Just Noticeable Difference Modeling for Face Recognition System

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

Abstract—High-quality face images are required to guarantee the stability and reliability of automatic face recognition (FR) systems in surveillance and security scenarios. However, a massive amount of face data is usually compressed before being analyzed due to limitations on transmission or storage. The compressed images may lose the powerful identity information, resulting in the performance degradation of the FR system. Herein, we make the first attempt to study just noticeable difference (JND) for the FR system, which can be defined as the maximum distortion that the FR system cannot notice. More specifically, we establish a JND dataset including 3530 original images and 137,670 compressed images generated by advanced reference encoding/decoding software based on the Versatile Video Coding (VVC) standard (VTM-15.0). Subsequently, we develop a novel JND prediction model to directly infer JND images for the FR system. In particular, in order to maximum redundancy removal without impairment of robust identity information, we apply the encoder with multiple feature extraction and attention-based feature decomposition modules to progressively decompose face features into two uncorrelated components, i.e., identity and residual features, via self-supervised learning. Then, the residual feature is fed into the decoder to generate the residual map. Finally, the predicted JND map is obtained by subtracting the residual map from the original image. Experimental results have demonstrated that the proposed model achieves higher accuracy of JND map prediction compared with the state-of-the-art JND models, and is capable of saving more bits while maintaining the performance of the FR system compared with VTM-15.0.

Index Terms—Just noticeable distortion, face recognition system, deep neural network, image and video coding.

I. INTRODUCTION

RECENTLY, advances in the field of face recognition (FR) have made it the prominent biometric technique for identity authentication in many areas [1], [2], [3], [4], [5], [6], [7], such as public security, finance, and human-computer interaction. Since FR systems refer to computing the similarity of discriminative features between two images, high-quality face images are of paramount importance for achieving the superior performance of FR systems. In general, as illustrated in Fig. 1(a), FR systems deployed in the cloud server perform recognition over two images sampled from the gallery set and probe set, respectively. The gallery set is composed of face data with known subjects stored in the cloud databases, and the probe set is collected from various remote devices such as smartphones, monitors and laptops. In some application scenarios, the probe set is required to be compressed before being transmitted to the cloud server to facilitate this remote computing process. Unfortunately, the FR system might not correctly recognize the compressed images with incomplete identity information from corrupted images [8]. Therefore, it is highly desirable to find the optimal solution for balancing coding bits and the performance of the FR system.

Just noticeable difference (JND), which characterizes the minimum perceptual threshold of signal change below which the change cannot be perceived by the ultimate receiver, has been widely used for improving the coding efficiency of image/video compression algorithms [9], [10], [11], [12], [13]. Therefore, we consider performing JND prediction before image compression over the probe set, as shown in Fig. 1(b). According to different ultimate receivers of images/video signals, existing JND models can be divided into two categories: JND models for the human visual system (HVS) and JND models for machine vision. Regarding JND models for the HVS, we can summarize them according to the domain over which the JND threshold is computed into two categories: 1) pixel-domain JND models [14], [15], [16], [17], [18], which predict the JND threshold for each pixel of the image; 2) subband domain JND models [19], [20], which predict the JND threshold for each sub-band by transforming the pixel-domain image into a specific transform domain (e.g., discrete cosine transformation (DCT), Karhunen-Loève transformation (KLT)). However, those HVS-oriented JND models are not proper models to estimate JND for the FR system, as the FR systems focus on discriminative information for analysis and recognition instead of luminance and texture information that most HVS-orientated models rely on.

Regarding JND models for machine vision, Zhang et al. [21] studied the impact of image/video compression on the performance of networks for image classification and object detection tasks. They firstly established a large-scale JND-annotated dataset including over 340,000 images for image classification and object detection tasks. The details can be found in Table I. Subsequently, two JND prediction models are proposed under no-reference and few-reference circumstances, named JND-MV-NR and JND-MV-FR, respectively. Jin et al. [22] proposed a JND prediction model for the
image classification task, which employs a semantic-guided redundancy assessment strategy to estimate the JND map via unsupervised learning. It is worth mentioning that the above-mentioned works investigate the JND on datasets with a large inter-variance, resulting in producing a large JND value for each image. Considering that face images shall have a small inter-class distance, more robust and discriminative features should be explored in JND prediction for the FR system.

In this paper, we make the first attempt to study the JND for the FR system. The major contributions can be summarized as follows:

- We establish a JND dataset served for the FR system. In particular, the proposed dataset consists of 3520 original images collected from the MegaFace dataset and 137,670 compressed images using VTM-15.0 intra-coding with 39 QP values ranging from 13 to 51. Subsequently, we conduct extensive recognition tests to search JND images for the FR system.

- We develop a novel JND prediction framework for face recognition systems. The fundamental idea behind the proposed framework is to perform identity information preservation and distortion maximization simultaneously when inferring JND images. Therefore, we devise attention-based feature decomposition (AFD) modules and employ the self-supervised optimization strategy.

- Experimental results show that the proposed JND prediction model achieves better performance than the state-of-the-art methods. Furthermore, compared with VTM-15.0 intra coding, the image coding framework combined with our proposed model can save more bits while maintaining the performance of the FR system.

II. RELATED WORKS

A. Deep Face Recognition

FR can be classified as face identification and face verification, both of which can essentially boil down to measuring the similarity of two face images. Traditional FR algorithms mainly use handcrafted features (e.g., Gabor [23], histogram of oriented gradients [24] and local binary patterns [25]) to calculate the similarity of two face images. Due to the limited representational capacity of handcrafted features, deep neural network (DNN) based FR algorithms have been widely studied, which are commonly optimized with a well-designed loss function to obtain more discriminative and generalizable feature representation. For example, Taigman et al. [3] trained the AlexNet [26] with a softmax loss function to propose the first DNN-based FR algorithm DeepFace. Schroff et al. [2] proposed a triplet loss to optimize the face embeddings extracted by GoogLeNet-24 [27] by minimizing the distance between an anchor and a positive sample and maximizing the distance between the anchor and a negative sample. Wen et al. [28] proposed a center loss to learn the center of each class and joint it with the softmax loss function to enhance the discriminative power of LeNet [29].

Recently, a popular line of designing the FR algorithm is introducing a large margin to separate hard samples strictly. Liu et al. [30] proposed the angular softmax loss (SphereFace) to learn the angularly discriminative feature on a hypersphere manifold. Following the work in [30], Wang et al. [4] proposed a new large margin loss (LMCL) by introducing an additive cosine margin penalty, and Deng et al. [5] proposed an additive angular-based margin loss (ArcFace). Based on the fact that easy samples and hard samples are of different importance during the training phase, Huang et al. [6] developed a curriculum learning loss (CurricularFace) to adjust the importance of different samples adaptively. To mitigate the overfitting problem, Zhong et al. [7] introduced the sigmoid-constrained hypersphere loss (SFace) to control intra-class and inter-class distances on the hypersphere manifold. Extending from the study of SphereFace, Liu et al. [31] incorporated the multiplicative margin to propose an improved SphereFace (SphereFaceR).

B. Just Noticeable Difference

According to the Weber’s law [32], JND can be defined as the minimum amount by which stimulus intensity must be changed to produce a noticeable variation in sensory experience. In the literature, the majority of pixel- and subband-domain JND research is based on human perception experience. Regarding the pixel-domain JND estimation, considering the effects of the luminance adaption (LA) and contrast masking (CM) on the threshold sensitivities of the HVS, Chou et al. [33] proposed a pioneer JND model to estimate the visibility threshold associated with each pixel of the image. Afterward, Yang et al. [34] re-formulated the CM model by taking the edge information into account. Liu et al. [16] further optimized the CM estimation algorithm by decomposing CM into two parts, i.e., edge masking (EM) and texture masking (TM). Since the HVS pays more attention to orderly information than disorderly information, Wu et al. [17] proposed a JND model based on the free-energy principle to compute JND thresholds in terms of the orderly contents and the disorderly contents of the original image. Inspired by the orientation selectivity mechanism in the primary visual cortex, Wu et al. [18] combined the pattern complexity with the luminance contrast to build a novel pattern masking for the JND prediction. Shen et al. [13] considered the fact that the JND profile is highly related to the local image content of the image and thus
proposed an adaptive block transform including two DCT block sizes of 16×16 and 8×8 to improve the accuracy of the JND estimation model. Bae et al. [19] developed a new texture complexity metric for estimating the JND, which can reflect the perceived complexity. Jiang et al. [20] estimated the JND threshold of the image via the minimum number of spectral components in KLT.

III. THE PROPOSED DATASET

Our proposed dataset is established with the aim of investigating the impact of image compression on the performance of the FR system. Therefore, both original face images and their corresponding compressed versions are included in our proposed dataset.

A. Data Collection

We first collect 3530 face images of 80 identities from the probe set in the publicly available refined version [5] of the MegaFace dataset [37] as original images. Those images cover various variations such as ages, poses, and styles and have a fixed resolution of 112×112. Subsequently, we apply the advanced reference encoding/decoding software based on the Versatile Video Coding (VVC) standard (VTM-15.0) [38] to encode original face images using different QPs. More specifically, we convert each original face image into YUV format and compress it with 39 QPs ranging from 13 to 51, where the larger QP value indicates the more severe signal-level corruption. The compressed YUV data is collected and converted back into RGB format to obtain compressed face images at different distortion levels. As a result, there are 3530 original face images and 137,670 compressed face images in total for the JND study.

B. Study of Just Noticeable Difference

Let x and y denote two images sampled from the probe and gallery sets, respectively. Moreover, \( h_t(x, y) \) represents the FR system, which determines whether two images are of the same identity at a given threshold \( t \). In particular, \( h_t(x, y) = 1 \) if the distance between feature representations of \( x \) and \( y \) is less than the threshold \( t \), which means \( x \) and \( y \) are of the same identity. By contrast, \( h_t(x, y) = 0 \) if the distance between feature representations of \( x \) and \( y \) is not less than the threshold \( t \), which implies \( x \) and \( y \) have different identities. In this study, our purpose is to find an image \( \hat{x} \) to satisfy \( h_t(\hat{x}, y) = h_t(x, y) \), where \( \hat{x} \) represents the distorted version of the original image \( x \) containing the maximum signal loss caused by image compression that the FR system does not perceive.

To this end, we conduct several recognition tests with different probe sets. More specifically, we first split all images in our dataset into 40 probe sets \( R = \{R(0)\}_{k=0}^{39} \) according to different QP values. The \( k \)-th probe set \( R(0) \) is in the form of \( \{r(0)_{s,i}\}_{i=1}^{N_s} \), where \( k \) and \( S \) mean the compression level and the number of identities, respectively, and \( N_s \) is the number of images of the \( s \)-th identity. The probe set is composed of all original (uncompressed) images when \( k = 0 \). Subsequently, in order to establish the gallery set, we check the original gallery set in the MegaFace dataset and find that there are many blurry face images and even some images without the human face. Herein, we use a face image quality metric [42] to evaluate the quality of all images within the original gallery set and sort them according to their corresponding quality scores from high to low. The top 10,000 face images with high recognizability \( D = \{d_m\}_{m=1}^{10^4} \) are selected. We combine them with all original images to establish the final gallery set \( M = R(0) \cup D \) of our recognition tests. During the recognition test on the \( k \)-th probe set, each image within the probe set is paired with images with the same identity within the gallery set \( M \) to form positive pairs \( P(k) = \{\{r(0)_{s,i}, r(0)_{s,j}\}_{i,j=1}^{N_s-1}\}_{i,s} \) and paired with images with different identities to form negative pairs \( C(k) = \{\{r(k)_{s,i}, d_m\}_{i=1}^{N_s} \}_{i,s} \). Then, we employ the FR system to verify all positive and negative pairs. After performing all recognition tests, for the \( i \)-th original image of the \( s \)-th identity, we seek to find a maximum QP (denoted as \( \hat{q} \)) such that \( r(\hat{q})_{s,i} \) satisfies the following constraints:

\[
\sum_{j \neq i}^{N_s-1} h_{t_0}(r(\hat{q})_{s,i}, r(\hat{q})_{s,j}) \geq \sum_{j \neq i}^{N_s-1} h_{t_0}(r(0)_{s,i}, r(0)_{s,j}), \tag{1}
\]

\[
\sum_{m=1}^{10^4} h_{t_0}(r(\hat{q})_{s,i}, d_m) \leq \sum_{m=1}^{10^4} h_{t_0}(r(0)_{s,i}, d_m), \tag{2}
\]

where \( t_0 \) represents the threshold with respect to \( 10^{-4} \) false acceptance rate (FAR) on \( C(0) \). As a result, \( \hat{q} \) and \( r(\hat{q})_{s,i} \) are the corresponding JND point and JND image of the original image \( r(0)_{s,i} \), respectively.

We initially enroll five FR algorithms trained on different circumstances (i.e., loss functions and backbone architectures), both of which are pre-trained algorithms provided by the
TABLE I
COMPARISON OF EXISTING BENCHMARK JND DATABASES FOR MACHINE VISION.

| Dataset          | # of original images | Source of original images | Distortion type | QP | Ground-truth JND | Machine vision task                      |
|------------------|----------------------|---------------------------|-----------------|----|------------------|------------------------------------------|
| Zhang et al. [21]| 355,541              | PASCAL VOC [39] and MS COCO [40] | HEVC [41]       | [18, 22, 27, 32, 38, 41, 43, 45, 47, 49, 51] | Available | Image classification/object detection |
| Proposed         | 3,520                | MegaFace [37]             | VTM-15.0 [58]   | [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51] | Avaliable | Face recognition |

Fig. 3. Results of the JND noticeable difference study for the FR system. (a) The JND points distribution of original face images in the proposed dataset; (b) BPP and “Accuracy” comparison among JND and different QPs.

corresponding authors. Fig. 2 shows the recognition performance of all candidate FR algorithms when the probe images are compressed at different QP values, where the “Accuracy” refers to the true acceptance rate (TAR) of the FR algorithm at its corresponding threshold $t_0$. From the experimental results, one can easily observe that the recognition performances of all candidate FR systems decrease as the QP value increases. When the loss function is fixed, the FR algorithm employing the ResNet-100 [37] network achieves higher accuracy and robustness than the FR algorithm employing other architectures. For example, when the QP value enlarges from 13 to 37, ArcFace-IR100 only reduces 1.3% of accuracy while ArcFace-MobileNet reduces 7.14% of accuracy. Additionally, we observe that although there are a variety of loss functions, FR algorithms with the same backbone architecture show a significant performance degradation at the same QP value, and the recognition performance of all algorithms is catastrophically compromised when the QP value rises to 51. Based on the above observations, we employ ArcFace-IR100 as the default algorithm of our FR system. Fig. 3(a) is the JND points distribution of original face images in our proposed dataset. From the results of Fig. 3(a), we can see that the distribution of JND points spans a wide range, which implies that the levels of discriminative information carried by original images in the proposed dataset are various. Subsequently, we encode all original images with their corresponding JND points and different QP points, respectively. The average bits per pixel (BPP) over the compressed images and the accuracy performance of the FR system on the compressed images are shown in Fig. 3(b). It can be clearly find that the JND study sheds light on the optimization of the image coding framework. In particular, the image coding framework using JND prediction leads to 81% bit-rate savings at least while keeping the performance of the FR system, and it achieves 5% of performance improvement under the same bit-rate level.

IV. THE PROPOSED MODEL

A. Motivation

Let $x$ denote an input face image. The distorted image $\hat{x}$ can be represented as the linear combination of $x$ and $e$, i.e., $\hat{x} = x + e$, where $e$ is the disturbance caused by image compression. According to the general definition of the JND [43], the JND prediction problem for the FR system can be formulated as follows,

$$\max e, \text{ s.t. } \|x_{id} - \hat{x}_{id}\|_2 \leq \epsilon,$$

where $x_{id}$ and $\hat{x}_{id}$ represent the identity information of $x$ and $\hat{x}$, respectively. $\epsilon$ is the threshold of the distance between $x_{id}$ and $\hat{x}_{id}$ below which $\hat{x}$ and $x$ can be recognized as being from the same subject. Our goal is to design a JND prediction model for the FR system, which can automatically infer the JND version of each input face image. To this end, we
we employ the encoder and decoder layers. Regarding the encoder, we
encoder-decoder architecture with skip connections between
various face variations, we embed the AFD module in the
positive and negative pairs are introduced to optimize the
residual feature to generate the residual map without losing
encoder to separate the mixed feature into identity and resid-
ual components. In this manner, the decoder can utilize the
 encoder to jointly tackles them with two parts: 1) identity information
preservation; 2) distortion maximization. Regarding identity
information preservation, considering the face image can be
jointly represented by the intrinsic identity information and
various face variations, we embed the AFD module in the
encoder to separate the mixed feature into identity and resid-
ual components. In this manner, the decoder can utilize the
residual feature to generate the residual map without losing
the identity information. Regarding distortion maximization,
positive and negative pairs are introduced to optimize the
learning procedure of identity and residual features in a self-
supervised manner. The overview of our proposed framework
is shown in Fig. 4. To be specific, the input of the AFD module is the mixed
feature (denoted as $F_{(s,i)^{(0)}}$) extracted by the previous
module (i.e., EnConv or Basic Block). We first aggregate spatial features of each channel through mean and standard
deviation calculations (denoted as Mean and Std), generating
two spatial descriptors (denoted as $F_{avg} \in \mathbb{R}^{C \times 1 \times 1}$ and $F_{std} \in \mathbb{R}^{C \times 1 \times 1}$). Then, we concatenate $F_{avg}$ and $F_{std}$ and pass the resulted feature through two 1 × 1 convolutions with
stride of 1 (denoted as Conv1×1) to generate the attention
feature. Subsequently, we fuse the attention feature and the
input feature $F$ to produce the identity feature (denoted as
$F_{id} \in \mathbb{R}^{C \times H \times W}$). As suggested in [49], [50], the residual
feature (denoted as $F_{res} \in \mathbb{R}^{C \times H \times W}$) can be generated as:
$F_{res} = F - F_{id}$. Regarding the decoder, it contains multiple
upsampling layers for upsampling the spatial resolution of the
input feature and deconvolutional units (denoted as DeConv)
for reducing the size of channels. In particular, the input
feature first passes through the upsampling layer, which is
design it based on ResNet-50 [5], [44], which consists of a
convolution unit (denoted as EnConv), four basic blocks and
five AFD modules. EnConv represents a 3 × 3 convolution with
stride of 1 followed by a batch-normalization (BN) layer [45]
and PReLU activation function. The structure of the basic
block is similar to the residual block of ResNet-50 in [44],
which is stacked by cascaded convolutional layers and the
residual units. The human perception process presents the
importance of the attention mechanism [46]. We use the AFD
module [47], [48] to capture the discriminative feature that the
FR system relies on for image understanding and recognition.

The detailed structure of the AFD module is shown in
Fig. 5. To be specific, the input of the AFD module is the mixed
feature (denoted as $F \in \mathbb{R}^{C \times H \times W}$) extracted by the previous
module (i.e., EnConv or Basic Block). We first aggregate
spatial features of each channel through mean and standard
deviation calculations (denoted as Mean and Std), generating
two spatial descriptors (denoted as $F_{avg} \in \mathbb{R}^{C \times 1 \times 1}$ and $F_{std} \in \mathbb{R}^{C \times 1 \times 1}$). Then, we concatenate $F_{avg}$ and $F_{std}$ and pass the resulted feature through two 1 × 1 convolutions with
stride of 1 (denoted as Conv1×1) to generate the attention
feature. Subsequently, we fuse the attention feature and the
input feature $F$ to produce the identity feature (denoted as
$F_{id} \in \mathbb{R}^{C \times H \times W}$). As suggested in [49], [50], the residual
feature (denoted as $F_{res} \in \mathbb{R}^{C \times H \times W}$) can be generated as:
$F_{res} = F - F_{id}$. Regarding the decoder, it contains multiple
upsampling layers for upsampling the spatial resolution of the
input feature and deconvolutional units (denoted as DeConv)
for reducing the size of channels. In particular, the input
feature first passes through the upsampling layer, which is
implemented by the bilinear interpolation. The upsampled feature is concatenated to the residual feature of the encoder at the corresponding location via the skip connection. It is worthwhile noting that the skip connections make only the residual features available for the decoder, avoiding the introduction of identity features during the process of residual map generation.

C. Loss Function

1) Identity Loss: The identity information that the FR system relies on for recognition is hidden in the feature space and does not have the corresponding ground-truth representation. This motivated us to extract the intrinsic identity feature in a self-supervised learning manner. Benefiting from the properties of identity consistency and identity diversity, we employ self-supervised contrastive learning [51], [52], [53] to capture inter-person and intra-person correspondences. Given an anchor image, we first find the appropriate positive and negative samples with respect to the anchor image. To be specific, considering that the more robust feature can be mined via the more significant intra-person variation, we use the positive image with the largest variation to the anchor image. The negative sample is produced within the training minibatch by shuffling training samples of a batch. After that, we maximize information from identity features within positive samples while minimizing those within negative samples. Following [51], [52], the identity loss is formulated as follows:

\[
L_{id} = \log \left(1 + e^{\|v - v^+\|_1}\right) + \log \left(1 + e^{-\|v - v^-\|_1}\right),
\]

where \(v, v^+\) and \(v^-\) are identity features of the anchor, positive and negative images extracted by the fifth AFD module, respectively. \(\| \cdot \|_1\) denotes \(\ell_1\) distance.

2) Pixel Loss: In order to force the predicted JND image to be closer to the ground-truth JND image, the straight-forward method is to constrain the distance between the two images in the pixel domain. Herein, the pixel loss is computed as:

\[
L_{pixel} = \| r_{s,i}^{(q)} - r_{s,i}^{(0)} \|_1,
\]

where \(r_{s,i}^{(0)}\) is the predicted JND image of the original probe image \(r_{s,i}^{(0)}\).

3) Content Loss: We use the content loss [54], [55] to make the predicted JND image consistent with the ground-truth JND image from the aspects of semantic content. The content loss is calculated as follows,

\[
L_{content} = \sum_{k=1}^{K} \left\| \phi_k \left( r_{s,i}^{(q)} \right) - \phi_k \left( r_{s,i}^{(0)} \right) \right\|_2^2,
\]

where \(\phi_k\) indicates the mapping function that project the image to the feature representation at the \(k\)-th layer of the VGG space and \(K\) is the total number of layers. In this work, we adopt feature representations from the Relu_1_1, Relu_2_1, Relu_3_1, Relu_4_1, and Relu_5_1 layers in VGG-19 network.

4) Total Loss: By jointly considering identity loss, pixel loss, and content loss, the final loss is defined as the weighted sum of these losses, that is,

\[
L_{total} = \lambda_{id}L_{id} + \lambda_{pixel}L_{pixel} + \lambda_{content}L_{content},
\]

where \(\lambda_{id}, \lambda_{pixel},\) and \(\lambda_{content}\) are weighting parameters to balance the relative importance of \(L_{id}, L_{pixel},\) and \(L_{content}\).

V. EXPERIMENT

In this section, we first present the implementation details of the proposed model. We then compare our proposed model with the existing JND models in terms of accuracy performance on predicting the JND image. Furthermore, we demonstrate the effectiveness of the proposed method in optimizing image coding for the FR system.

A. Implementation Details

We implement our model using PyTorch [57]. The inputs of our network are resized to 112 \(\times\) 112 \(\times\) 3, and the minibatch size is set as 12. The whole network is trained via the Adam [58] optimizer. The maximum epoch is 100. The learning rate is set by 1e-4 at the first 50 epochs. After 50 epochs, the learning rate is reduced to its 1/5 every 1 epoch. The weighting parameters \(\lambda_{id}, \lambda_{pixel},\) and \(\lambda_{content}\) in Eqn. (7) are set as 0.5, 5.0, and 0.1. In the experiments, to reduce the negative impact of excessive inter-class distance on the JND study for the FR system, we remove the original probe images that the FR system can not recognize correctly from our proposed dataset. As a result, 3501 original images and 136,539 compressed images are employed.

B. Comparisons With State-of-the-Art Methods

In this experiment, we aim to explore the prediction accuracy of the proposed model in directly predicting JND images for the FR system. Several classical and state-of-the-art HVS-oriented JND prediction models, including Zhang05 [56], Yang05 [14], Liu10 [16], Wu17 [18], Shen21 [13], Jiang22 [20], are employed for performance comparison. To ensure a fair comparison, all the source codes of the competing models are obtained from their authors. We randomly split our dataset into five subsets to conduct five-fold cross-validation. In each trial, four subsets are selected for training the proposed JND prediction model, and the remaining one subset is used for testing. Five objective metrics are selected as the evaluation criteria. They are peak signal-to-noise ratio (PSNR), learned perceptual image patch similarity (LPIPS) metric, similarity distribution distance for face image quality assessment (SDD-FIQA) metric [42], stochastic embedding robustness for face image quality (SER-FIQ) [59] evaluation metric and FaceQnet_V1 [60], the latter three of which are no-reference image quality assessment metrics specifically designed to estimate the recognizability of face images for the FR systems. Regarding PSNR and LPIPS, we compute the average similarity yielded between the predicted JND images and their ground truth JND images over the test set. Regarding SDD-FIQA, SER-FIQ and FaceQnet_V1, we
compare the average difference between recognizability scores of the predicted JND images and those of the ground truth JND images over the test set. The higher values of PSNR and the lower value of LPIPS, \( \Delta \text{SDD-FIQA}, \Delta \text{SER-FIQ} \) and \( \Delta \text{FaceQnet}_V1 \) represent better performance, which indicates that the predicted JND image from the model is closer to the ground truth JND image. The results are listed in Table I1 where \( G_k \) means the \( i \)-th trial, the first-, second- and third-best performances are highlighted in red, blue and black bold, respectively.

From the experimental results, one can obviously observe that compared with the competing JND prediction models, our proposed model achieves higher PSNR values and lower LPIPS, \( \Delta \text{SDD-FIQA}, \Delta \text{SER-FIQ} \) and \( \Delta \text{FaceQnet}_V1 \) values on both \( G_1, G_2, G_3, G_4 \) and \( G_5 \). The main reason is the gap in the perceptual characteristics between the FR system and the HVS. By considering the properties of the FR system in recognizing the face image, we combine the low-level structure and high-level semantic information into the JND image inference procedure. Therefore, the proposed model can be more consistent with the perception of the FR system than the classical and state-of-the-art JND prediction models.

To make a more explicit comparison between these HVS-orientated JND prediction models and our proposed model, we provide some examples of the predicted JND and residual images as shown in Fig.6 and Fig.7. Comparing the predicted results of our proposed model (see Fig.6(m)) with that of Zhang05, Wu17 and Shen21 (see Fig.6(h), (j) and (k)), one can find that Zhang05, Wu17 and Shen21 deem that the low luminance area contains more redundant information. This is because Zhang05, Wu17 and Shen21 are modeled by considering the LA mechanism of the HVS. Such luminance-dependent JND prediction models fail to estimate the accurate JND map under the low luminance environment. Since the HVS is sensitive to the foreground regions of ordered textures and insensitive to the complex textured regions, those HVS-orientated JND prediction models cannot correctly find discriminative regions that the FR system focuses on specifically when estimating the occluded face images. For example, the complex textured regions (e.g., hair, eyebrows) that might provide some useful clues about the identity of the subject are underestimated (see Fig.7(i), (j), (k)), and the irrelevant information carried by the occlusion (i.e., the mobile phone in the Fig.7(a)) are received overmuch attention (see Fig.7(i) and (l)). All these reveal that our proposed model can obtain better consistency with the perception of the FR system and make an accurate estimation.
Fig. 6. Visual comparison of JND images and residual images generated by different JND models. (a): A low-brightness original face image selected from our proposed dataset; (b)&(h): the ground-truth JND image and the residual image; (c)&(i): the JND image and the residual image generated by Zhang05 [56], the PNSR value between (b) and (c) is 31.4499dB; (d)&(j): the JND image and the residual image generated by Wu17 [18], the PNSR value between (b) and (d) is 28.3692dB; (e)&(k): the JND image and the residual image generated by Shen21 [13], the PNSR value between (b) and (e) is 27.9426dB; (f)&(l): the JND image and the residual image generated by Jiang22 [20], the PNSR value between (b) and (f) is 33.4728dB; (g)&(m): the JND image and the residual image generated by our proposed model, the PNSR value between (b) and (g) is 36.1064dB.

Fig. 7. Visual comparison of JND images and residual images generated by different JND models. (a): An occluded original face image selected from our proposed dataset; (b)&(h): the ground-truth JND image and the residual image; (c)&(i): the JND image and the residual image generated by Zhang05 [56], the PNSR value between (b) and (c) is 26.0927dB; (d)&(j): the JND image and the residual image generated by Wu17 [18], the PNSR value between (b) and (d) is 24.3461dB; (e)&(k) are the corresponding the JND image and the residual image generated by Shen21 [13], the PNSR value between (b) and (e) is 26.7643dB; (f)&(l): the JND image and the residual image generated by Jiang22 [20], the PNSR value between (b) and (f) is 26.8419dB; (g)&(m): the JND image and the residual image generated by our proposed model, the PNSR value between (b) and (g) is 29.8513dB.

C. Application in Remote FR System

In this subsection, we evaluate the performance of the proposed JND model in improving the efficiency of image compression. Here, we denote \( x \) and \( x^{(q)} \) as the original face image and its corresponding compressed image with QP equals to \( q \), respectively, \( \hat{x} \) is the predicted JND image produced by the JND prediction model. As such, we can search the optimal \( \hat{q} \) of the predicted JND image \( \hat{x} \) as follows,

\[
\hat{q} = \arg \max_q g \left( x^{(q)}, \hat{x} \right),
\]

where \( g \left( x^{(q)}, \hat{x} \right) \) denotes PNSR value of \( x^{(q)} \) and \( \hat{x} \). The optimal \( \hat{q} \) is viewed as the predicted JND point of \( x \).

We compare our proposed model with two JND models for machine vision, i.e., JND-MV-NR [21] and JND-MV-FR [21]. It is worth mentioning that JND-MV-NR and JND-MV-FR are composed of multiple classifiers, which can directly predict the optimal QP value for each input image. JND-MV-NR and JND-MV-FR trained on our proposed dataset are employed for comparison. We conduct the evaluation experiment by applying each JND model to VTM-15.0 intra coding framework. More specifically, we select the VTM 5.0 intra coding frameworks with QPs of 24, 32, 35, 36, 37, 40, 41 and 42 as the anchors. All test images in each trial are compressed by the VTM-15.0 intra coding with their corresponding optimal QP values produced by the JND model and the employed anchor QP values, respectively. The compressed images are fed into the FR system to examine whether they can be recognized correctly. The evaluation results in terms of BPP and “Accuracy” (denoted ACC.) are reported in Table III, from
which one can observe that 1) compared with the anchors under the similar bit-rate level, JND-MV-NR and JND-MV-FR degrade the accuracy performance of the FR system; 2) our proposed coding framework achieves bit-rate savings while improving the accuracy compared with QP35 of $G_1$, QP36 of $G_2$, $G_3$, $G_4$ and $G_5$, which means our proposed coding framework is able to achieve a better balance between the accuracy of the FR system and bit-rate saving than constant QP coding; 3) compared with QP37 and QP42, where the former is the point at which the performance of the FR system begins to decrease significantly while the latter is the one with the largest distribution of JND points for all original images in the proposed database, the proposed coding framework achieves higher accuracy, which means our proposed coding framework can assign a more reasonable QP for each image. Those results provide useful evidence that the proposed model has great potential in image coding for the FR system.

D. Ablation Studies

1) Optimization Strategy Analysis: To investigate the contribution of the self-supervised contrastive learning strategy, we perform the ablation study based on different optimization strategies. Considering that the crucial role of the self-supervised contrastive learning strategy is to optimize the learned feature space, we compare the distances between the identity features of inter-class and intra-class pairs. The results are listed in Table [V] where $Dist_{pos}$ represents the average $\ell_1$ distance of identity features between all predicted JND images and their positive samples, and $Dist_{neg}$ represents the average $\ell_1$ distance of identity features between all predicted JND images and their negative samples. The lower $Dist_{pos}$ and higher $Dist_{neg}$ values indicate that the predicted JND images contain identity information with higher robustness and higher discriminability, respectively. Ours w/o $L_{id}$ is the proposed model without using the contrastive learning strategy, i.e., the whole framework is only trained with $L_{pixel}$ and $L_{content}$. From the results of Table [V] one can observe the proposed model using self-supervised contrastive learning strategy achieves lower $Dist_{pos}$ and higher $Dist_{neg}$ values.

2) Architecture Analysis: To better understand the individual contributions of the AFD module and skip connection, we conduct several comparisons between our proposed model and its two variants: 1) Ours w/o AFD and Skip Connection: removing both the AFD modules and skip connections. That is an AFD model is embedded between $G_2$, $G_3$, $G_4$ and $G_5$; 2) Ours w/ AFD, w/o Skip Connection: removing the skip connections. That is an AFD model is embedded between two convolutional blocks of the encoder, and the input of the decoder is the residual feature decomposed from the feature of the last stage of the encoder to generate the residual map by the decoder. Noted that the $L_{id}$ is associated with the AFD module, we also remove the $L_{id}$; 2) Ours w/ AFD, w/o Skip Connection: removing the skip connections. That is an AFD model is embedded between two convolutional blocks of the encoder, and the input of the decoder is the residual feature decomposed from the feature of the last stage of the encoder. The average PSNR, LPIPS and $\Delta$SDD-FIQA results on $G_1$, $G_2$, $G_3$, $G_4$ and $G_5$ are listed in Table [V] from which one can easily see that compared with the two variants, our proposed model achieves around 10.89% and 10.59% improvements on PSNR, 51.48% and 51.26%, reductions on LPIPS, 43.65% and 39.42% reductions on $\Delta$SDD-FIQA. Therefore, our proposed model using the AFD module and skip connection allows multi-level residual features learned from the encoder to be embedded in the decoder so that it can improve the accuracy of the JND images.

### Table III

| Index | Measures | JND-MV-NR [21] | JND-MV-FR [21] | Proposed | QP24 | QP32 | QP35 | QP36 | QP37 | QP40 | QP41 | QP42 |
|-------|----------|----------------|----------------|----------|------|------|------|------|------|------|------|------|
| $G_1$ | BPP      | 0.21           | 0.22           | 0.38     | 1.44 | 0.60 | 0.42 | 0.37 | 0.33 | 0.24 | 0.22 | 0.20 |
|       | ACC.     | 88.9%          | 91.41%         | 99.23%   | 99.75%| 99.53%| 99.17%| 99.00%| 98.50%| 95.08%| 92.09%| 87.61%|
| $G_2$ | BPP      | 0.21           | 0.22           | 0.36     | 1.44 | 0.59 | 0.41 | 0.37 | 0.33 | 0.24 | 0.22 | 0.20 |
|       | ACC.     | 90.82%         | 88.52%         | 98.71%   | 99.62%| 99.26%| 98.82%| 98.66%| 97.99%| 94.17%| 91.05%| 86.13%|
| $G_3$ | BPP      | 0.22           | 0.22           | 0.38     | 1.44 | 0.59 | 0.42 | 0.37 | 0.33 | 0.24 | 0.22 | 0.20 |
|       | ACC.     | 91.33%         | 90.95%         | 98.85%   | 99.75%| 99.54%| 99.03%| 98.63%| 98.16%| 94.98%| 92.47%| 88.75%|
| $G_4$ | BPP      | 0.22           | 0.21           | 0.36     | 1.43 | 0.59 | 0.42 | 0.37 | 0.33 | 0.24 | 0.22 | 0.20 |
|       | ACC.     | 90.96%         | 88.83%         | 98.97%   | 99.64%| 99.39%| 99.01%| 98.75%| 98.35%| 94.91%| 92.31%| 88.02%|
| $G_5$ | BPP      | 0.22           | 0.23           | 0.34     | 1.43 | 0.59 | 0.42 | 0.37 | 0.33 | 0.24 | 0.22 | 0.20 |
|       | ACC.     | 90.20%         | 92.24%         | 98.71%   | 99.65%| 99.53%| 99.11%| 98.70%| 98.38%| 94.60%| 92.02%| 88.25%|
| Average| BPP      | 0.22           | 0.22           | 0.36     | 1.44 | 0.59 | 0.42 | 0.37 | 0.33 | 0.24 | 0.22 | 0.20 |
|       | ACC.     | 90.44%         | 90.39%         | 98.90%   | 99.68%| 99.45%| 99.03%| 98.75%| 98.28%| 94.86%| 91.99%| 87.75%|

### Table IV

| Method                                      | Dist$_{pos}$ | Dist$_{neg}$ |
|---------------------------------------------|--------------|--------------|
| Ours w/o $L_{id}$                          | 0.8834       | 1.4146       |
| Ours                                        | 0.8807       | 1.4147       |

### Table V

| Method                                      | PSNR | LPIPS | $\Delta$SDD-FIQA |
|---------------------------------------------|------|-------|------------------|
| Ours w/o AFD and Skip Connection            | 27.9070 | 0.2665 | 16.2935         |
| Ours w/ AFD, w/o Skip Connection            | 27.9842 | 0.2642 | 15.1565         |
| Ours                                        | 30.9470 | 0.1293 | 9.1811          |
prediction.

VI. CONCLUSION

We have made several efforts in the JND research of FR systems: 1) establish a JND dataset; 2) propose a JND prediction model; 3) incorporate the model into the remote FR system. Regarding the JND dataset, 3530 original images and 137,670 compressed images generated by VTM-15.0 are available for research usage. Regarding the JND prediction model, the main novelty is that our model is able to preserve the identity information and remove the redundancy simultaneously in a self-supervised manner. Extensive experiments conducted over our proposed JND prediction model and the state-of-the-art JND prediction models have clearly demonstrated that the predicted JND images by our proposed model are much closer to the ground-truth JND images in terms of low-level and high-level features. In addition, we also show that the image coding framework combined with our proposed model can gain performance improvement. Namely, the image coding algorithm guided by our proposed model is able to assign more appropriate QP values to face images, thereby achieving bit-rate savings while maintaining similar accuracy to the FR system.

REFERENCES

[1] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” Adv. Neural Inf. Process. Syst., vol. 27, pp. 1988–1996, 2014.
[2] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” IEEE Conf. Comput. Vis. Pattern Recog., pp. 815–823, 2015.
[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” IEEE Conf. Comput. Vis. Pattern Recog., pp. 1701–1708, 2014.
[4] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” IEEE Conf. Comput. Vis. Pattern Recog., pp. 5265–5274, 2018.
[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” IEEE Conf. Comput. Vis. Pattern Recog., pp. 4690–4699, 2019.
[6] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, “Curricularface: Adaptive curriculum learning loss for deep face recognition,” IEEE Conf. Comput. Vis. Pattern Recog., pp. 5901–5910, 2020.
[7] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen, “Face: Sigmoid-constrained hypersphere loss for robust face recognition,” IEEE Trans. Image Process., vol. 30, pp. 2587–2598, 2021.
[8] S. Wang, S. Wang, W. Yang, X. Zhang, S. Ma, and W. Gao, “Towards analysis-friendly face representation with scalable feature and texture compression,” IEEE Trans. Multimedia, pp. 1–1, 2021.
[9] X. Zhang, S. Wang, K. Gu, W. Lin, S. Ma, and W. Gao, “Just noticeable difference-based perceptual optimization for jpeg compression,” IEEE Signal Process. Lett., vol. 24, no. 1, pp. 96–100, 2017.
[10] T. Tian, H. Wang, L. Zhu, C.-C. J. Kuo, and S. Kwong, “Just noticeable difference level prediction for perceptual image compression,” IEEE Trans. Broadcast., vol. 66, no. 3, pp. 690–700, 2020.
[11] H. Liu, Y. Zhang, H. Zhang, C. Fan, S. Kwong, C.-C. J. Kuo, and X. Fan, “Deep learning-based picture-wise just noticeable distortion prediction model for image compression,” IEEE Trans. Image Process., vol. 29, pp. 641–656, 2020.
[12] M. Zhou, X. Wei, S. Kwong, W. Jia, and B. Fang, “Just noticeable distortion-based perceptual rate control in hevc,” IEEE Trans. Image Process., vol. 29, pp. 7603–7614, 2020.
[13] X. Shen, Z. Ni, W. Yang, X. Zhang, S. Wang, and S. Kwong, “Just noticeable distortion profile inference: A patch-level structural visibility learning approach,” IEEE Trans. Image Process., vol. 30, pp. 26–38, 2021.
[14] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, “Motion-acknowledged residue preprocessing in video coding based on just-noticeable-distortion profile,” IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 6, pp. 742–752, 2005.
[15] S. Wang, L. Ma, Y. Fang, W. Lin, S. Ma, and W. Gao, “Just noticeable difference estimation for screen content images,” IEEE Trans. Image Process., vol. 25, no. 8, pp. 3838–3851, 2016.
[16] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, “Just noticeable difference for images with decomposition model for separating edge and textured regions,” IEEE Trans. Circuits Syst. Video Technol., vol. 20, no. 11, pp. 1648–1652, 2010.
[17] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi, “Just noticeable difference estimation for images with free-energy principle,” IEEE Trans. Multimedia, vol. 15, no. 7, pp. 1705–1710, 2013.
[18] J. Wu, L. Li, W. Dong, G. Shi, W. Lin, and C.-C. J. Kuo, “Enhanced just noticeable difference model for images with pattern complexity,” IEEE Trans. Image Process., vol. 26, no. 6, pp. 2682–2693, 2017.
[19] S.-H. Bae and M. Kim, “A novel generalized dt-based jnd profile based on an elaborate cm-jnd model for variable block-sized transforms in monochrome images,” IEEE Trans. Image Process., vol. 23, no. 8, pp. 3227–3240, 2014.
[20] Q. Jiang, Z. Liu, S. Wang, F. Shao, and W. Lin, “Towards top-down just noticeable difference estimation of natural images,” IEEE Trans. Image Process., 2022.
[21] R. Zhang, S. Wang, X. Zhang, S. Ma, and W. Gao, “Just recognizable distortion for machine vision oriented image and video coding,” Int. J. Comput. Vis., vol. 129, p. 2889–2906, 2021.
[22] J. Jin, X. Zhang, X. Fu, H. Zhang, W. Lin, J. Lou, and Y. Zhao, “Just noticeable difference for deep machine vision,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 6, pp. 3452–3461, 2022.
[23] C. Liu and H. Wechsler, “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition,” IEEE Trans. Image Process., vol. 11, no. 4, pp. 467–476, 2002.
[24] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, “On effectiveness of histogram of oriented gradient features for visible to near infrared face matching,” Int. Conf. Pattern Recogn., pp. 1788–1793, 2014.
[25] T. Ahoonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, 2006.
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Adv. Neural Inf. Process. Syst., vol. 25, pp. 1097–1105, 2012.
[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” IEEE Conf. Comput. Vis. Pattern Recogn., pp. 1–9, 2015.
[28] 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.
[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
[30] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” IEEE Conf. Comput. Vis. Pattern Recogn., July 2017.
[31] W. Liu, Y. Wen, B. Raj, R. Singh, and A. Weller, “Sphereface reved: Unifying hyperspherical face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., 2022.
[32] Weber’s Law of Just Noticeable Difference, 2017. [Online]. Available: http://apps.usd.edu/doglab/WeberLaw.html
[33] C.-H. Chu and Y.-C. Li, “A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile,” IEEE Trans. Circuits Syst. Video Technol., vol. 5, no. 6, pp. 467–476, 1995.
[34] X. Yang, W. Ling, Z. Lu, E. Ong, and S. Yao, “Just noticeable distortion model and its applications in video coding,” Signal Process. Image Commun., vol. 20, no. 7, pp. 662–680, 2005.
[35] Z. Wei and K. N. Ngan, “Just noticeable distortion profile for grey scale image/video in dct domain,” IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 3, pp. 337–346, 2009.
[36] L. Ma, K. N. Ngan, F. Zhang, and S. Li, “Adaptive block-size transform based just-noticeable distortion difference model for images/videos,” Signal Process. Image Commun., vol. 26, no. 3, pp. 162–174, 2011.
[37] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard, “The megamark benchmark: 10 million faces for recognition at scale,” IEEE Conf. Comput. Vis. Pattern Recogn., pp. 4873–4882, 2016.
[38] Versatile Video Coding, Standard Rec. ITU-T H.266 and ISO/IEC 23090-3, 2020.
[39] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” Eur. Conf. Comput. Vis., pp. 740–755, 2014.

[41] High Efficiency Video Coding, Standard Rec. ITU-T H.265 and ISO/IEC 23008-2, 2014.

[42] F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y.-G. Wang, “Ssd-fiqa: Unsupervised face image quality assessment with similarity distribution distance,” IEEE Conf. Comput. Vis. Pattern Recog., pp. 7670–7679, 2021.

[43] W. Liu and G. Ghinea, “Progress and opportunities in modelling just noticeable difference (jnd) for multimedia,” IEEE Trans. Multimedia, 2021.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[45] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” Int. Conf. Learn. Represent, vol. 37, pp. 448–456, 2015.

[46] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, “Recurrent models of visual attention,” Neural Inf. Process. Syst., pp. 2204–2212, 2014.

[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cham: Convolutional block attention module,” Eur. Conf. Comput. Vis., pp. 3–19, 2018.

[48] 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.

[49] H. Wang, D. Gong, Z. Li, and W. Liu, “Decorrelated adversarial learning for age-invariant face recognition,” IEEE Conf. Comput. Vis. Pattern Recog., June 2019.

[50] Z. Huang, J. Zhang, and H. Shan, “When age-invariant face recognition meets face age synthesis: A multi-task learning framework,” IEEE Conf. Comput. Vis. Pattern Recog., pp. 7282–7291, June 2021.

[51] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, “Learning deep representations by mutual information estimation and maximization,” Int. Conf. Learn. Represent., 2018.

[52] E. H. Sanchez, M. Serrurier, and M. Ortmann, “Learning disentangled representations via mutual information estimation,” Eur. Conf. Comput. Vis., 2019.

[53] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2019.

[54] J. Justin, A. Alexandre, and F.-F. Li, “Perceptual losses for real-time style transfer and super-resolution,” Eur. Conf. Comput. Vis., pp. 694–711, 2016.

[55] Z. Ni, W. Yang, S. Wang, L. Ma, and S. Kwong, “Unpaired image enhancement with quality-attention generative adversarial network,” in ACM Int. Conf. Multimedia, 2020, pp. 1697–1705.

[56] X. Zhang, W. Lin, and P. Xue, “Improved estimation for just-noticeable visual distortion,” Signal Process., vol. 85, no. 4, pp. 795–808, 2005.

[57] 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.

[58] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” Int. Conf. Learn. Represent., 2015.

[59] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuecher, and A. Kuijper, “Serfiq: Unsupervised estimation of face image quality based on stochastic embedding robustness,” IEEE Conf. Comput. Vis. Pattern Recog., pp. 5651–5660, 2020.

[60] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay, “Biometric quality: review and application to face recognition with faceqnet,” arXiv preprint arXiv:2006.03298v2, 2020.