The main reason is that most models are designed for evaluating the quality of natural images whose distortions differ widely from GAN-generated distortions. In particular, GAN-generated distortions are related to the network architectures of generative models, which are diverse and difficult to predict. Besides, the features extractor of GMM-GIQA is the pre-trained Inception-v3 network [23] designed for image classification tasks, ignoring the influence of human visual perception. To further demonstrate the effectiveness of our proposed model specifically for evaluating restored images, we

TABLE III
PERFORMANCE COMPARISONS OF THE PROPOSED MODEL AND THE STATE-OF-THE-ART FR IQA MODELS ON GFIS GENERATED BY PENNET IN THE GFID TESTING DATASET. THE PROPOSED MODEL IS TRAINED WITH GFIS PAIRS OF PROGAN, STYLEGAN, AND STARGAN IN THE GFID TRAINING DATASET.

Methods	SRCC	PLCC
PSNR	0.4891	0.4776
SSIM [12]	0.3218	0.3358
IW-SSIM [46]	0.5781	0.5629
VSI [25]	0.4493	0.3084
VIF [26]	0.4046	0.4036
AttGF-IQA	**0.6952**	**0.6718**

TABLE IV
SRCC AND PLCC RESULTS OF THE PROPOSED MODEL WHEN TRAINING ON SOURCE DOMAINS IN THE GFID TRAINING DATASET AND TESTING ON BOTH SOURCE AND UNSEEN DOMAINS IN THE GFID TESTING DATASET.

Unseen	SRCC			
Domain	StyleGAN	StarGAN	ProGAN	PENNet
StyleGAN	0.7326	0.5967	0.7200	0.6856
StarGAN	0.7045	0.6459	0.7576	0.7428
ProGAN	0.7015	0.6409	0.7728	0.7733
PENNet	0.6858	0.6096	0.7520	0.6952
Unseen	PLCC			
Domain	StyleGAN	StarGAN	ProGAN	PENNet
StyleGAN	0.7538	0.5789	0.7194	0.6748
StarGAN	0.7815	0.6184	0.7791	0.7310
ProGAN	0.7717	0.6599	0.7638	0.7704
PENNet	0.7359	0.5718	0.7408	0.6718

also compare the proposed method with several state-of-the-art FR-IQA models, including the peak signal-to-noise ratio (PSNR), SSIM [12], IW-SSIM [46], VSI [25], VIF [26]. Table III documents the performance results of all models evaluated on GFIs generated by PENNet in the GFID testing dataset, where AttGF-IQA is the proposed model trained on GFIs pairs of StyleGAN, StarGAN, and ProGAN in the GFID training dataset. From the results, the proposed AttGF-IQA still acquire the best performance without the reference information of pristine images. Besides, from the results of Table II and Table III, we notice that although the GFID training dataset uses the quality scores predicted by GMM-IQA and IWSSIM as the pseudo-MOS, the proposed AttGF-IQA can achieve better performance than GMM-IQA and IWSSIM on the GFID testing dataset. This is because the proposed AttGF-IQA learns from quality-discriminable image pairs with relative quality rankings such that it can avoid the overfitting problem caused by overconfidence in the pseudo-MOS of each GFI.

Table IV shows the results when evaluating the models on both seen and unseen domains. In particular, each time we select on GAN algorithm as the unseen domain, and train the proposed model on the corresponding three source domains in the GFID training dataset. Given the trained model, we evaluate the performance on each domain of the testing data. From the experimental results, we can find that those models are still able to deliver promising performance for both seen and unseen domains when using the testing data with ground-truth MOS for evaluation. However, we can also observe relatively low SRCC and PLCC values on the quality assessment of GFIs generated by StarGAN. The reason is that most distortions in GFIs generated by StarGAN appear in facial contour and global spatial information, and the proposed framework fails to extract global image descriptor and reflect the correlation between facial attributes in the GFI. In Fig. 8, we further show the scatter plots of the MOS values against the objective scores as predicted by the RankIQA, GMM-GIQA, and AttGF-IQA, where the AttGF-IQA is trained on GFIs pairs of PENNet, StyleGAN, and StarGAN in the GFID training dataset, and the objective scores are the predicted results of GFIs of all GAN models in the GFID testing dataset. As shown in Fig. 8, compared with Meta-IQA, RankIQA, and GMM-GIQA, the proposed AttGF-IQA can better align with the MOS values. It implies the proposed AttGF-IQA is more consistent with visual quality of GFIs.

TABLE V
COMPARISON OF SRCC AND PLCC OF THE PROPOSED FRAMEWORK USING DIFFERENT OPTIMIZATION STRATEGIES ON UNSEEN DOMAINS IN THE GFID TESTING DATASET.

Methods	SRCC			
	Unseen Domain			
	StyleGAN	StarGAN	ProGAN	PENNet
s-StyleGAN	–	0.4512	0.6271	0.4817
s-StarGAN	0.4900	–	0.5726	0.4231
s-ProGAN	0.6222	0.5215	–	0.6582
s-PENNet	0.6309	0.4912	0.7052	–
w/o meta-learning	0.7026	0.5826	0.7304	0.6392
AttGF-IQA	**0.7326**	**0.6459**	**0.7728**	**0.6952**
Methods	PLCC			
	Unseen Domain			
	StyleGAN	StarGAN	ProGAN	PENNet
s-StyleGAN	–	0.4585	0.6328	0.4723
s-StarGAN	0.5457	–	0.5938	0.4232
s-ProGAN	0.6929	0.5296	–	0.6477
s-PENNet	0.7022	0.5381	0.7244	–
w/o meta-learning	0.7450	0.5921	0.7353	0.6330
AttGF-IQA	**0.7538**	**0.6495**	**0.7791**	**0.6718**

D. Ablation Studies

In this subsection, we conduct a series of ablation experiments to confirm the contribution of meta-learning strategy and different components of the proposed model on assessing GFIs quality.

1) Contribution of Meta-Learning Strategy: To investigate the contribution of meta-learning on the performance of our proposed model, we perform the ablation study based on different optimization strategies. Table V lists SRCC and PLCC results of the model evaluated on the unseen domain

TABLE VI
PERFORMANCE COMPARISONS OF THE PROPOSED MODEL WITH DIFFERENT COMPONENTS WHEN TRAINING WITH GFIS PAIRS OF STYLEGAN, STARGAN, AND PENNET IN THE GFID TRAINING DATASET AND TESTING WITH GFIS OF PROGAN IN THE GFID TESTING DATASET.

Methods	SRCC	PLCC
AttGF-IQA w/o CBA	0.7396	0.7496
AttGF-IQA w/o ABA	0.6923	0.6922
AttGF-IQA	**0.7728**	**0.7791**

in the GFID testing dataset when using different optimization strategies. In this table, **s-StyleGAN**, **s-StarGAN**, **s-ProGAN**, and **s-PENNet** represent models trained with GFIs pairs of StyleGAN, StarGAN, ProGAN, and PENNet in the GFID training dataset, respectively. From the results, one can easily observe that the proposed models trained on a single source domain have performance degradation at different degrees. This phenomenon reveals that there exists a domain gap between GFIs generated by different generative models. Moreover, the three source domains are merged as the training data, and we evaluate the model performance without meta-learning strategy. In particular, **w/o meta-learning** is the model trained with GFIs pairs of three corresponding source domains in the GFID training dataset and directly optimized by an Adam optimizer, and **AttGF-IQA** is the proposed model trained with GFIs pairs of three corresponding source domains in the GFID training dataset and optimized by meta-learning strategy. Compared with the results of **w/o meta-learning** and **AttGF-IQA**, we can observe that the proposed model with meta-learning achieves around 4.0%, 9.8%, 5.4%, 8.0% improvements on SROCC and 1.2%, 8.8%, 5.6%, 4.0% improvements on PLCC. Therefore, the proposed model using the meta-learning optimization strategy achieves higher prediction performance in this application.

2) Contribution of Different Components: To evaluate the contributions of different components, we conduct experiments to confirm the benefit from each component of the proposed AttGF-IQA. After removing CBA and ABA components, respectively, the results of those models when training on GFIs pairs of StyleGAN, StarGAN, and PENNet in the GFID training dataset and testing on GFIs of ProGAN in the GFID testing dataset are shown in Table VI. More specifically, **AttGF-IQA w/o CBA** means that the AttGF-IQA model removes all CBA modules. **AttGF-IQA w/o ABA** means that the AttGF-IQA model replaces two AGT of every ABA module with two batch normalization layers. From the table, we can see that the model removing the CBA or ABA module will cause performance degradation.

VI. CONCLUSION

In this paper, we focus on studying the GFIs quality assessment from both subjective and objective perspectives. Specifically, we establish the first database GFID for the GFIs quality assessment, which consists of training GFIs pairs with relative quality rankings and testing GFIs with human-annotated scores. Moreover, we design an objective AttGF-IQA model based on the special characteristics of face images and employ the meta-learning optimization strategy to improve generalization ability of the prediction model. Extensive simulation results demonstrate that the proposed AttGF-IQA model achieves higher prediction accuracy and generalization capability on the quality assessment of GFIs than state-of-the-art IQA methods.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Adv. Neural Inform. Process. Syst.*, pp. 2672–2680, 2014.

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *Adv. Neural Inform. Process. Syst.*, pp. 5767–5777, 2017.

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *Int. Conf. Learn. Represent.*, 2018.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4401–4410, 2019.

[5] T. Karras, S. Laine, M. Attala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8110–8119, 2020.

[6] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Starganv2: Diverse image synthesis for multiple domains," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8188–8197, 2020.

[7] A. Lahiri, A. K. Jain, S. Agrawal, P. Mitra, and P. K. Biswas, "Prior guided gan based semantic inpainting," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 13 696–13 705, 2020.

[8] Z. Ni, W. Yang, S. Wang, L. Ma, and S. Kwong, "Towards unsupervised deep image enhancement with generative adversarial network," *IEEE Trans. Image Process.*, vol. 29, pp. 9140–9151, 2020.

[9] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5549–5558, 2020.

[10] N. Yang, Z. Zheng, M. Zhou, X. Guo, L. Qi, and T. Wang, "A domain-guided noise-optimization-based inversion method for facial image manipulation," *IEEE Trans. Image Process.*, vol. 30, pp. 6198–6211, 2021.

[11] Y. Shen, C. Yang, X. Tang, and B. Zhou, "Interfacegan: Interpreting the disentangled face representation learned by gans," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[12] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2016.

[13] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, 2014.

[14] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[15] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structure similarity estimation," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, 2012.

[16] S. Golestaneh and L. J.Karam, "Reduced-reference quality assessment based on the entropy of dwt coefficients of locally weighted gradient magnitudes," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5293–5303, 2016.

[17] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14 143–14 152, 2020.

[18] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," *Int. Conf. Comput. Vis.*, 2017.

[19] J. Ma, J. Wu, L. Li, W. Dong, X. Xie, G. Shi, and W. Lin, "Blind image quality assessment with active inference," *IEEE Trans. Image Process.*, vol. 30, pp. 3650–3663, 2021.

[20] B. Chen, H. Li, H. Fan, and S. Wang, "No-reference screen content image quality assessment with unsupervised domain adaptation," *IEEE Trans. Image Process.*, 2021.

[21] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Adv. Neural Inform. Process. Syst.*, pp. 2234–2242, 2016.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Adv. Neural Inform. Process. Syst.*, pp. 6626–6637, 2017.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2818–2826, 2016.

[24] S. Gu, J. Bao, D. Chen, and F. Wen, "Giqa: Generated image quality assessment," *Eur. Conf. Comput. Vis.*, pp. 369–385, 2020.

[25] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, 2014.

[26] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[27] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: Jpeg2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, 2005.

[28] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 717–728, 2009.

[29] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, 2011.

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

[31] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

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

[33] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, 2020.

[34] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3667–3676, 2020.

[35] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, 2017.

[36] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," *Inf. Conf. Mach. Learn*, pp. 89–96, 2005.

[37] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1486–1494, 2019.

[38] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5400–5409, 2018.

[39] P. Yang and W. Gao, "Multi-view discriminant transfer learning," *IJCAI*, pp. 1848–1854, 2013.

[40] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba1, "Undoing the damage of dataset bias," *Eur. Conf. Comput. Vis.*, pp. 158–171, 2012.

[41] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," *Int. Conf. Comput. Vis.*, pp. 5542–5550, 2017.

[42] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," *Int. Conf. Learn. Represent.*, 2018.

[43] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," *Int. Conf. Learn. Represent.*, 2021.

[44] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," *AAAI*, 2018.

[45] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3516–3525, 2020.

[46] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, 2011.

[47] C. Yu, J. Wan, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," *Eur. Conf. Comput. Vis.*, pp. 325–341, 2018.

[48] "Methodology for the subjective assessment of the quality of television pictures," *ducument Rec. ITU-R BT.500-13*, 2012.

[49] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "Esim: Edge similarity for screen content image quality assessment," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4818–4831, 2017.

[50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *Eur. Conf. Comput. Vis.*, pp. 3–19, 2018.

[51] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," *Int. Conf. Learn. Represent.*, 2020.

[52] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Describable visual attributes for face verification and image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1962–1977, 2011.

[53] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2337–2346, 2018.

[54] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *Int. Conf. Comput. Vis.*, 2017.

[55] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, and Q. He, "Deep subdomain adaptation network for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1713–1722, 2020.

[56] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," *Eur. Conf. Comput. Vis.*, pp. 499–515, 2016.

[57] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, 2017.

[58] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, 2018.

[59] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Trans. Image Process.*, vol. 30, pp. 3474–3486, 2021.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, and Z. L. et al., "Pytorch: An imperative style, high-performance deep learning library," *Adv. Neural Inform. Process. Syst.*, vol. 32, pp. 8026–8037, 2019.

[61] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *Int. Conf. Learn. Represent.*, 2015.

[62] V. Q. E. Group, "Final report from the video quality experts group on the validation of objective models of video quality assessment," *VQEG meeting*, 2020.

[63] Z. Ni, H. Zeng, L. Ma, J. Hou, J. Chen, and K.-K. Ma, "A gabor feature-based quality assessment model for the screen content images," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4516–4528, 2018.