Subjective and Objective Quality Assessment of Multi-Attribute Retouched Face Images

Guanghui Yue, *Member, IEEE*, Honglv Wu, Weiqing Yan, Tianwei Zhou, Hantao Liu, and Wei Zhou

***Abstract*—Facial retouching, aiming at enhancing an individual's appearance digitally, has become popular in many parts of human life, such as personal entertainment, commercial advertising, etc. However, excessive use of facial retouching can affect public aesthetic values and accordingly induce issues of mental health. There is a growing need for comprehensive quality assessment of Retouched Face (RF) images. This paper aims to advance this topic from both subjective and objective studies. Firstly, we generate 2,500 RF images by retouching 250 high-quality face images from multiple attributes (i.e, eyes, nose, mouth, and facial shape) with different photo-editing tools. After that, we carry out a series of subjective experiments to evaluate the quality of multi-attribute RF images from various perspectives, and construct the Multi-Attribute Retouched Face Database (MARFD) with multi-labels. Secondly, considering that retouching alters the facial morphology, we introduce a multi-task learning based No-Reference (NR) Image Quality Assessment (IQA) method, named MTNet. Specifically, to capture high-level semantic information associated with geometric changes, MTNet treats the alteration degree estimation of retouching attributes as auxiliary tasks for the main task (i.e., the overall quality prediction). In addition, inspired by the perceptual effects of viewing distance, MTNet utilizes a multi-scale data augmentation strategy during network training to help the network better understand the distortions. Experimental results on MARFD show that our MTNet correlates well with subjective ratings and outperforms 16 state-of-the-art NR-IQA methods.**

***Index Terms*—Image quality assessment, subjective and objective quality assessment, multi-attribute facial retouching, multi-task learning.**

I. Introduction

FACIAL retouching is a digital beauty technology that has been ubiquitous in personal entertainment and commercial advertising. This technology can edit the face image digitally and make changes in the spatial domain, e.g., slimming of facial contours, enlarging of the eyes, correction of the mouth, causing facial alterations similar to those achieved by plastic surgery [1]. Although improving attractiveness and beauty of photos, it has also raised controversy because of luring people into pursuing an unrealistic representation of physical beauty, which may mislead public aesthetic values and result in mental anxiety for body control [2]. Additionally, inappropriate use of Retouched Face (RF) images brings a challenge for face recognition and poses security risks in some specific application scenarios, such as authentication, transaction, sentiment analysis, plastic surgery planning, etc. Recently, there have been increasing calls for legislation that the RF images should be labeled when used due to concerns for the general issue of truth in advertising and for public health. In this context, quality assessment of RF images is emerged as the times require [3].

Image quality assessment (IQA), as an essential research topic for many image processing algorithms, can be divided into subjective and objective methods [4]–[8]. The former evaluates image quality using subjective experiments with strict scoring rules. Specifically, a group of qualified observers is recruited to rate the images. This method directly reflects the results of human perception and provides reliable quality scores. It has two essential roles. On the one hand, the subjective method is usually used to generate IQA databases, in which the rating scores can be serve as the ground truth to validate and compare objective methods. The predictive scores of a superior objective method have higher consistency with the subjective rating scores. On the other hand, the analysis of subjective behaviors can provide inspirations for designing effective objective methods. Researchers can add specific focuses in objective method design by analyzing the relationship between image distortions and rating scores.

In the literature, the subjective method has been widely adopted to investigate the perceptual quality of different types of images, such as natural scene images [9], [10], screen content images [11], animation images [12], etc., resulting in many IQA databases. However, related works regarding RF images are quite scarce. Recently, researchers have begun to detect the presence or absence of facial retouching through a binary classification task [13]–[15]. However, such methods cannot inform photo editors and observers of how much an RF image has strayed from reality. Generally, a continuous rating can directly indicate the amount by which a person's appearance has been perceptually altered, helping audiences better understand the authenticity of content when using RF images. More recently, Kee and Farid proposed to subjectively evaluate the alteration degree of retouched images [3] and reported a database with 468 raw and retouched image pairs collected from various on-line sources. Since this database was

This work was supported in part by National Natural Science Foundation of China (Nos. 62001302, 62371305, 62103286), in part by Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011348), in part by the Youth Innovation Team Project of Higher Education Institutions in Shandong Province under Grant 2022KJ268, and in part by the Science and Technology Innovation Development Plan Projects of Yantai under Grant 2023JCYJ046.. (Corresponding author: Weiqing Yan.)

G. Yue and H. Wu are with the School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen 518060, China (e-mail: yueguanghui@szu.edu.cn; kailyn_wu@126.com).

W. Yan is with the School of Computer and Control Engineering, Yantai University, Yantai, 261400, China (e-mail: wqyan@tju.edu.cn).

T. Zhou is with the College of Management, Shenzhen University, Shenzhen 518060, China (e-mail: tianwei@szu.edu.cn).

H. Liu and W. Zhou are with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, United Kingdom (e-mail: liuh35@cardiff.ac.uk; zhouw26@cardiff.ac.uk).

not specialized in facial retouching and the number of images was small, more specific databases are required to promote the development of quality assessment of RF images.

Existing objective works mainly focus on natural scene images (NSIs), where distortions are usually characterized by compression, noise, blur, contrast change, etc. To quantify the distortions, early researches conducted extensive attempts in mining effective handcrafted features. These attempts were mainly from two perspectives: extracting features based on natural scene statistics (NSS) and extracting features from the mathematical model of human visual system (HVS). The former is motivated by the observation that the presence of distortions can measurably modify the regular statistical properties of natural images [16]. Representative NSS-based features are the local normalized intensity coefficients in the spatial domain [17], in the DCT domain [18], etc., and the global statistical coefficients in the spatial and transform domains, such as intensity histogram [19], high-order statistics [20], and entropy [21]. The latter is inspired by the fact that features from the well-built HVS model can, to some extent, reflect the perceptual response of how a distorted image affects visual experience. Among HVS-inspired methods, representative works mainly extracted biologically inspired features by simulating the responses of opponent cells [22], considering luminance masking and contrast sensitivity effects [23], building contrast sensitivity function [24], and so on. Although these methods have achieved remarkable success in some simple IQA scenarios, e.g., pre-defined synthetic distortions that are with inherent regular properties, they usually fail in handling the IQA tasks in more complex scenarios, e.g., those have authentic and geometric distortions [25]. Different from NSIs, the distortions in RF images are caused by beauty algorithms or photo-editing tools, usually resulting in geometric modifications on key facial features, e.g., deformation of eyes and protrusion of the nose. As such, more advanced IQA methods are required.

Recently, deep neural networks (DNNs) have been considered as a powerful alternative to traditional handcrafted feature based methods in complex IQA scenarios because they can automatically extract and integrate low-level spatial distortions and high-level semantic distortions of an input image [26]. Due to the limited labeled data, early works mainly began by dividing the input image into multiple small patches for data augmentation and building very shallow convolutional neural networks (CNNs) for quality estimation [27]. Considering the position relationships between patches and their entire image, the adaptive weighting strategy was utilized when integrating predictive scores from different patches [28]. Later, pre-trained deep CNNs designed specifically for image classification were adopted and modified from different aspects, such as fully connected layers, global and local feature fusion, etc., to meet the requirement of IQA [29], [30]. In order to fully train the network, the input image was randomly cropped multiple times, generating many large patches. To learn more discriminative feature representation, the multi-task learning was adopted by setting relevant auxiliary task, such as distortion recognition [31], saliency detection [32], etc., for the main task (i.e., the overall quality estimation). While praising the achieved remarkable success, we have to notice that CNN-based methods often possess limited ability in characterizing global distortions because of the limited receptive field of convolutional operations. More recently, motivated by the fact that Transformer is able to model long-range dependencies, increasing efforts have been made in utilizing Transformer in a reasonable manner for accurate IQA [33]. Representative works were reported from leveraging multi-scale representation [34] and utilizing attention-panel mechanism [35]. Some works also incorporated the strengths of CNN and Transformer for more accurate predictions [36], [37]. Although these DNN-based methods have performed effectively on various conventional IQA tasks, their effectiveness in handling RF data requires further investigation.

This paper makes a comprehensive study on the quality assessment of multi-attribute RF images. On the one hand, we selected 250 high-quality face images and utilized four popular photo-editing tools to retouch them automatically in various styles, resulting in a collection of 2,500 retouched images. Then, we carried out a series of subjective experiments to evaluate the perceptual quality of RF images from multiple perspectives. Based on the subjective data, we constructed a Multi-Attribute Retouched Face Database (MARFD) with multi-labels for RF images, and concluded the primary factors impacting the perceptual quality. Our findings and insights highlight the key points of quality assessment of RF and provide inspiration for the follow-up design of objective IQA methods. On the other hand, according to these findings, we introduced a simple yet effective multi-task learning based IQA method, termed MTNet, for evaluating the quality of RF images in a no-reference (NR) manner. MTNet can objectively measure how much an RF image has strayed from reality, providing a certain reference for audiences. More specifically, it utilized a classical CNN as the feature extractor and adopted a multi-task learning strategy to simultaneously estimate the alteration degree of facial features and perceptual quality score of the RF image. Inspired by the perceptual effects of viewing distance, a multi-scale training strategy is applied to help the network better understand the distortions in RF images. The contributions of this paper are as follows:

- We conduct a series of subjective studies to comprehensively evaluate the perceptual quality of multi-attribute RF images from diverse perspectives and construct a new IQA database, namely MARFD[1]. A total of 2,500 RF images are generated by processing 250 high-quality face images with 4 popular photo-editing tools under different settings. Each RF image is labeled with 5 continuous quality scores from different perspectives, including the overall quality and alteration degree of each attribute (i.e., eyes, nose, mouth, and facial shape).
- We undertake a thorough discussion of the newly constructed MARFD, highlighting the key points we should notice when designing objective methods. We also evaluate 16 state-of-the-art NR-IQA methods on MARFD, providing a thorough summary of the benchmarking

[1]This database will be available at https://github.com/hhhollyjones for academic purposes upon the acceptance of this paper.

results and insights into areas for improvement.
- According to the findings from subjective experiments, we introduce a multi-task IQA network, termed MTNet. Considering the geometric distortions of RF images, MTNet utilizes four auxiliary tasks, i.e., alteration degree estimation of retouching attributes, to enhance the feature representation and improve the performance of the main task, i.e., overall image quality prediction. Extensive experiments demonstrate that our MTNet correlates well with subjective ratings and is more competent for the RF IQA task than 16 state-of-the-art NR-IQA methods.

II. Related Works

A. Image Quality Assessment Databases

In the past decades, the literature has accumulated remarkable achievements in constructing IQA databases for different application scenarios. Among numerous attempts, research on NSIs receives the earliest attention [38]. For instance, considering the distortions occurring at the stage of image acquisition, storage, and transmission, Sheikh *et al*. [9] constructed the legend IQA database LIVE, including five types of synthetic distortions. Similarly, Ponomarenko *et al*. [10] proposed the TID2013 that consists of 3,000 distorted images generated from 24 raw images using 25 synthetic distortions. In view of the actual situations, database with authentically distorted images was later reported [39]. Motivated by the success of relevant works on NSIs, many efforts have also been made in constructing IQA databases for other types of images/videos, such as screen content images [11], medical images [40], panoramic videos [41], etc. Recently, researchers began to investigate the perceptual quality assessment issues of images distorted by post-processing algorithms, e.g., stitching [25], super-resolution [42], enhancement [43], and so on.

Face IQA is a fundamental yet important research area in identity recognition system. For a long time, face IQA mainly aims to evaluate the impact degree of a distorted face image on the accuracy of identity recognition systems. To propel the development of this field, many databases have been reported recently, such as Adience [44], Cross-Quality LFW [45], AgeDB-30 [46], and so on. However, with the wide usage of RF images in personal entertainment and commercial advertising, it is highly desired to evaluate the quality assessment of an RF image from the perspective of digital photo alterations. In this context, different from previous works, the quality assessment discussed in this study refers to the perceptual alteration degree of a face image after the retouching operations. Recently, Kee and Farid [3] made a pioneering attempt to evaluate the alteration degree of retouched images through subjective experiments and constructed an IQA database, which includes 468 original and retouched images. Since each original image was only retouched once and most images included geometric alterations on the whole body rather than only on the facial regions, this database is not suitable for comparing and validating quality assessment methods for RF images.

B. Objective Image Quality Assessment Methods

Face IQA is a hot topic in recent years [47]. It involves assessing factors like clarity, lighting, and absence of occlusions or distortions that might impede accurate recognition. For instance, supervised face IQA methods like SDD-FIQA [48], MagFace [49], and CR-FIQA [47] have seen widespread use. These methods, along with unsupervised techniques like SER-FIQ [50] and FaceQAN [51], have demonstrated effectiveness across different face recognition models and datasets. Different from these works, the RF quality assessment discussed in this study aims to evaluate perceptual digital alteration degree caused by retouching operations.

For accurate image quality prediction, one of the key points is extracting effective features for distortion representation. Inspired by the statistical properties of natural images, Mittal *et al*. [17] measured the distortion degree of an image via the local normalized intensity coefficients in the spatial domain. Lamichhane *et al*. [52] introduced a NR quality metric for light field images via analysis of spatial and angular characteristics. Apart from NSS-based works, effective IQA methods have also been designed from the perspective of modeling the characteristics of HVS. Given that tone-mapping usually causes color distortions, Yue *et al*. [22] proposed to simulate the responses of the brain in processing color information and extract statistical features from the response maps to represent the distortions in tone-mapped images. Gu *et al*. [53] utilized both classical HVS-inspired features and free-energy-based features to measure the distortions in an image and build an IQA model by fusing these features via support vector regression.

Benefiting from the strong capability of automatic feature extraction and fusion, DNN-based IQA methods have attracted increasing attention and have gradually become an expected substitute for traditional handcrafted feature-based IQA methods. Kang *et al*. [27] introduced a shallow CNN that takes the image patches as the inputs for blindly evaluating image quality. Pan *et al*. [36] proposed a multi-branch network, in which the spatial-domain features, the gradient-domain features, and the weighting information were extracted in parallel. The features were fused based on the weighting information to obtain the image quality. Su *et al*. [29] introduced a self-adaptive hyper network that separated the NR IQA procedure into the stages of content understanding, perception rule learning, and quality predicting. Given that CNNs may not be sensitive to global distortions due to the limited receptive field of convolutional operations, You and Korhonen [33] added the Vision Transformer (ViT) on the top of a feature map extracted by CNN, and the features from ViT were fused via a multi-layer perception head to yield the image quality. Zhou *et al*. [7] proposed a U-shaped Transformer network for evaluating image quality in a NR manner. To thoroughly extract local and global distortions, Golestaneh *et al*. [36] incorporated CNNs and Transformer to extract both local and global features for quality assessment.

Fig. 1. Examples in our constructed MARFD: (a) raw face images, (b) retouched face images generated under different settings.

III. Multi-attribute Retouched Face Database

A. Image Collection and Processing

Generally, a standard IQA database should include diverse image contents to represent real-world scenarios as much as possible. In view of this, we collected 250 high-quality face images[2] from the FFHQ database [54]. According to practical considerations, certain standards are followed during data collection. Firstly, we focused on selecting images of Asians due to the shortage of databases specifically designed for Asian subjects in the literature. Secondly, we only chose images with no facial obstructions, adequate lighting, and neutral expressions, which made it easier to modify them using photo editing softwares and ensured clear observation of any possible retouching artifacts. Thirdly, images with rich contents and details in facial features (e.g., mouth, nose, eyes, etc.) and shapes, were primarily chosen, and all images maintained a relatively balanced age and gender distribution. To facilitate the design of subsequent objective IQA methods, all images have the resolution of 1024×1024 with the face centered. We treated these images as raw face images. Some examples are shown in Fig. 1(a).

After collecting the raw images, we employed photo-editing tools to retouch these images automatically in different styles. Four well-known photo-editing tools were utilized in this study: Adobe Photoshop[3], MeiTu[4], Luminar Neo[5], and Arcsoft Portrait+3[6]. These tools were selected for their user-friendly interfaces and comprehensive portrait editing functionalities that facilitate the automatic creation of diverse and refined facial images. Generally, geometric changes in facial structure have a more significant impact on appearance and can seriously tamper with identity information, posing challenges to identification. In addition, considering that retouched images with shape adjustment become increasingly popular in many aspects of social media, e.g., Twitter, Instagram, WeChat, we focused primarily on geometric changes in facial features (eyes, nose, mouth) and facial shapes, rather than photometric changes such as skin smoothing or brightening. The detailed settings for each photo-editing tool employed in our experiment are documented in Table I. It is worth noting that, we only considered the widely used functions of these tools in face retouching during parameter settings. This approach allowed us to generate a wide array of retouched images, each varying in the degree and style of alteration, thereby providing a comprehensive dataset for our analysis. The diversity in retouching styles, stemming from the distinct characteristics of each software, enriches our dataset and contributes to the robustness of our study in assessing the quality of retouched face images.

Based on the settings in Table I, we generated a total of 2,500 multi-attribute RF images from 250 raw face images. Fig. 1(b) provides some examples of RF images, based on which we have the following observations: 1) The retouching degree is highly related to the predefined parameters of photo-editing tools, and each tool has its own unique style for changing facial appearance. 2) The retouching effect is influenced by the image content. Images with different contents may have different response even processed by the same photo-editing tool with the same settings. 3) Due to the retouching on different attributes, both local and global geometric distortions exist in the RF images. 4) The retouching artifacts of facial parts are different. Overall, these observations highlight the significant challenges associated with assessing the perceptual quality of multi-attribute RF images. To better understand the main factors that impact the perceptual image quality, it is essential to undertake comprehensive subjective studies on all generated multi-attribute RF images. Such insights can then inform and guide the design of subsequent objective IQA methods, which would enable more accurate assessment of RF image quality.

B. Subjective Experiment

We recruited 23 observers (aged 20 to 30) to rate the perceptual quality of all collected multi-attribute RF images. All observers signed the written consent form. The subjective experiment was approved by the Medical Ethical Committee Approval of Shenzhen University Health Science Center (Number: PN-202400002). All observers had normal or correct-to-normal visions and had exposure to retouching

[2]All images are copyright of their rightful owners, and no copyright infringement is intended.

[3]https://www.adobe.com/products/photoshop.html?promoid=RBS7NL7F

[4]https://mt.meipai.com/

[5]https://skylum.com/luminar

[6]https://arcsoft-portrait.software.informer.com/3.0/

TABLE I
THE SETTINGS OF PHOTO-EDITING TOOLS.

Num.	Photo-editing Tools	Settings
1	Luminar Neo	[Eyes] Iris visibility: 80, Eye magnification: 80, Eye enhancement: 20.
2	Arcsoft Portrait+3	[Nose] Nose bridge: 50; [Mouth] Smile: 50.
3	MeiTu	[Eyes] Size: 100, Height: 80; [Nose] Size: 70, Nostrils: 100, Nose bridge: 100; [Mouth] Size: -80; [Facial shape] Width: 100, Jaw: 80.
4	Adobe Photoshop	[Eyes] Size: 50 50, Eye width: 50 50, Eye height: 50 50; [Nose] Nose width: -50; [Mouth] Smile: 30, Lip width: -100; [Facial shape] Chin height: 50, Jaw: -50, Face width: -50.
5	Adobe Photoshop	[Eyes] Size: 50 50, Eye width: 50 50, Eye height: 50 50; [Nose] Nose width: -200; [Mouth] Smile: 60, Lip width: -200; [Facial shape] Chin height: 50, Jaw: -50, Face width: -50.
6	Luminar Neo	[Eyes] Enlarge eyes: 170; [Facial shape] Slim face: 100.
7	Adobe Photoshop	[Eyes] Size:100 100, Eye width: 100 100, Eye height: 100 100; [Nose] Nose width: -100; [Mouth] Smile: 15, Lip width: -50; [Facial shape] Chin height: 50, Jaw: -50, Face width: -50.
8	MeiTu	[Eyes] Size: 180, Eye height: 150; [Nose] Size: 120, Nostrils: 150, Nose bridge: 150; [Mouth] Size: -150; [Facial shape] Width: 100, Jaw: 80.
9	MeiTu	[Eyes] Size: 180, Eye height: 150; [Nose] Size: 120, Nostrils: 150, Nose bridge: 150; [Mouth] Size: -150; [Facial shape] Width: 180, Jaw: 180, Chin height: 100.
10	Adobe Photoshop	[Eyes] Size:100 100, Eye width: 100 100, Eye height: 100 100; [Nose] Nose width: -50; [Mouth] Smile: 15, Lip width: -50; [Facial shape] Chin height: 100, Jaw: -100, Face width: -100.

images in daily life. According to ITU recommendations [55], during the subjective experiment, observers were asked to rate the quality of each RF image at a viewing distance of three times the image height. All images were presented on a 1920×1080 HP 27-inch screen using a designed rating software. Before the formal rating, a training session was prepared for all the observers. Specifically, we first introduced the experiment's goals and rules by PowerPoint presentations, showcased the differences in various levels of image retouching, and demonstrated how to use the rating software, ensuring that the observers fully understood the task requirements. Then, the observers were asked to score 20 raw/RF image pairs. Only qualified observers with rating accuracy over 80% were allowed to participate in the formal rating session. The number of final qualified observers was 20. Notably, these training images were not included in the subsequent formal rating stage as well as the final database.

Fig. 2 illustrates the graphical user interface of the used rating software. Specifically, the raw face image was displayed on the left, while two multi-attribute RF images were displayed on the right. By comparing the difference degree between left and right image, observers were asked to first rate the quality of the right image separately from four attributes (eyes, nose, mouth, facial shape), and then rate the overall quality. Since the retouching level of each attribute was limited during the RF image generation, the rating range here was relatively narrow, within a range of 1∼3. However, when incorporating the settings of all attributes, the generated RF images became more distinct and complex. In view of this, we set the rating range of the overall quality to 1∼5. A higher rating score indicates a larger difference between the RF image and its raw image. The presentation order of the images was randomized to minimize memory effects on ratings. Once the "Next" button was pressed, the rating software automatically recorded the rating scores of current two RF images. It is noteworthy that we randomly presented two RF images simultaneously to better compare their differences and reduce the likelihood of identical ratings. The observers were advised to take a break every 20 minutes to relieve accumulated visual fatigue and to ensure that their ratings remained reliable. In view of the number of generated RF images, we set six rating sessions for each participant on different days, and the image number in each rating session was 410, 410, 420, 420, 420, and 420, respectively. The average number of rated images per session was 417, and it took each participant approximately three hours to complete a rating session. Overall, a total of 250,000=2,500×20×5 subjective ratings were collected, where 2,500, 20, and 5 denote the image number, observer number, and ratings number per image, respectively.

Fig. 2. Graphical user interface designed for our subjective experiment. This interface displays three images side-by-side for comparison: the original reference image on the left, and two randomly retouched versions of this original image in the middle. Participants in the study are instructed to compare each retouched image with the original and assign attribute scores to the retouched images. These scores reflect the degree of retouching; a higher score indicates more alteration.

C. Mean Opinion Score Computation

Since observers may have different judgements on the same image due to their inconsistent understanding even though they receive the same training. Therefore, a data processing process is required before determining the final quality score of each RF image. First, we removed the outliers based on

data statistics. For a given image j, its mean score $\overline{\mu}_j$ and standard deviation s_j across all observers are calculated by

$$\overline{\mu}_j = \frac{1}{N}\sum\nolimits_{i=1}^{N} r_{ij}, \tag{1}$$

$$s_j = \sqrt{\frac{1}{N-1}\sum\nolimits_{i=1}^{N}(\overline{\mu}_j - r_{ij})^2}, \tag{2}$$

where N denotes the number of observers who participated in the experiment. r_{ij} is the rating score of the j-th image by the i-th observer. The confidence interval for the j-th image is defined as $[\overline{\mu}_j - \sqrt{20}s_j, \overline{\mu}_j + \sqrt{20}s_j]$. If a rating score for an image falls outside the confidence interval, it is considered an outlier. Let P_i and Q_i denote the total number of outliers for the i-th observer, respectively. An observer is removed if more than 5% of his/her rating scores are outliers and the ratio $\left|\frac{(P_i-Q_i)}{(P_i+Q_i)}\right|$ is less than 30%. As a result, none of the observers was rejected among 20 observers. Next, the qualified rating score r_{mj} of each qualified observer needs to be converted to a Z-score $Z_{m,j}$ by :

$$Z_{m,j} = \frac{r_{mj} - \overline{\mu}_m}{s_m}, \tag{3}$$

where $\overline{\mu}_m$ and s_m are the mean score and standard deviation across all qualified rating scores of the m-th observer.

After that, a linear mapping function is used to scale the Z-Score to the rating range, i.e., [1, 5] for the overall quality and [1, 3] for all attributes (eyes, nose, mouth, and facial shape). The mean opinion score (MOS) Q_j for the j-th image is finally calculated as the average value of scaled Z-score $Z'_{m,j}$:

$$Q_j = \frac{1}{M}\sum\nolimits_{m=1}^{M} Z'_{m,j} \tag{4}$$

where M is the number of qualified observers. Through the above process, we have five MOS values for each RF image in total.

D. Subjective Data Analysis

Fig. 3 shows the distributions of MOS values. As can be seen, the distributions of MOS values rated from different attributes are different. Such an observation indicates that the attention paid on these attributes should be different when retouching face images as well as designing objective IQA methods. In view of this, we further investigate the order of importance of these attributes by calculating the distance between the MOS distributions using the Kullback-Leibler (KL) divergence. If the two distributions are closer, the KL divergence will be smaller, otherwise, the KL divergence will be larger. Experimental results show that, the KL divergence is 0.0125, 0.0109, 0.0106, and 0.0049 between the MOS values distribution of the overall quality and that of eyes, nose, mouth, and facial shape, respectively. This demonstrates that, among these four attributes, facial shape is the most expressive and attractive factor affecting subjective rating, followed by the mouth, nose, and eyes.

IV. Proposed Method

A. Motivation

In practice, retouching tools not only produce glamorous and visually striking appearance for publishers, but also bring great difficulty to audience to determine how much a retouched image has strayed from reality. Excessive use of retouching techniques can cause false information, which may deceive the audience, such as online dating scams. This leads to a surge in the requirement of effective quality assessment methods for RF images. Considering that geometric retouching functions, e.g., eye enlargement, face lifting, etc., are more easily cause appearance change than photometric retouching functions, e.g., smoothing and whitening, this study mainly evaluates the perceptual quality of RF images from the perspective of geometric alterations. For this purpose, a simple yet effective multi-task learning based NR-IQA method is introduced.

The motivations of our method are that: 1) Since audience cannot access the pristine counterparts of RF images in most short-video platforms, e.g., Facebook Reels, YouTube Shorts, and TikTok, a NR IQA method is much desired to reflect how humans make predictions without reference to the original faces. 2) As described previously in Section III-A, the geometric alterations of an RF image may be the result of retouching operations on facial features, such as eyes, mouth, nose, etc. Therefore, analyzing the alteration degree of each attribute can help us better understand the perceptual quality of an RF image. 3) According to the principle of multi-task learning, the network can learn more appropriate feature representations to obtain better performance for the main task, thanks to the assistance of well-designed auxiliary tasks.

B. Network Architecture

Our MTNet works in a multi-task manner, where the main and auxiliary tasks are the prediction of the overall quality score and alteration degree estimations of four facial features (i.e., eyes, nose, mouth, and facial shape), respectively. Fig. 4 presents the architecture of MTNet, which consists of image feature extraction and image quality prediction.

1) Image Feature Extraction: For an RF image with the size of $H \times W \times 3$, we feed it into the backbone and extract a set of features ($F_i \in \mathbb{R}^{H/2^i \times W/2^i \times C_i}$, $i \in \{1, 2, 3, 4\}$) from the backbone. In this study, considering that the RF images usually include both global and local semantic distortions, we select the popular ResNet50 [56] as the backbone for feature extraction by reserving five residual blocks and removing fully connected layers, instead of a shallow network that takes small image patches as the input and only consists of very few convolution layers. This is because taking small image patches is not conducive to the network capturing global distortions, and the shallow network has limited ability in extracting semantic concepts. It is worth noting that, other mainstream CNN-based and Transformer-based networks that take relatively large image patches or the whole image as the input can also be applied. We leave the optimal backbone selection as a possible future work as it is not the focus of this study.

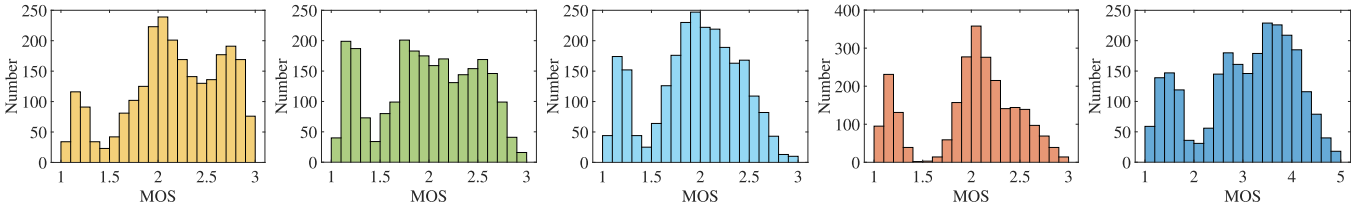

Fig. 3. Distribution of MOS values in the MARFD. Each sub-figure is labeled with the name of the corresponding retouching attribute for easy comparison.

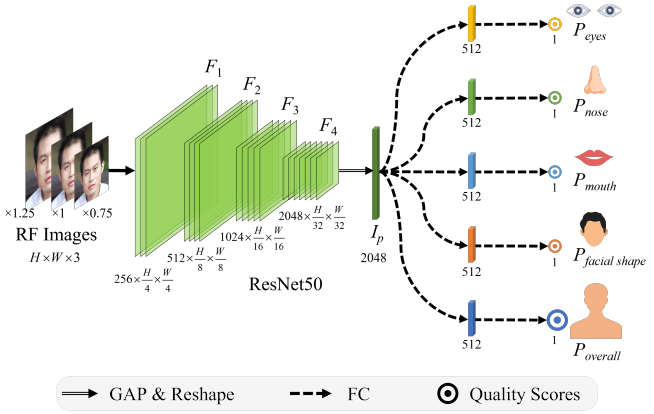

Fig. 4. Architecture of MTNet. Given an RF image, it is sent to the feature extractor first to obtain the semantic feature F_4. Then, F_4 is processed by the GAP and reshape operations. Finally, the resulting I_p is fed into the quality regressor to estimate the alteration degree P_i (i∈{eyes, nose, mouth, facial shape}) of four facial features (eyes, nose, mouth, facial shape) and the overall quality score of the RF image simultaneously.

2) Image Quality Prediction: Since F_4 contains the semantic concepts of the input RF image, we process it with a global average pooling (GAP) operation and reshape to compress it in the spatial domain. After that, the resultant feature I_p is sent to the quality regressor to predict image quality scores from different perspectives. Specifically, the quality regressor follows a five-branch architecture, in which the top four branches predict the alteration degree P_i (i∈{eyes, nose, mouth, facial shape}) of four facial features, respectively, and the bottom branch estimates the perceptual quality score $P_{overall}$ (P_O) of the RF image. All five branches have the same structure, including two fully connected layers with the (input, output) neural nodes of (2048, 512) and (512, 1), respectively. After the first fully connected layer, we add a Hardswish activation function. By setting these four auxiliary tasks, the network is able to better extract discriminative features from different attributes and achieve better prediction consistency with the subjective ratings. In addition, it can help us better understand which attributes have been modified and to what extent they have been modified.

C. Loss Function

To optimize the network, we adopt a multi-task learning strategy. The overall loss function for MTNet is defined as:

$$\mathcal{L} = \mathcal{L}_q(G_O, P_O) + \sum_{i=1}^{4} \mathcal{L}_q(G_i, P_i), \tag{5}$$

where G_O is the MOS value of the overall quality, and G_i is the MOS value of the i-the attribute ($i \in$ {eyes, nose, mouth, facial shape}) of the input RF image. In Eq. (5), the first item is used for supervising the main task that predicts the overall quality score of an RF image, while the second item is used for supervising the auxiliary tasks that separately estimate the alteration degree of all attributes. In this study, following previous works, we choose the widely used mean square error loss as $\mathcal{L}_q(\cdot,\cdot)$.

D. Implementation Details

In the training stage, the backbone uses the pre-trained weights by ImageNet and the other layers of our MTNet are randomly initialized. The network is learnt by minimizing Eq. (5), with a batch size of 32. A total of 60 training epochs are carried out, utilizing the Adam optimizer. The initial learning rate is set at 1e-4, decaying by 0.5 after every 10 epochs. All the input RF images are resized to 512×512, and the randomly horizontal flip is used to augment training images. To fully capture the distortions in RF images, we scale the images in each batch by ratios of 1, 0.75, and 1.25, and train each batch through three separate iterations. In the inference stage, we resize the testing RF images to 512×512 and pass them to the well-trained model to get the predicted scores. Our NR-IQA network is implemented on the PyTorch framework, and all the experiments are conducted on a workstation equipped with two Intel XEON 4210R CPUs and one NVIDIA RTX 3090 GPU.

V. Experiments and Results

A. Experimental Setup

1) Train-Test Split of the Database: Following the general practice in the IQA field, we randomly divide the constructed MARFD into two non-overlapping subsets based on the contents of the raw face images. Specifically, 2,000 RF images from 200 out of 250 raw face images are selected as the training subset, and the 500 RF images from the remaining 50 raw face images are chosen as the testing subset. For fair comparisons, all IQA models are trained (tested) on the same training (testing) subset. Also, to eliminate the performance bias from a specific train-test split, we execute this random train-test split procedure 10 times and report the mean value of the results across 10 trials.

2) Evaluation Metrics: We choose four mainstream evaluation metrics in the IQA field [57] to report the quantitative results of IQA methods, including SRCC, KRCC, PLCC, and RMSE. Specifically, the first two metrics measure the prediction monotonicity, while the last two metrics evaluate the prediction accuracy. As suggested by the video quality experts group [58], a nonlinear fitting function is used to map objectively predicted scores to subjective rating scores before computing PLCC and RMSE:

$$f(s) = \kappa_2 + \frac{\kappa_1 - \kappa_2}{1 + e^{\kappa_4(s-\kappa_3)}}, \tag{6}$$

where s are the predictive scores computed by an objective IQA method, and $f(s)$ are the accordingly fitted scores. $\kappa_1 \sim \kappa_4$ are fitting parameters that can be estimated by minimizing the sum of squared errors between $f(s)$ and subjective ratings. Generally, a superior IQA method provides predictive scores that are more consistent with subjective ratings, achieving higher SRCC, KRCC, and PLCC values yet a lower RMSE value.

3) Compared Methods: In this study, we compare our introduced MTNet with 16 state-of-the-art NR-IQA methods, including BRISQUE [17], GM-LOG [59], GWH-GLBP [60], FRIQUEE [61], BIQME [62], BMPRI [63], MDM [64], CNNIQA [27], MB-CNN [28], HyperIQA [29], GraphIQA [30], VIPNet [37], StairIQA [65], MUSIQ [34], TReS [36], and DEIQT [35]. The first seven methods are traditional handcrafted feature-based methods, while the others are DNN-based methods. Among these DNN-based competitors, CNNIQA and MB-CNN divide the image into small patches (with the size of 32×32) and take them as the input, while the others take the cropped large patches (usually with the size of 224×224 or larger) as the input. In addition, the first five DNN-based methods utilize CNN for distortion-aware feature extraction and fusion, the last two DNN-based methods use pure Transformer networks, while the middle two methods consider both CNN and Transformer for image quality prediction. All these methods are retrained and tested on the divided data in Section V-A1 using the official source codes released by authors with their default settings. For methods (VIPNet and DEIQT) that rely on combined synthetic distortion databases for training, we directly load the pre-trained modules (i.e., the DPM of VIPNet and the encoder of DEIQT) provided by the authors and retrain them on our constructed dataset. This is because our constructed MARFD does not include the synthetic distortions they needed.

B. Comparison on the Whole Database

Table II shows the results of our MTNet and 16 competing NR-IQA methods on the constructed database in terms of four evaluation metrics. The best result of each evaluation metric is marked in bold. From Table II, we have the following observations. First of all, conventional NR-IQA methods are not qualified for the perceptual assessment of multi-attribute RF images. For instance, even though BRISQUE achieves the best performance among the six methods, it only obtains 0.241, 0.303, 0.164, and 0.940 in SRCC, PLCC, KRCC, and RMSE, respectively. This is because these methods are specifically designed for evaluating synthetically distorted images based on the assumptions that features, e.g., natural scene statistics [17], [61], local structures [59], [60], or information amount [64], are measurably modified by the presence of distortions, so they, perhaps not surprisingly, are inclined to mediocrity when coping with the complex geometric distortions of RF images. Secondly, DNN-based methods possess the superior capability for accurately evaluating RF images and generally perform better than conventional methods. A possible reason for this is that DNN can automatically extract and fuse distortion-aware features. Last but not the least, our MTNet performs better than competing methods. For instance, our MTNet has the increment of 0.048 in SRCC, 0.043 in PLCC, and 0.070 in KRCC, respectively, over the popular method HyperIQA. Additionally, compared with the recently reported methods VIPNet and DEIQT, our MTNet still has clear leading advantages in accurately evaluating the quality of RF images. For instance, it surpasses VIPNet and DEIQT by approximately 14.8%, 4.1% in SRCC, 9.9%, 3.5% in PLCC, and 18.1%, 6.1% in KRCC, respectively. As shown in Fig. 5, compared to 9 DNN-based competing methods, the proposed MTNet can produce scatter points that are closer to the fitted curve. This indicates that MTNet correlates better with human visual perception in evaluating the quality of RF images.

TABLE II
PERFORMANCE COMPARISONS ON THE CONSTRUCTED DATABASE. ↑ (↓) INDICATES THAT THE HIGHER (LOWER) VALUE IS, THE BETTER PERFORMANCE IS.

	Methods	SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
Conventional NR-IQA	BRISQUE [17]	0.241	0.303	0.164	0.940
	GM-LOG [59]	0.214	0.271	0.145	0.949
	GWH-GLBP [60]	0.213	0.246	0.145	0.956
	FRIQUEE [61]	0.214	0.225	0.149	0.960
	BIQME [62]	0.115	0.133	0.082	0.977
	MDM [64]	0.073	0.072	0.058	0.984
	BMPRI [63]	0.207	0.270	0.140	0.476
DL-based NR-IQA	CNNIQA [27]	0.266	0.295	0.182	0.940
	MB-CNN [28]	0.265	0.298	0.180	0.941
	HyperIQA [29]	0.880	0.903	0.699	0.423
	GraphIQA [30]	0.466	0.487	0.323	0.816
	VIPNet [37]	0.780	0.847	0.588	0.524
	TReS [36]	0.833	0.862	0.643	0.500
	StairIQA [65]	0.471	0.486	0.327	0.861
	MUSIQ [34]	0.839	0.866	0.651	0.491
	DEIQT [35]	0.887	0.911	0.708	**0.179**
	MTNet (Ours)	**0.928**	**0.946**	**0.769**	0.318

C. Comparison on Each Photo-Editing Setting

Four popular photo-editing tools are selected and operated under different retouching settings to generate RF images, as shown in Table I. It is meaningful to compare the performance of NR-IQA methods on images generated under each setting. Table III tabulates the experimental results, in which only SRCC and PLCC are given due to space limitations. As can be seen, the results of each NR-IQA method vary greatly across different retouching settings. For example, among these NR-IQA methods, the achieved best SRCC values are 0.463 and 0.768 in Setting 2 and Setting 8, respectively, across all 10 settings. Most competing methods produce unsatisfactory performance (with SRCC lower than 0.15 and PLCC lower

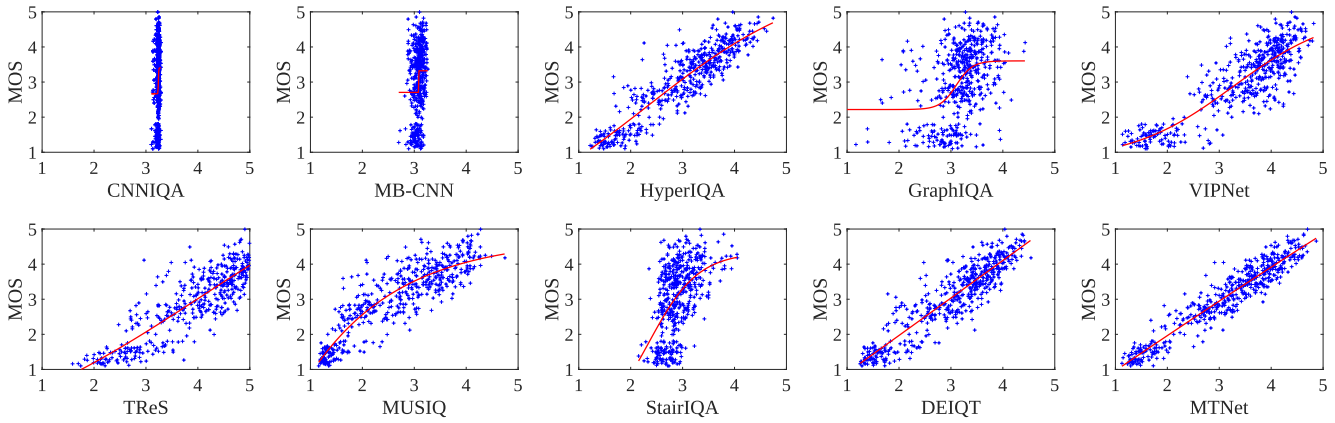

Fig. 5. Scatter plots of the performance of 9 DNN-based NR-IQA methods and our proposed MTNet on the constructed database.

than 0.2) in most settings. TReS performs best in Setting 9, obtaining the SRCC and PLCC values of 0.702 and 0.722. Despite this, our MTNet exhibits stable leading advantages and performs the best in 9 out of 10 settings. More specifically, it achieves the best 18 times and the second best 2 times in a total of 20 comparisons. In contrast, the second-best method TReS only performs the best 2 times and the second best 8 times. This indicates that our MTNet is a more suitable for assisting the audience in estimating the alteration degree of an RF image than competing NR-IQA methods. According to the descriptions in Table I, four photo-editing tools are used for generating RF images. Here, we also report the results on each photo-editing tool. As shown in Table IV, our proposed MTNet also shows superior performance than competing methods.

TABLE IV
QUANTITATIVE RESULTS ON EACH INDIVIDUAL PHOTO-EDITING TOOL. ↑ INDICATES THAT THE HIGHER VALUE IS, THE BETTER PERFORMANCE IS. WE MARK THE BEST RESULT OF EACH EVALUATION METRIC IN **BOLDFACE** FOR CONVENIENT COMPARISONS.

Methods	Luminar Neo		Arcsoft Portrait+3		MeiTu		Adobe Photoshop	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BRISQUE [17]	0.295	0.289	0.100	0.123	0.113	0.128	0.070	0.062
GM-LOG [59]	0.239	0.205	0.106	0.087	0.139	0.134	0.064	0.060
GWH-GLBP [60]	0.222	0.202	0.048	0.070	0.145	0.145	0.091	0.084
FRIQUEE [61]	0.213	0.227	0.067	0.094	0.109	0.126	0.053	0.064
BIQME [62]	0.156	0.154	0.126	0.118	0.126	0.106	0.097	0.075
MDM [64]	0.071	0.075	0.088	0.097	0.050	0.055	0.079	0.078
BMPRI [63]	0.184	0.188	0.074	0.077	0.108	0.100	0.059	0.060
CNNIQA [27]	0.315	0.342	0.108	0.260	0.212	0.255	0.081	0.167
MB-CNN [28]	0.303	0.334	0.151	0.258	0.220	0.244	0.073	0.125
HyperIQA [29]	0.796	0.799	0.153	0.185	0.834	0.823	0.672	0.684
GraphIQA [30]	0.359	0.390	0.086	0.237	0.362	0.379	0.129	0.170
VIPNet [37]	0.796	0.841	0.112	0.224	0.733	0.732	0.283	0.307
TReS [36]	0.792	0.788	0.275	0.358	0.797	0.799	0.557	0.588
StairIQA [65]	0.448	0.450	0.112	0.296	0.471	0.479	0.247	0.274
MUSIQ [34]	0.474	0.516	0.155	0.303	0.404	0.458	0.291	0.371
DEIQT [35]	0.819	0.826	0.318	0.377	0.841	0.845	0.680	0.705
MTNet (Ours)	**0.882**	**0.906**	**0.463**	**0.530**	**0.898**	**0.904**	**0.798**	**0.822**

D. Computational Complexity

In this section, we carry out more experiments to compare our method with DNN-based methods in terms of computational complexity. The floating-point operations per second (FLOPs) and the total number of parameters (Params) are selected as evaluation metrics. Table V presents the comparison results, in which we can find that our MTNet does not show significant advantages in terms of these two metrics and ranks in the middle position among competing DNN-based methods. Despite this, our proposed MTNet achieves better consistency with subjective rating scores than competing methods, as shown in Table II, Table III, and Fig. 5. It is worth noting that our current focus is to propose a high-accuracy IQA model for RF images. In the future, we plan to improve our model to further enhance its efficiency, potentially reducing its computational load without compromising performance. This could involve exploring more advanced network pruning techniques or more efficient architectures.

TABLE V
THE COMPLEXITY OF THE DNN-BASED NR-IQA ALGORITHMS.

Methods	FLOPs	Params
CNNIQA [27]	**2.377M**	**724.901K**
MB-CNN [28]	168.810G	66.201M
HyperIQA [29]	4.335G	27.375M
GraphIQA [30]	4.154G	45.752M
VIPNet [37]	19.105G	47.191M
TReS [36]	8.387G	34.457M
StairIQA [65]	5.110G	31.799M
MUSIQ [34]	126.521G	125.563M
DEIQT [35]	4.256G	22.774M
MTNet (Ours)	21.591G	28.756M

E. Ablation Studies

In our introduced MTNet, we propose two feasible strategies for accurate image quality prediction, i.e., using the multi-task learning (MTL) and applying the multi-scale training (MST). The MTL is used to learn more discriminative feature representations through four auxiliary tasks, and the MTL aims to help the network better understand the distortions inspired by the perceptual effects of viewing distance on quality assessment. Here, we further conduct several ablation studies to investigate the contribution of MTL and MST. All experiments are conducted with the same experimental settings as the main experiment, as introduced in Section V-A1.

1) Effectiveness of MTL: To validate the effectiveness of MTL, we learn a variant IQA model by abandoning the MTL strategy from the standard MTNet. Specifically, we directly remove the top four branches and only retain the last branch

TABLE III
QUANTITATIVE RESULTS ON EACH PHOTO-EDITING SETTING. ↑ INDICATES THAT THE HIGHER VALUE IS, THE BETTER PERFORMANCE IS. WE MARK THE BEST RESULT OF EACH EVALUATION METRIC IN **BOLDFACE** FOR CONVENIENT COMPARISONS.

Methods	Setting 1		Setting 2		Setting 3		Setting 4		Setting 5		Setting 6		Setting 7		Setting 8		Setting 9		Setting 10	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BRISQUE [17]	0.143	0.106	0.123	0.100	0.111	0.122	0.113	0.121	0.102	0.101	0.167	0.147	0.080	0.111	0.122	0.144	0.123	0.128	0.075	0.074
GM-LOG [59]	0.084	0.072	0.087	0.106	0.067	0.087	0.137	0.153	0.108	0.129	0.153	0.127	0.130	0.136	0.105	0.073	0.106	0.082	0.137	0.128
GWH-GLBP [60]	0.117	0.087	0.070	0.048	0.145	0.149	0.107	0.143	0.099	0.150	0.107	0.133	0.082	0.126	0.113	0.101	0.115	0.112	0.105	0.125
FRIQUEE [61]	0.090	0.085	0.094	0.067	0.100	0.078	0.117	0.098	0.101	0.091	0.157	0.164	0.099	0.059	0.090	0.078	0.076	0.089	0.083	0.077
BIQME [62]	0.117	0.131	0.118	0.126	0.146	0.100	0.125	0.149	0.157	0.159	0.158	0.161	0.114	0.101	0.134	0.170	0.164	0.205	0.109	0.148
MDM [64]	0.129	0.141	0.097	0.088	0.139	0.145	0.113	0.108	0.128	0.141	0.144	0.150	0.155	0.156	0.083	0.094	0.149	0.133	0.161	0.151
BMPRI [63]	0.155	0.115	0.077	0.074	0.140	0.117	0.121	0.123	0.154	0.173	0.102	0.116	0.157	0.167	0.143	0.114	0.147	0.117	0.138	0.139
CNNIQA [27]	0.107	0.224	0.108	0.260	0.108	0.204	0.101	0.209	0.096	0.236	0.170	0.238	0.126	0.263	0.150	0.226	0.178	0.283	0.114	0.253
MB-CNN [28]	0.145	0.249	0.150	0.256	0.132	0.255	0.061	0.149	0.099	0.196	0.157	0.236	0.078	0.187	0.223	0.260	0.223	0.269	0.100	0.238
HyperIQA [29]	0.470	0.464	0.166	0.217	0.509	0.492	0.343	0.329	0.433	0.428	0.634	0.630	0.505	0.487	0.700	0.701	0.668	0.676	0.498	0.465
GraphIQA [30]	0.081	0.230	0.096	0.258	0.120	0.260	0.105	0.251	0.085	0.243	0.167	0.279	0.117	0.247	0.232	0.322	0.228	0.297	0.155	0.275
VIPNet [37]	0.241	0.344	0.098	0.241	0.241	0.341	0.139	0.200	0.187	0.302	0.403	0.446	0.234	0.300	0.398	0.459	0.467	0.498	0.188	0.311
TReS [36]	0.543	0.582	0.295	0.357	0.553	0.584	0.349	0.368	0.316	0.376	0.718	0.732	0.531	0.535	0.710	0.736	**0.702**	**0.722**	0.440	0.468
StairIQA [65]	0.315	0.343	0.112	0.296	0.291	0.341	0.172	0.281	0.147	0.285	0.499	0.524	0.273	0.387	0.493	0.519	0.443	0.484	0.177	0.286
MUSIQ [34]	0.474	0.516	0.155	0.303	0.404	0.458	0.291	0.371	0.400	0.497	0.604	0.647	0.442	0.528	0.579	0.622	0.563	0.599	0.381	0.472
DEIQT [35]	0.525	0.539	0.275	0.310	0.550	0.563	0.377	0.404	0.406	0.438	0.664	0.683	0.532	0.516	0.723	0.743	0.682	0.704	0.482	0.479
MTNet (Ours)	**0.561**	**0.599**	**0.463**	**0.530**	**0.633**	**0.650**	**0.533**	**0.635**	**0.534**	**0.579**	**0.729**	**0.753**	**0.621**	**0.651**	**0.762**	**0.784**	0.683	0.718	**0.548**	**0.597**

of the regressor in MTNet. As shown in Table VI, there are obvious performance decrements if we abandon the MTL strategy. For example, compared with the standard MTNet, the network without MTL shows a decrement of 2.2%, 2.0%, and 3.3% in terms of SRCC, PLCC, and KRCC, respectively. This indicates that the used MTL strategy plays a positive role in obtaining good performance. In addition, we further investigate the performance of our method in evaluating the quality score of each attribute. As shown in Fig. 6, our MTNet is able to understand the alteration degree of each attribute well and achieves good predictions, with PLCC, SRCC, and KRCC greater than 0.92, 0.87, and 0.68, respectively. More specifically, our proposed MTNet has better predictions, with higher SRCC, PLCC, KRCC values, when evaluating the alteration degree from nose than other three attributes, i.e., mouth, eyes, and facial shape. The performance difference motivates us to add different weights for different tasks in Eq. (5) for further overall performance improvement in the future. How to select the optimal weights is our future direction.

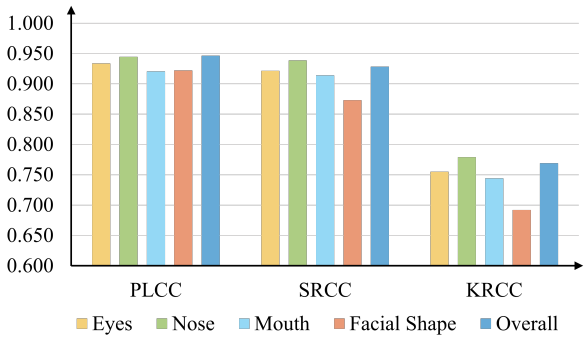

Fig. 6. Results of quality prediction of each attribute.

2) Effectiveness of MST: In this study, to help the network better understand the distortions, we scale each image into three ratios, i.e., 1.25, 1, and 0.75, and take them as the inputs of the network. To validate the effectiveness of such an operation, we learn a variant IQA model by only retaining the scaling ratio of 1. Through the comparisons between results in the last two rows of Table VI, we can see that MTNet has a decrement of 0.9%, 1.0%, and 1.6% in terms of SRCC, PLCC, and KRCC, respectively, if we do not use the MST strategy. Overall, both MTL and MST strategies play a positive role in accurately evaluating the quality of multi-attribute RF images.

TABLE VI
ABLATION ANALYSIS OF MULTI-TASK LEARNING (MTL) AND MULTI-SCALE TRAINING (MST). THE SYMBOL "×" ("✓") DENOTES THE COMPONENT IN THE COLUMN IS EXCLUDED (INCLUDED) IN THE STANDARD MTNET.

Baseline	MTL	MST	SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
✓	×	×	0.882	0.900	0.702	0.429
✓	×	✓	0.906	0.926	0.736	0.372
✓	✓	×	0.919	0.936	0.753	0.347
✓	✓	✓	**0.928**	**0.946**	**0.769**	**0.318**

3) Impact of Backbone: We further investigate the impact of the selected backbone. Table VII presents the performance of various network architectures used as the backbone. Here, "Single Task" denotes using the proposed MTNet without employing MTL and MST strategies, and "Multi-Task" denotes using the standard MTNet. As seen, the backbone, related to the feature representation capability, directly affects the prediction performance. Among the selected six backbones, i.e., ResNet18 [56], ResNet50 [56], ResNet101 [56], AlexNet [66], MobileNetV2 [67], and DenseNet121 [68], ResNet50 performs the best in case of "Single Task" and the second best in case of "Multi-Task", respectively. Moreover, ResNet50 also achieves the best results when averaging the results of two cases. Therefore, we choose ResNet50 as the backbone of our MTNet. It is worth noting that other popular network architectures can also be selected as the backbone of our method and the focus of current study is to design an effective quality assessment model for RF images. We leave the optimal backbone selection as a future work.

VI. DISCUSSIONS

Nowadays, owing to the usage of photo retouching tools, advertisements and short-video platforms routinely depict impossibly beautiful facial appearance to attract audience attention. The reasonable usage of such tools, to some extent, can increase audiences' shopping enthusiasm and online

TABLE VII
ABLATION ANALYSIS OF BACKBONE.

Methods	Single Task		Multi-Task		Average	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
ResNet18 [56]	0.849	0.872	0.924	0.940	0.887	0.906
ResNet50 [56]	**0.882**	**0.900**	0.928	0.946	**0.905**	**0.923**
ResNet101 [56]	0.876	0.895	**0.930**	0.945	0.903	0.920
AlexNet [66]	0.798	0.824	0.847	0.870	0.823	0.847
MobilenetV2 [67]	0.785	0.814	0.901	0.921	0.843	0.868
Densenet121 [68]	0.856	0.879	**0.930**	**0.947**	0.893	0.913

activity, improving life quality. However, as a negative effect, excessively using photo retouching techniques not only brings tough challenges to face recognition for important scenarios, e.g., online identity authentication, but also provides incorrect inducement information that may cause audiences' mental anxiety about body control and that may lead to online dating scams. In this context, reliable evaluation algorithms for RF images are highly required. An algorithm that simply labels an image as digitally altered or not would have limited efficacy because it cannot explicitly tell us how much an RF image has strayed from reality [13]. In contrast, the algorithm that can provide a continuous rating score for the RF image is more desired to protect the interests of audiences. However, this topic has been largely overlooked in the past.

In this study, we advance this topic from both subjective and objective studies. Specifically, considering that there are very few IQA databases in this field, a multi-attribute retouched face image database, named MARFD, is first constructed through strict subjective experiments. MARFD consists of 2,500 RF images and associated continuous quality scores. The RF images are the results of 4 popular photo-editing tools on 250 face images under different settings. Unlike existing databases that only include the raw and retouched image pairs for binary retouching detection [15] or consider the whole body retouching images from the online sources [3], MARFD focuses on RF images generated by several popular photo-editing tools and provides continuous rating scores for each RF image from different perspectives. With these scores, it helps us better understand which attributes have been modified and to what extent they have been modified. We evaluate 16 state-of-the-art NR-IQA methods to investigate their effectiveness on multi-attribute RF data and their responses to the generated images of different photo-editing tools. The experimental results arouse our thinking as follows.

DNN-based methods are the better choice for RF image quality estimation than handcrafted feature-based methods. The reasons are from two aspects. On the one hand, geometric distortions from retouching operations are quite complex, and it is very challenging to mine effective and robust handcrafted features to measure such distortions. On the other hand, DNN-based methods can automatically extract and integrate distortion-aware features, thanks to the data-driven property of deep learning (DL). For better performance, efforts paid in feature representation, e.g., multi-task learning [28], content understanding [29], backbone modification [36], multi-scale inputs [34], etc., are the key points of designing such methods. Although very popular, their acceptance in the practical context is often limited by the fact that data-driven methods lack interpretability. Considering the subjective rating scores from multi-attributes, we introduce a multi-task learning based method for evaluating the quality of RF images. The arrangement of four auxiliary tasks helps us better understand the distortion information of an RF image and achieves accurate quality prediction. Nevertheless, the interpretability from the network itself is still missing. As an exciting signal, the recent advances in the field of explainable DL have the prospect to remedy the deficiency of current works by allowing the audiences to align more image processing knowledge with indicative features for accurate image quality estimation.

Since the literature lacks large-scale IQA databases, data augmentation is the prior concern of designing effective DNN-based IQA methods. Existing DNN-based methods can be broadly divided into two categories, i.e., cutting-based methods and cropping-based methods, based on the way of generating network inputs. The former divides the image evenly into equally small patches (with the size of 32×32), while the latter randomly crops the image many times to obtain large patches (usually with the size of 224×224 or larger). Although these two categories of methods have been validated effectively in some synthetically distorted NSI databases, their effectiveness on RF data is different. According to the results in Section V-B, cutting-based methods (i.e., CNNIQA and MB-CNN) are generally inferior to cropping-based methods (i.e., HyperIQA, GraphIQA, VIPNet, MUSIQ, StairIQA, TReS, and DEIQT). The reasons for this can be attributed to two aspects. First, dividing images into small patches splits the connectivity between different regions of the image, making the network cannot understand the global semantic concepts. Note that, multi-attribute retouching mainly changes the geometric features of face and brings both local and global semantic distortions. Second, to train the network, the obtained patches are labeled with the same quality score as the whole image. Such an operation brings label noise to the network, which is adverse to effective model generation. Compared to small patches, the large patches have relatively smaller label noise. As a consequence, cropping-based methods have generally better results than cutting-based methods. To avoid such a problem, the proper use of original-sized image is highly recommended.

With the above observations, we introduce a simple yet effective NR-IQA method, named MTNet, for multi-attribute RF images. Our MTNet benefits from two strategies, i.e., the multi-task learning strategy and the multi-scale training strategy. Different from existing works that take the distortion recognition as the auxiliary task [31], our MTNet takes the alteration degree estimation of four facial attributes as auxiliary tasks. The use of such auxiliary tasks in our method has two important roles. On the one hand, as shown by the results of ablation studies in Table VI, it helps improve the performance of the main task, i.e., overall quality prediction, through multi-task learning. On the other hand, it helps users better understand which attributes have been modified and to what extent they have been modified. In addition, contrary to existing works that divide or crop images for data augmentation, we utilize the multi-scale training strategy to fully capture the distortion characteristics of RF images. Results of ablation studies in Table VI demonstrate the effectiveness of

these two strategies. We also investigate the performance of these considered methods on evaluating RF images generated by each photo-editing tool. Experimental results show that no NR-IQA method can always achieve the best results in handling the image generated by all retouching settings. This indicates that, for always reliably evaluating the quality of RF images, we need to implement a set of algorithms. Moreover, as shown in Table III, our MTNet achieves better performance in evaluating RF generated under most settings than competing methods. This demonstrates that it can be a more ideal choice than others if only one algorithm can be deployed owing to the limited source.

Different from existing face retouch detection methods that only detect whether the retouching effect exists or not [69], our method, i.e., MTNet, can provide a score of how much an RF image has strayed from reality. This merit makes it have broader application prospects. Here, we further investigate its effectiveness in the binary-classification face retouch detection task and compare it with existing face retouch detection methods. For this purpose, we conduct some experiments on our constructed MARFD dataset and one public dataset (FaceForensics++ [70]). Since MARFD is originally constructed for the regression task, we make some modifications on it for the binary-classification task. In real-world applications, e.g., face image sharing in social networks, slight retouching is preferred to increase personal attractiveness, while excessive retouching should be avoided to reduce difficulties for authentication. In view of this, we divide the RF images into two groups. Specifically, we set a threshold to 3 and mark the RF image as real (if its subjective rating score is below the threshold) or as fake (if its subjective rating score is over or equal to the threshold). For the label conversion of other four attributes, a similar approach is taken using a threshold of 2. The reason why we set the thresholds at 3 and 2 is that the rating range of the overall quality is [1, 5] and the rating range of each attribute is [1, 3]. Taking the median as the threshold can help us fairly partition the data into slight retouching groups and excessive retouching groups. After that, the MARFD dataset is randomly split in an 80/20 ratio for training/testing without content duplication. To conduct the detection experiment task on MARFD, we transform the original five regression heads of MTNet into five classification heads and replace the mean square error loss with the Cross Entropy loss. FaceForensics++ contains 5,000 videos, including 1,000 original and 4,000 fake videos created using four deepfake technologies. To reduce time costs, we randomly extract 36 or 9 frames from each real/fake video for analysis. The dataset is also randomly divided into training and testing subsets according to the video ID, and the split ratio is 4:1. The real instances are labelled to 1 while the others are labelled to 0. To carry out the detection experiment on FaceForensics++, we slightly modify our MTNet by removing the prediction branches for the four attributes and only preserving the prediction branch for the overall quality. In addition, we replace the regression head with the classification head and use the Cross Entropy loss during network training. Four face retouch detection methods are selected for comparisons, including Scattering ResNet [71], XceptionNet [72], Two-Stream Net [73], and CADDM [74]. All these methods are implemented using their default settings. Table VIII tabulates the results. As seen, our proposed MTNet is also competent for the binary-classification face retouch detection task, with superior performance over four competing methods.

TABLE VIII
ACCURACY COMPARISON ON FAKEFACE DETECTION TASKS

Methods	MARFD	FaceForensics++ [70]
Scattering ResNet [71]	0.568	0.681
XceptionNet [72]	0.890	0.800
Two-Stream Net [73]	0.886	0.762
CADDM [74]	0.914	0.745
MTNet (Ours)	**0.918**	**0.885**

Our study dovetails with current discourse in digital media, noticing the important effects of facial retouching within public and private spheres, from social media to professional settings. An effective RF IQA method plays a crucial role in various sectors including personal use and entertainment, legal and government document processing, as well as in the social media and advertising industries. In the personal realm, it aids users in selecting more authentic photos for life documentation; in legal and government documentation, it ensures the authenticity of official documents; and in the realm of social media and advertising, it contributes to maintaining the genuineness of content and prevents the misleading of consumers.

VII. CONCLUSION

This paper conducts an in-depth research on perceptual quality assessment of RF images. Firstly, considering that there are very few IQA databases in this field, we generate 2,500 multi-attribute RF images using 4 photo-editing tools under different settings and construct a new IQA database by conducting subjective studies. The constructed database MARFD provides a reliable platform to validate and compare the effectiveness of objective IQA methods. Secondly, we introduce a simple yet effective multi-task learning based NR-IQA method, named MTNet. A multi-scale training strategy is applied to help the network better understand the retouching distortions. Extensive experiments on MARFD show that our MTNet is qualified for the RF IQA task, with superior performance than 16 state-of-the-art NR-IQA methods. Results of ablation studies also demonstrate the effectiveness of the multi-task learning and multi-scale training strategies. Nevertheless, there is still much room for improvement, especially on tests of each photo-editing setting. The release of the newly constructed database and benchmarking results are expected to pave the way for proposing specific IQA methods for multi-attribute RF images, and the design concept of MTNet can provide a reference for the follow-up research.

REFERENCES

[1] P. Majumdar, A. Agarwal, M. Vatsa, and R. Singh, "Facial retouching and alteration detection," in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer International Publishing Cham, 2022, pp. 367–387.

[2] H. Dittmar, "How do body perfect ideals in the media have a negative impact on body image and behaviors? factors and processes related to self and identity," *Journal of Social and Clinical Psychology*, vol. 28, no. 1, pp. 1–8, 2009.

[3] E. Kee and H. Farid, "A perceptual metric for photo retouching," *Proceedings of the National Academy of Sciences*, vol. 108, no. 50, pp. 19 907–19 912, 2011.

[4] S. Lang, X. Liu, M. Zhou, J. Luo, H. Pu, X. Zhuang, J. Wang, X. Wei, T. Zhang, Y. Feng *et al.*, "A full-reference image quality assessment method via deep meta-learning and conformer," *IEEE Transactions on Broadcasting*, accepted, in press, DOI: 10.1109/TBC.2023.3308349, 2023.

[5] C. Jin, Z. Peng, F. Chen, and G. Jiang, "Subjective and objective video quality assessment for windowed-6dof synthesized videos," *IEEE Transactions on Broadcasting*, vol. 68, no. 3, pp. 594–608, 2022.

[6] X. Lan, M. Zhou, X. Xu, X. Wei, X. Liao, H. Pu, J. Luo, T. Xiang, B. Fang, and Z. Shang, "Multilevel feature fusion for end-to-end blind image quality assessment," *IEEE Transactions on Broadcasting*, vol. 69, no. 3, pp. 801–811, 2023.

[7] M. Zhou, L. Chen, X. Wei, X. Liao, Q. Mao, H. Wang, H. Pu, J. Luo, T. Xiang, and B. Fang, "Perception-oriented u-shaped transformer network for 360-degree no-reference image quality assessment," *IEEE Transactions on Broadcasting*, vol. 69, no. 2, pp. 396–405, 2023.

[8] Y. Chang, S. Li, A. Liu, W. Zhang, J. Jin, and W. Xiang, "Bidirectional feature aggregation network for stereo image quality assessment considering parallax attention-based binocular fusion," *IEEE Transactions on Broadcasting*, accepted, in press, DOI: 10.1109/TBC.2023.3278096, 2023.

[9] H. Sheikh, "Live image quality assessment database release 2," *http://live. ece. utexas. edu/research/quality*, 2005.

[10] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[11] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4408–4421, 2015.

[12] G. Yue, C. Hou, and K. Gu, "Subjective quality assessment of animation images," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[13] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer, "Detecting facial retouching using supervised deep learning," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1903–1913, 2016.

[14] K. Sharma, G. Singh, and P. Goyal, "Ipdcn2: Improvised patch-based deep cnn for facial retouching detection," *Expert Systems with Applications*, vol. 211, p. 118612, 2023.

[15] C. Rathgeb, C.-I. Satnoianu, N. E. Haryanto, K. Bernardo, and C. Busch, "Differential detection of facial retouching: A multi-biometric approach," *IEEE Access*, vol. 8, pp. 106 373–106 385, 2020.

[16] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems*, vol. 5, no. 4, p. 517, 1994.

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

[18] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[19] G. Yue, C. Hou, K. Gu, and N. Ling, "No reference image blurriness assessment with local binary patterns," *Journal of Visual Communication and Image Representation*, vol. 49, pp. 382–391, 2017.

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

[21] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.

[22] G. Yue, C. Hou, K. Gu, S. Mao, and W. Zhang, "Biologically inspired blind quality assessment of tone-mapped images," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 3, pp. 2525–2536, 2018.

[23] G. Ginesu, F. Massidda, and D. D. Giusto, "A multi-factors approach for image quality assessment based on a human visual system model," *Signal Processing: Image Communication*, vol. 21, no. 4, pp. 316–333, 2006.

[24] X. Wang, S. Kwong, and Y. Zhang, "Considering binocular spatial sensitivity in stereoscopic image quality assessment," in *2011 Visual Communications and Image Processing (VCIP)*. IEEE, 2011, pp. 1–4.

[25] W. Yan, G. Yue, Y. Fang, H. Chen, C. Tang, and G. Jiang, "Perceptual objective quality assessment of stereoscopic stitched images," *Signal Processing*, vol. 172, p. 107541, 2020.

[26] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[27] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[28] Z. Pan, F. Yuan, X. Wang, L. Xu, S. Xiao, and S. Kwong, "No-reference image quality assessment via multi-branch convolutional neural networks," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 1, pp. 148–160, 2023.

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

[30] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "Graphiqa: Learning distortion graph representations for blind image quality assessment," *IEEE Transactions on Multimedia*, accepted, in press, DOI: 10.1109/TMM.2022.3152942, 2022.

[31] C. Zhang, Z. Huang, S. Liu, and J. Xiao, "Dual-channel multi-task cnn for no-reference screen content image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5011–5025, 2022.

[32] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1383–1391.

[33] J. You and J. Korhonen, "Transformer for image quality assessment," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1389–1393.

[34] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.

[35] G. Qin, R. Hu, Y. Liu, X. Zheng, H. Liu, X. Li, and Y. Zhang, "Data-efficient image quality assessment with attention-panel decoder," *arXiv preprint arXiv:2304.04952*, 2023.

[36] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1220–1230.

[37] X. Wang, J. Xiong, and W. Lin, "Visual interaction perceptual network for blind image quality assessment," *IEEE Transactions on Multimedia*, accepted, in press, DOI: 10.1109/TMM.2023.3243683, 2023.

[38] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 520–531, 2015.

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

[40] G. Yue, D. Cheng, T. Zhou, J. Hou, W. Liu, L. Xu, T. Wang, and J. Cheng, "Perceptual quality assessment of enhanced colonoscopy images: A benchmark dataset and an objective method," *IEEE Transactions on Circuits and Systems for Video Technology*, accepted, in press, DOI: 10.1109/TCSVT.2023.3260212, 2023.

[41] Y. Zhang, Y. Wang, F. Liu, Z. Liu, Y. Li, D. Yang, and Z. Chen, "Subjective panoramic video quality assessment database for coding applications," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 461–473, 2018.

[42] Q. Jiang, Z. Liu, K. Gu, F. Shao, X. Zhang, H. Liu, and W. Lin, "Single image super-resolution quality assessment: a real-world dataset, subjective studies, and an objective metric," *IEEE Transactions on Image Processing*, vol. 31, pp. 2279–2294, 2022.

[43] W. Lin, Y. Wu, L. Xu, W. Chen, T. Zhao, and H. Wei, "No-reference quality assessment for low-light image enhancement: Subjective and objective methods," *Displays*, vol. 78, p. 102432, 2023.

[44] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[45] M. Knoche, S. Hormann, and G. Rigoll, "Cross-quality lfw: A database for analyzing cross-resolution image face recognition in unconstrained environments," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–5.

[46] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 51–59.
[47] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer, "Cr-fiqa: face image quality assessment by learning sample relative classifiability," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5836–5845.
[48] F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y.-G. Wang, "Sdd-fiqa: unsupervised face image quality assessment with similarity distribution distance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7670–7679.
[49] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 225–14 234.
[50] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Serfiq: Unsupervised estimation of face image quality based on stochastic embedding robustness." in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5651–5660.
[51] Ž. Babnik, P. Peer, and V. Štruc, "Faceqan: Face image quality assessment through adversarial noise exploration," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 748–754.
[52] K. Lamichhane, M. Neri, F. Battisti, P. Paudyal, and M. Carli, "No-reference light field image quality assessment exploiting saliency," *IEEE Transactions on Broadcasting*, vol. 69, no. 3, pp. 790–800, 2023.
[53] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2014.
[54] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.
[55] R. BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, vol. 4, 2002.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[57] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of visual communication and image representation*, vol. 22, no. 4, pp. 297–312, 2011.
[58] V. Q. E. Group *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment," in *VQEG meeting, Ottawa, Canada, March, 2000*, 2000. [Online]. Available: http://www.vqeg.org/
[59] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.
[60] Q. Li, W. Lin, and Y. Fang, "No-reference quality assessment for multiply-distorted images in gradient domain," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 541–545, 2016.
[61] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32–32, 2017.
[62] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 1301–1313, 2018.
[63] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.
[64] H. Z. Nafchi and M. Cheriet, "Efficient no-reference quality assessment and classification model for contrast distorted images," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 518–523, 2018.
[65] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 6, pp. 1178–1192, 2023.
[66] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.
[67] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
[68] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
[69] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "Deepfake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18 757–18 775, 2022.
[70] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1–11.
[71] E. Oyallon, S. Zagoruyko, G. Huang, N. Komodakis, S. Lacoste-Julien, M. Blaschko, and E. Belilovsky, "Scattering networks for hybrid representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2208–2221, 2019.
[72] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.
[73] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 317–16 326.
[74] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3994–4004.

Guanghui Yue received the B.S. degree in communication engineering from Tianjin University in 2014, and the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2019. He was a joint Ph.D. student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore, from September 2017 to January 2019.

He is currently an Associate Professor with the School of Biomedical Engineering, Health Science Center, Shenzhen University. His research interests include medical image analysis, bioelectrical signal processing, image quality assessment, 3D image visual discomfort prediction, pattern recognition, and machine learning.

Honglv Wu received the B.S. degree in Communication Engineering from Chang'an University, Xi'an, China, in 2021. She is currently pursuing the Master's degree in Electronic and Information Engineering at Shenzhen University, China. Her research interests include image quality assessment, image enhancement, deep learning, and so on.

Weiqing Yan received the PhD. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2017. She was a visiting PhD student at visual spatial perceived lab, University of California, Berkeley, CA, USA from September 2015 to September 2016.

Tianwei Zhou received the B.S. degree in automation from Tianjin University in 2014 and the Ph.D. degree in control science and engineering from Tianjin University, Tianjin, China, in 2019. She was a joint Ph.D. student with the Department of Electrical & Computer Engineering, National University of Singapore from August 2017 to August 2018.

She is currently an Assistant Professor with the College of Management, Shenzhen University. Her current research interests include event-triggered control, intelligent scheduling, image processing, and medical image analysis.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of the IEEE Transactions on Human Machine Systems and the IEEE Transactions on Multimedia.

Wei Zhou is an Assistant Professor at Cardiff University, United Kingdom. Dr Zhou was a Postdoctoral Fellow at University of Waterloo, Canada. Wei received the PhD degree from the University of Science and Technology of China in 2021, joint with the University of Waterloo from 2019 to 2021. Dr Zhou was a visiting scholar at National Institute of Informatics, Japan, a research assistant with Intel, and a research intern at Microsoft and Alibaba. Wei's research interests span multimedia computing, perceptual image processing, and computational vision.