Understanding Deep Face Representation via Attribute Recovery

Min Ren, Yuhao Zhu, Yunlong Wang, Yongzhen Huang, *Senior Member, IEEE*, and Zhenan Sun, *Senior Member, IEEE*

***Abstract*— Deep neural networks have proven to be highly effective in the face recognition task, as they can map raw samples into a discriminative high-dimensional representation space. However, understanding this complex space proves to be challenging for human observers. In this paper, we propose a novel approach that interprets deep face recognition models via facial attributes. To achieve this, we introduce a two-stage framework that recovers attributes from the deep face representations. This framework allows us to quantitatively measure the significance of facial attributes in relation to the recognition model. Moreover, this framework enables us to generate sample-specific explanations through counterfactual methodology. These explanations are not only understandable but also quantitative. Through the proposed approach, we are able to acquire a deeper understanding of how the recognition model conceptualizes the notion of "identity" and understand the reasons behind the error decisions made by the deep models. By utilizing attributes as an interpretable interface, the proposed method marks a paradigm shift in our comprehension of deep face recognition models. It allows a complex model, obtained through gradient backpropagation, to effectively "communicate" with humans. The source code is available here, or you can visit this website: *https://github.com/RenMin1991/Facial-Attribute-Recovery*.**

***Index Terms*— Interpretability, face recognition, facial attribute, counterfactual sample.**

I. Introduction

In THE last few years, the realm of deep learning has witnessed remarkable advancements in bolstering the precision facet across a myriad of endeavors. When it comes to the field of facial recognition, the profundity of deep learning models lies in their ability to map facial images into intricate high-dimensional deep face representations, thereby facilitating similarity assessments [1], [2], [3]. Notwithstanding their notable advancements, current deep learning-based face recognition methods face serious bottlenecks and numerous challenging issues. For instance, improving performance on unseen samples is challenging without expanding the training set; deep learning models are susceptible to adversarial attacks, where minor perturbations to input samples can deceive the models; moreover, data imbalance persists, such as racial bias. These issues necessitate diagnosing and addressing the shortcomings of deep learning models. However, the main obstacle lies in the high-dimensional representations that are not interpretable for humans. Researchers struggle to grasp the analytical relationship between facial images and deep face representations, hindering the understanding of decision-making processes and the crucial attributes of facial images in these decisions. The boundaries of potential errors for deep learning models are also difficult to anticipate. Additionally, interpretability is crucial for ensuring a trustworthy system, particularly in security-sensitive scenarios. For instance, if a face recognition system's predictions are used to identify someone as a criminal, it is imperative to understand why the probe and gallery faces appear similar to prevent false convictions or acquittals.

Therefore, interpreting deep learning-based face recognition models effectively is a prerequisite and vital step in addressing the current bottlenecks in facial recognition tasks. Moreover, it holds promise as a key technology for constructing trustworthy face recognition systems, especially in security-critical scenarios.

We believe that the desired interpreting method for deep face recognition models should possess three key characteristics: Firstly, its mode of interpreting should be comprehensible to humans, as this is a fundamental aspect. Secondly, the explanation should be quantitative for measurement. Finally, the interpreting method should be capable of providing both model and sample-specific explanations. The former allows us to comprehend how the recognition model conceptualizes the notion of "identity", whereas the latter helps us understand which underlying factors are influential during a specific instance of recognition. Regarding sample-specific interpretation, sufficiency and necessity are also requisite qualities for the desired interpretation, indicating that the interpretation sufficiently supports the results and contains no redundancy.

Manuscript received 13 December 2023; revised 25 April 2024 and 20 June 2024; accepted 30 June 2024. Date of publication 5 July 2024; date of current version 22 July 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC3310400; in part by China Postdoctoral Science Foundation under Grant BX20230044 and Grant 2023M730290; in part by the National Natural Science Foundation of China under Grant 62276025, Grant U23B2054, and Grant 62276263; and in part by Shenzhen Technology Plan Program under Grant KQTD20170331093217368. The associate editor coordinating the review of this article and approving it for publication was Prof. Linke Guo. *(Min Ren and Yuhao Zhu contributed equally to this work.) (Corresponding author: Yongzhen Huang.)*

Min Ren and Yongzhen Huang are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100088, China (e-mail: renmin@bnu.edu.cn; huangyongzhen@bnu.edu.cn).

Yuhao Zhu is with the Institute of Computing Technologies, China Academy of Railway Sciences Corporation Ltd., Beijing 100081, China (e-mail: zhuyuhao@rails.cn).

Yunlong Wang and Zhenan Sun are with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yunlong.wang@cripac.ia.ac.cn; znsun@nlpr.ia.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2024.3424291, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2024.3424291

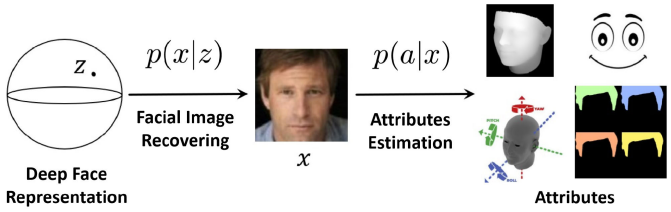

Fig. 1. The proposed two-stage attribute recovery framework. The facial image is recovered in the first stage. The attributes are estimated and the significance of the attributes for the recognition result is quantified.

A multitude of endeavors have been undertaken to interpret models based on deep learning for face recognition. The majority of current approaches strive to interpret the deep neural networks by ascribing the ultimate decisions to local regions or sets of pixels within the input images [4], [5], [6], [7], [8], [9], [10]. This methodology for attribution analysis is quite superficial. For instance, they might assert that the region of the nose holds the utmost significance for the recognition outcome, yet how essential attributes like the shape, texture, or positioning of the nose affects recognition results remains unclear. Some researchers attempt to explain the results by the facial attributes [11]. They create counterfactual samples by altering facial attributes to explore the correlation between facial attributes and recognition outcomes. However, this method is constrained to altering a limited set of facial attributes, and struggles with analyzing crucial attributes such as facial shape. Moreover, it can only offer sample-specific explanations. Research on the interpretability of deep learning models has received widespread attention in recent years. Most studies attempt to attribute model outputs to local regions of input images [12], [13], [14], [15], [16], [17], [18], [19], [20], while some visualize the hidden representations of models to understand their behavior patterns [21], [22]. Recently, some researchers have tried to explore the key factors influencing model outputs through counterfactual samples [11], [23], [24], [25]. Existing methods still have some shortcomings, such as a lack of ability to quantitatively analyze models and difficulty in off-the-shelf applications to face recognition models.

In order to provide a better solution, we propose a two-stage framework to interpret the deep learning based face recognition models via facial attributes, as shown in Fig. 1. In the initial stage, the aim is to recover all the attributes of the input facial image from the deep face representation as authentic as possible, i.e. to recover the input facial image itself. In the second stage, the attributes are estimated according to the recovered image and the significance of the attributes for the recognition result is quantified. In the process of feature extraction in facial recognition models, facial images are mapped to deep face representations, which can be seen as compressing and encoding information of facial images. The facial attribute information relevant to recognition tasks is encoded into deep face representations, while irrelevant information is discarded. Therefore, the key to understanding facial recognition models lies in discovering which facial attribute information is encoded in deep face representations and to what extent. Thus, performing attribute recovery based on deep face representations can be viewed as a reverse decoding process, where facial attributes encoded in deep face representations are revealed and quantitatively measured to understand deep face representations. This two-step framework is highly scalable, enabling the estimation of ***any*** desired facial attribute from the recovered images. Furthermore, this framework maximizes the use of existing facial attribute estimation techniques, thereby bypassing the arduous task of independently obtaining a recovery model for each specific facial attribute.

Within this framework, the facial attributes embedded in the deep face representation are visually manifested in the recovered image, for human-friendly comprehension. To quantitatively evaluate the significance of facial attributes for the recognition model as a model-specific explanation, we introduce a method based on mutual information to gauge the informative content of the attributes within the deep face representation.

To achieve sample-specific explanation, the methodology of counterfactual explanation is employed. Existing methods for constructing counterfactual samples typically alter the attributes of interest in the original sample to assess their significance. However, such methods can only examine a single attribute at a time and fail to explore the combined effects of multiple attributes on the output. Therefore, we employ adversarial examples as a means to generate adaptive counterfactual explanations. Adversarial examples entail slight modifications to the facial image that alter the prediction of the recognition model. This approach adaptively modifies the original sample, effectively overcoming the limitations of existing methods. The reason adversarial examples have not been used as counterfactual samples in existing research is that humans cannot comprehend the changes induced by adversarial examples to the deep face representation. However, by utilizing the proposed framework, the alterations in deep face representation prompted by adversarial examples can be visualized and quantitatively interpreted in terms of facial attributes.

Therefore, the proposed approach possesses the three characteristics that an ideal explanatory approach should have.

The main contributions of this paper can be summarized as follows:

- This paper proposes the use of facial attributes to understand deep face representations and introduces a framework for quantitatively measuring the relationship between facial attributes and deep face representations using mutual information.
- The proposed framework allows a black-box deep learning model to "communicate" with humans through perceptible attributes, thereby providing an interface that is intelligible to human understanding. It may bring about a paradigm shift in the comprehension of deep face recognition models.
- The proposed framework facilitates a profound comprehension of how the recognition model "comprehends" the concept of identity, thus granting us a deeper understanding of its limitations and boundaries. This understanding can guide improvements in recognition models.

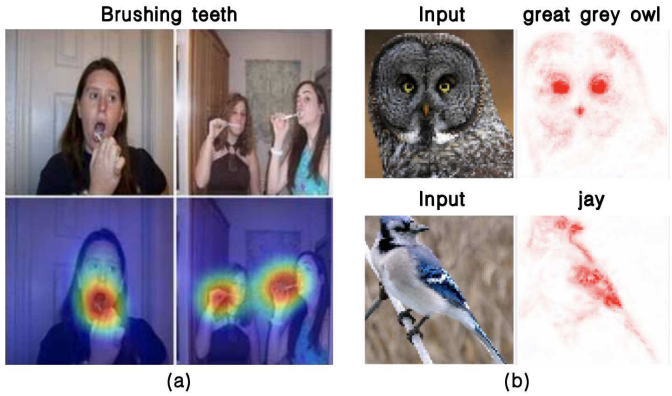

Fig. 2. Some methods attempt to establish the correlation between predictions and local regions or pixels in images. (a): CAM [15]; (b):CoDA-Nets [20].

II. Related Work

A. Efforts to Interpret Deep Learning Models

In the computer vision and machine learning communities, extensive exploration of various methods is underway to enhance the credibility and interpretability of deep learning models [26]. The goal is to achieve a heightened understanding of the fundamental processes involved in recognition.

One straightforward approach to comprehending the deep learning model revolves around the visualization of the features acquired and internalized by the model. This technique allows for a more instinctive comprehension of the acquired representations. In the nascent stages, the adoption of deconvolution networks [27] was prevalent in visualizing the artificial neurons [21], [22]. Subsequently, a number of researchers embarked on the endeavor of visualizing deep features through optimization techniques [28], [29], [30]. These approach involved iteratively modifying the input image to maximize the response of specific neurons or to elicit desired features. Through this optimization process, the hidden representations within the neural network could be unveiled.

Establishing the correlation between predictions and local regions or pixels in images has emerged as another prominent avenue of research [12], [13], [14], [15], [16], [17], [18], [19], [20], as shown in Fig. 2. This line of inquiry aims to uncover the relationship between the model's output and specific regions of the input image. By analyzing the impact of local regions or pixels on the final prediction, researchers gain insights into the decision-making process of the model and can provide more interpretable explanations for its outputs.

Moreover, there has been a research focus on establishing the correlation between predictions and the individual neurons within deep models. This analysis allows for insights into the neural activity patterns that drive the model's predictions and provides valuable insights into the features and representations learned by the network. Shrikumar et al. [31] propose a method to assign contribution scores to neurons. Brendel and Bethge [32] propose a method that combines the strategy of Bag of Feature (BoF) [33], which represents images as histograms of visual words, with deep learning to improve the interpretability. This fusion of methods allows for improved interpretability by providing insights into the visual words that contribute to the model's predictions.

Recently, counterfactual explanations have been gaining significant attention [11], [23], [24], [25]. Counterfactual explanations aim to provide insights into the causal relationship between input variables and the model's output by considering alternative scenarios. They attempt to answer questions such as "What changes to the input would have resulted in a different output?"

Most existing methods for interpreting deep learning models are only aimed at understandability. Such methods often lack an in-depth analysis of deep learning models, and merely encompass coarse observations of these models. Only a few methods attempt quantitative analysis, and to our knowledge, there are scarcely any methods that can provide both model-specific and sample-specific quantitative explanations.

B. Interpretability of Face Recognition

The methodology of interpreting general classification models has been extended to the field of face recognition. Earlier work devote their attention to attributing decisions to specific local regions in face images [4], [5], [6], [7], [8], [9]. This kind of approach aims to understand which specific local region contribute the most to the model's decision-making process. Attributing decisions to local regions in face images provides a convenient way to visualize the focuses of face recognition models. Nevertheless, the explanations provided by these methods are rather vague and incomplete. For example, they may claim that the region of the nose is the key region that leads to the prediction, but they cannot tell us whether the shape of the nose, the skin color, or the position of the nose is the decisive feature of the prediction.

Additionally, a handful of methodologies strive to examine the effects of facial attributes. Dhar et al. [34] analyze four facial attributes in multi-layer neural networks. Adudarham et al. [35] investigate the impacts of facial features that are used by humans on the deep face recognition model. These efforts to analyze the impacts of facial attributes in face recognition systems are valuable and promising. However, these methods remain significantly flawed, possessing a constrained ability to analyze facial attributes. For example, they struggle with assessing critical attributes like facial shape, and also exhibit a lack of expansibility.

More recently, uncertainty estimation has emerged as a research area for interpreting and improving the performance of the face recognition models [26], [36], [37]. These methods represent the input image as a distribution rather than a single point in the facial representation space. This distribution provides a measure of uncertainty, where the variance of the distribution reflects the uncertainty associated with the corresponding features. By incorporating uncertainty estimation, researchers can gain insights into the reliability and confidence of the model's predictions. Nevertheless, this kind of approach is mathematically sound, but remains incomprehensible to human comprehension. They do not provide explanations of deep face representations. Instead, their explanations are based on the interpretability of deep face representations.

III. Method

This section introduce the proposed framework to interpret the deep learning based face recognition models in terms of the attributes of facial images. This section first provide the notation used, followed by an introduction of the attribute recovery framework that translates deep face representation into facial attributes. Then, the method that quantifies the significance of attributes in determining recognition outcomes is introduced.

A. Notation

In the realm of facial recognition, the symbol x is assigned to represent a facial image. The deep face recognition model maps x to a deep face representation:

$$z = F(x) \tag{1}$$

where $F(\cdot)$ is the deep face recognition model, z is the deep face representation. The attributes of a facial image x are denoted by a set: $\{a_i\}$, such as shape, expression, head pose, etc. Each element of $\{a_i\}$ is an attribute of the facial image, it can be estimated as follows:

$$a = E(x) \tag{2}$$

where $E(\cdot)$ is the attribute estimator.

A facial image x, its deep face representation z, and one of its attributes a can be modeled by a Bayesian network:

$$a \longleftarrow x \longrightarrow z \tag{3}$$

Their joint probability distribution can be calculated by the product of the prior distribution of x and conditional distributions:

$$p(a, x, z) = p(x)p(a|x)p(z|x) \tag{4}$$

The conditional distribution $p(z|x)$ is parameterized by the deep learning model $F(\cdot)$, $p(z|x)$ can be rewritten as $p_F(z|x; \theta_F)$, where θ_F is the parameter of $F(\cdot)$.

B. Attributes Recovering

To estimate the relationship between the intricate deep face representation and the facial attributes, we propose an attribute recovery framework, as shown in Fig. 3. In order to recover the facial attributes to their utmost authenticity from the deep face representation, the proposed framework encompasses four fundamental components: the deep face recognition model, the latent space transformer, the facial image decoder, and the attribute estimator. The deep face recognition model extracts deep face representations from the input images. The facial image decoder is utilized to recover facial images. Due to the misalignment between the deep face representation space of the recognition model and the encoding space of the facial image decoder, it is not feasible to directly feed deep face representations into the facial image decoder. Therefore, the latent space transformer is employed to perform space transformations from the deep face representation space to the encoding space of the facial image decoder. Upon the recovery of the facial image by the facial image decoder, the attribute estimator then estimate the facial attributes, thereby culminating the process of attribute recovery.

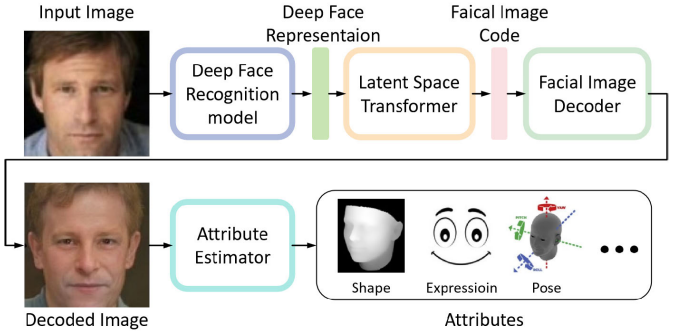

Fig. 3. The facial attribute recovery framework encompasses four fundamental components: the deep face recognition model, the latent space transformer, the facial image decoder, and the attribute estimator. The deep face representation extracted by the deep face recognition model undergoes a decoding process in order to faithfully recover the original facial image. Subsequently, the facial attributes are estimated according to the recovered facial image.

During the training process of the proposed framework, the facial image decoder and the latent space transformer are trained separately. The facial image decoder is trained first, aiming to obtain a model with strong facial image generation capabilities, thereby ensuring the framework's ability to recover facial attributes. Once the facial image decoder is obtained, the latent space transformer is trained to enable it to perform the transformation from the deep face representation space to the encoding space of the facial image decoder.

1) Facial Image Decoder: To accomplish attribute recovery, it is essential for the facial image decoder to exhibit exceptional capabilities in generating facial images. Particularly in the analysis of attributes characterized by finer details, the facial image decoder must be adept at accurately representing these attributes within the pixel space. Should the facial image decoder demonstrate subpar performance in attribute generation, it would significantly impede the overall efficacy of the proposed framework.

Therefore, we employ a StyleGAN2 [38] based decoder as the facial image decoder. StyleGAN2 is a generative model known for their exceptional generation capabilities, enabling them to effectively capture fine details of facial images while exhibiting strong diversity in attributes [11]. It is particularly suitable for the attribute recovery.

2) Latent Space Transformer: The latent space transformer, which transfers the deep face representations to the encoding space of facial image decoder, is a multilayer perceptron (MLP). Specifically, it consists of 8 fully connected layers to realize the transformation. Since the facial image code of StyleGAN2 consists of multiple sub-codes, the last four layers of the latent space transformer are divided into n separated branches, where n is the number of sub-codes. Each branch generates one sub-code. The structure of the latent space transformer is illustrated in the Fig. 5. The latent space transformer does not directly predict the latent code required for the facial image decoder. Instead, it predicts the residual between the latent code and the sampling average of latent

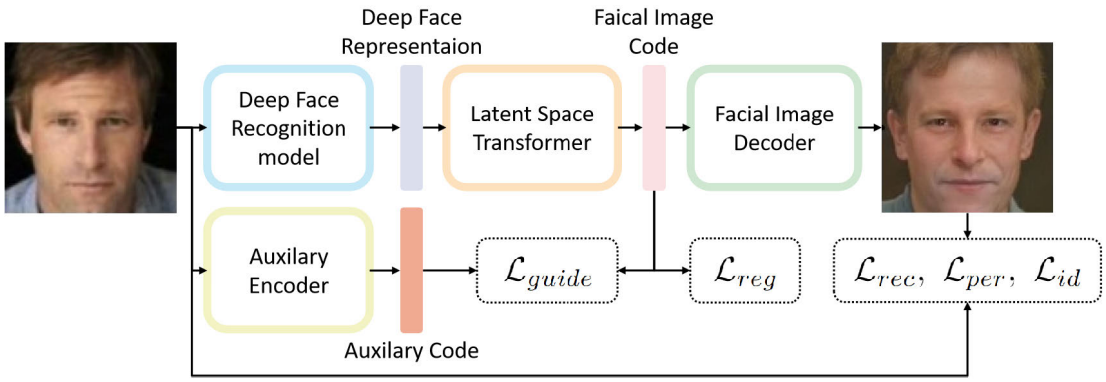

Fig. 4. The training of the latent space transformer.

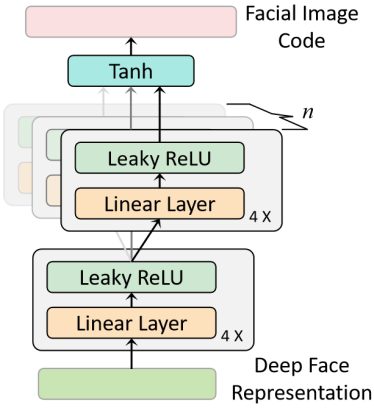

Fig. 5. The structure of the latent space transformer. Since the facial image code of StyleGAN2 consists of multiple sub-codes, the last four layers of the latent space transformer are divided into n separated branches, where n is the number of sub-codes.

codes. The sampling average of latent codes refers to the mean value of latent codes sampled from the encoding space of the facial image decoder:

$$l = l_{res} + l_{mean} \tag{5}$$

where l_{res} is the output of the latent space transformer, l_{mean} is the sampling average of latent codes, l is the latent code fed into the facial image decoder. By employing residual learning, the latent space transformer can converge faster and achieve greater stability during the training process.

The training objective of the latent space transformer encompasses five components. The first component of the training objective for the latent space transformer is the pixel-wise reconstruction loss:

$$\mathcal{L}_{rec} = ||x - x_{re}||_2 \tag{6}$$

where x is the original facial image, x_{re} is the recovered image by the facial image decoder. This loss function directly constrains the distance between the recovered facial images and the original images in the pixel space.

Next, there are two loss functions that constrain the differences between them in the feature domain:

$$\mathcal{L}_{per} = ||P(x) - P(x_{re})||_2 \tag{7}$$

$$\mathcal{L}_{id} = 1 - cosine(F(x), F(x_{re})) \tag{8}$$

where $\mathcal{L}_{per}$ is Learned Perceptual Image Patch Similarity (LPIPS) loss [39], $P(\cdot)$ denotes the perceptual feature extractor, $F(\cdot)$ denotes the deep face recognition model, $cosine(\cdot, \cdot)$ denotes cosine similarity of two face representations. The former emphasizes the preservation of the style of the original image, while the latter enforces the identity of the recovered image.

Moreover, it is beneficial to provide direct supervision on the output of the latent space transformer in the encoding space of the facial image decoder. To this end, an auxiliary encoder is introduced to the training process. This auxiliary encoder can be regarded as the reflection of the facial image decoder, mapping from the pixel space of the face to the encoding space of the facial image decoder. It takes the original facial images as input and produces latent codes in the encoding space as output, thus providing direct supervision for the latent code transformer in the encoding space of the facial image decoder:

$$\mathcal{L}_{guide} = ||l_{res} + l_{mean} - E_{au}(x)||_2 \tag{9}$$

where $E_{au}(\cdot)$ is the auxiliary encoder.

Besides, L_2 norm of l_{res} is adopted as a regularization term to improve the generalization and the stability of training.

$$\mathcal{L}_{reg} = ||l_{res}||_2 \tag{10}$$

The final objective function for the training of the latent space transformer is:

$$\mathcal{L} = \beta_1\mathcal{L}_{rec} + \beta_2\mathcal{L}_{per} + \beta_3\mathcal{L}_{id} + \beta_4\mathcal{L}_{guide} + \beta_5\mathcal{L}_{reg} \tag{11}$$

where β_1, β_2, β_3, β_4, and β_5 are the weights of losses.

3) Facial Attribute Estimation: Following the recovery of the facial image, it becomes viable to estimate the facial attributes based on the recovered image. It is important to note that the spectrum of potential attributes is limitless, encompassing any discernible attribute from the recovered facial image. One can examine common facial attributes such as facial shape, expression, age, and so forth, as demonstrated in Section IV-B, and also customize facial attributes, for instance, the ratio of nose width to mouth width, as demonstrated in Section IV-C. In the process of facial attribute estimation, it is feasible to utilize the existing attribute estimation techniques, and the proposed framework constitutes a highly scalable framework.

By following the aforementioned procedure, the recovering process from deep face representation to facial attributes has been accomplished. This recovering process establishes the groundwork for interpreting the recognition outcomes of deep face recognition models based on attributes.

C. Model-Specific Information Measurement

The model-specific interpretation focuses on providing an explanation for the recognition model, specifically by quantitatively assessing the significance of attributes for the recognition model. Conversely, the sample-specific interpretation places emphasis on individual samples, evaluating the importance of the attributes of a specific sample in determining its recognition outcome. This subsection will first introduce the quantitative interpretation methods for model-specific interpretation.

The mutual information between a certain attribute and the deep face representation is used to quantitatively measure the importance of the attribute for the deep recognition model:

$$\mathcal{I}(a; z) \tag{12}$$

The larger the mutual information between a specific attribute and the deep face representation, the more importance this attribute occupies in the deep face representation. However, calculating $\mathcal{I}(a; z)$ is challenging due to the high-dimensional and continuous nature of the deep face representation z, which usually follows a complex distribution.

It can be proven that $\mathcal{I}(a; \hat{a})$ is a tight lower bound of $\mathcal{I}(a; z)$:

$$\mathcal{I}(a; z) \geqslant \mathcal{I}(a; \hat{a}) \tag{13}$$

where $\hat{a}$ is the recovered attribute from z by the proposed framework (see the proof in the appendix). Therefore, $\mathcal{I}(a; \hat{a})$ is adopted to estimate $\mathcal{I}(a; z)$.

The gap between $\mathcal{I}(a; z)$ and $\mathcal{I}(a; \hat{a})$ is related to the recovering capability of the proposed framework and specifically, to the generation capability of the facial image decoder. The stronger the generation capability of the facial image decoder, the stronger the recovering capability of the proposed framework, resulting in a smaller gap (see the proof in the appendix). This is also empirically supported by the experiments conducted in this paper.

Generally, $\mathcal{I}(a; \hat{a})$ can be easily calculated or estimated. The mutual information of a and $\hat{a}$ can be calculated as follows:

$$\mathcal{I}(a; \hat{a}) = \mathcal{H}(a) + \mathcal{H}(\hat{a}) - \mathcal{H}(a, \hat{a}) \tag{14}$$

where $\mathcal{H}(\cdot)$ denotes the information entropy.

D. Sample-Specific Explanation

In this subsection, we introduce a sample-specific explanation method that determines the importance of attributes in the recognition result of a particular sample. A counterfactual explanation approach that utilizes adversarial examples as counterfactual samples is proposed. The concept of counterfactual explanation involves manipulating facial images to create counterfactual samples. If altering a certain attribute leads to a change in the recognition result, it indicates the importance of that attribute for the recognition result.

During face recognition, the similarity between a facial image x and the reference image x_{ref} is measured through the deep recognition model:

$$S(F(x), F(x_{ref})) \tag{15}$$

where x_{ref} denotes the reference image, $F(\cdot)$ denotes the deep recognition model, $S(\cdot, \cdot)$ denotes the similarity metric function. The recognition result depends on whether this similarity is larger than the pre-defined threshold. The adversarial examples are adopted to adaptively generate counterfactual samples:

$$x_{adv} = A(x, x_{ref}; F) \tag{16}$$

where A denotes the adversarial attack method, it aims to change the recognition result with minimum modifications, x_{adv} denotes the adversarial example. Therefore, x_{adv} serves as a counterfactual sample and modifies the deep face representation of x. This modification in the deep face representation is not only sufficient, as it leads to a changed recognition result, and also almost necessary, as it is crafted with minimal alterations. Utilizing adversarial examples offers an adaptive approach to generate counterfactual samples that are both sufficient and necessary.

The proposed framework allows us to comprehend the alterations induced by adversarial examples in terms of attributes. By comparing the attribute recovered from the deep face representation of x with that recovered from the deep face representation of x_{adv}, we can quantitatively measure the importance of the attribute for the recognition result:

$$s_a = \frac{||\hat{a} - \hat{a}_{adv}||}{||\hat{a} - \hat{a}_{ref}||} \tag{17}$$

where $\hat{a}$, $\hat{a}_{ref}$, and $\hat{a}_{adv}$ are attributes recovered from the deep face representation of x, x_{ref}, and x_{adv} respectively, s_a is the significance of the attribute to the recognition result. The numerator measures the change caused by the counterfactual sample. The denominator is a normalization factor of the significance of the attribute.

If $s_a > 1$, indicating that the alteration caused by the counterfactual sample exceeds the disparity between x and x_{ref}, the attribute is considered significant. If $s_a < 1$, however, the attribute is regarded as non-significant. The quantitative measurement allows us to assess the significance of each attribute to the recognition outcome. In this way, a quantitative sample-specific explanation is obtained.

IV. Experiments

This section conducts experiments to validate the proposed method in interpreting the deep facial recognition models. Two aspects are primarily examined. The first aspect is model-specific interpretation, which quantitatively measures the importance of attributes for the recognition model. The second aspect is sample-specific interpretation, which assesses the importance of attributes for the recognition outcome of a particular sample. Lastly, ablation studies are carried out on the proposed method.

Fig. 6. Examples of the original facial images and the corresponding recovered facial images. The first row displays the original facial images, while the second row showcases the recovered facial images.

A. Implementation Details

The facial image decoder based on StyleGAN2 was trained on the Flickr-Faces-HQ (FFHQ) dataset [40] for 550,000 iterations. We followed the configuration specified by Tero et al. [38] during the training process, with the exception of the generated image resolution. Instead of aiming for a resolution of 1024×1024, we opted for 128×128. We made this choice because a resolution of 128^2 is deemed sufficient to meet the requirements for face recognition tasks. ArcFace (R50) [3] is employed as the face recognition model, and the training dataset used is consistent with [3].

The latent space transformer is a Multi-Layer Perceptron (MLP) consisting of 8 fully connected layers. It takes a 512-dimensional input, which is the deep face representation, and generates a latent code that is compatible with StyleGAN2's $\mathcal{W}+$ space. The auxiliary encoder consists of a ResNet50 backbone, a feature pyramid structure based on FPN for feature refinement, and a set of non-linear mapping networks to predict the latent code. It was also trained on the FFHQ dataset. For more detailed information, please refer to Zhu et al. [41]. The weights in Eq. 11 are set as $\beta_1 = 0.1$, $\beta_2 = 0.8$, $\beta_3 = 8$, $\beta_4 = 0.1$, and $\beta_5 = 0.02$. The latent space transformer was trained on the Large-scale CelebFaces Attributes (CelebA) Dataset [42], with a batch size of 16. The initial learning rate was set to 0.01, and the Adam optimizer was used for parameter updates, and the model was trained for 20 epochs.

B. Model-Specific Deep Face Representation Parsing

The model-specific explanation focuses not on specific facial samples, but rather on understanding the recognition model itself in terms of the importance of facial attributes in its decision-making process. In this experiment, nine facial attributes are examined, which can be categorized into three groups: The first group comprises shape, expression, and head pose, which depict the spatial structure of faces. The second group encompasses skin color, hair color, and age, which describe the texture of faces. The final group of attributes characterizes facial accessories, including glasses, hats, and earrings. By considering these three groups of attributes, we are able to provide a comprehensive explanation of the facial recognition model from an attribute perspective. It is worth noting that the proposed methodology is not limited to these attributes. Any other facial attribute can be subjected to the same analysis. Our selection is merely a choice of the most illustrative attributes for demonstration purposes.

To assess the face shape, we employ 3DDFA [43], [44], [45], a model that predicts the 3D morphable model (3DMM) [46] of faces, as the attribute estimator. This enables us to represent the face shape by the reconstruction coefficients of the face shape principal components. For the estimation of expression, head pose, and age, we rely on a state-of-the-art commercial face analysis API.[1] Color moments are utilized to capture the color attributes of skin and hair. For further details on the Color moments, please refer to the appendix. As for accessories, we adopt a binary attribute to denote whether they are worn or not. Details on the calculation of mutual information can be found in the appendix. As for the test dataset used for analysis, we utilize LFW [47], which comprises 133,233 facial images belonging to 5,749 identities.

1) Qualitative Analysis: In order to gain an intuitive and qualitative understanding of the facial attributes contained in deep face representations, we first visualize the recovered images, as shown in Fig. 6. More recovered images can be found in the appendix. From the figure, it can be observed that the recovered images generally maintain the shape of the original faces, including the overall facial proportions, jawline contours, sizes and shapes of facial features, as well as the relative layouts of the features. This indicates that deep facial representations effectively encapsulate the shape information of the faces. Even in cases where self-occlusion occurs due to head poses, as seen in the third sample from the left, the recovered image still manages to reasonably restore its shape. This suggests that deep recognition models are capable of inferring and completing the shape of a face, even under challenging head pose conditions, which contributes to their robustness in face recognition.

On the other hand, the original expressions and head poses are not well preserved in the recovered facial images. For instance, in the first and second samples from the left, there are distinctions in expressions, but the recovered images exhibit similar expressions. Similarly, although the third sample from the left has a different head pose compared to the others, the recovered facial image shows minimal variation in head pose. This suggests that the deep recognition model, during the feature extraction process, essentially abandons information pertaining to facial expressions and poses.

[1] Face++ Research Toolkit: https://www.faceplusplus.com.cn/.

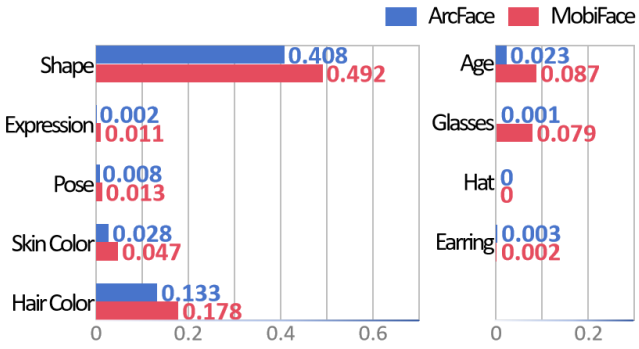

Fig. 7. Quantitative comparative analysis of ArcFace and MobiFace.

The hair color of the input facial image is generally preserved in the recovered image, yet it is also subject to interference from background colors or the color of hats, as seen in the fourth image from the left and the second from the right. In the case of the fourth sample from the right, the age is evidently older than the other samples. However, this age gap is not effectively reflected in the recovered images, indicating that the recognition model does not retain age-related information.

Furthermore, the three samples from the right wear accessories such as glasses, a hat, and an earring. However, these accessories are not reflected in the recovered facial images, indicating that the recognition model selectively gives up the accessory-related information.

2) Quantitative Analysis: Furthermore, we utilized the proposed mutual information based quantitative interpretation method to explore these nine attributes. The results are shown in Fig. 7. The quantitative interpretation confirms and validates the observations we made through the recovered images. Through this quantitative analysis, we can draw the following conclusions:

- The quantitative analysis reveals that the attributes contained in deep facial representations are not balanced. The importance of shape significantly outweighs the other attributes, with hair color ranking second, but significantly lower than shape. This suggests that the facial recognition model places a strong emphasis on extracting shape information during the feature extraction process. Additionally, it considers a partial amount of information related to hair color, while neglecting the other facial attributes during the recognition process.
- This quantitative analysis method allows us to compare the judgments of deep facial recognition models with those of humans in understanding facial recognition. This enables us to fundamentally assess the rationality of the recognition models. For instance, it is reasonable that attributes like expressions, head pose, and accessories are considered less important in most applications since they are not inherently tied to an individual's identity from a human perspective. However, the importance of some attributes like hair color may be debatable since it can be changed for the same person.
- Through this quantitative analysis, we can acquire a lucid comprehension of how the recognition model "comprehends" the notion of identity. Analyzing the model's "perception" of identity can provide valuable insights for various facial-related research fields. For instance, in tasks like face editing and talking face, which entail preserving or decoupling facial identity, researchers often resort to complex models and loss functions to constrain these models, particularly when the specific connotations of "identity" are not fully understood. [48], [49], [50]. Understanding the dominant attributes of identity can make such operations more convenient and targeted.

3) Differences Between Recognition Models: What differentiates one facial recognition model from another? Do different recognition models focus on different facial attributes? These questions have been difficult to answer explicitly, but the proposed method offers a possibility to address these questions.

The MobiFace [51], as a lightweight facial recognition model, has been used for comparative analysis with ArcFace, and the training dataset of MobiFace is consistent with ArcFace. They are different in terms of model structure and training objective. The quantitative analysis results are depicted in Fig. 7.

The quantitative comparison between ArcFace and MobiFace provides a clear understanding of the similarities and differences between them:

- The overall distribution of attributes that MobiFace and ArcFace focus on during the feature extraction process is similar. Both methods prioritize facial shape and hair color, while ignoring expressions and head pose.
- The differences between the two models are also evident. Overall, deep face representations of MobiFace contain more information compared to ArcFace, indicating that ArcFace applies a stricter information refinement process during feature extraction.
- There are significant differences between the two models in terms of age and glasses attributes. MobiFace incorporates more information about these two attributes compared to ArcFace, which is also evident in the recovered facial images, as shown in Fig 8. This may result in more noise interference for MobiFace when dealing with glasses disturbances or performing cross-age recognition.

In order to further validate the discrepancies observed between ArcFace and MobiFace, both models were subjected to testing on two datasets, namely LFW and AgeDB-30 [52]. The LFW dataset functioned as a benchmark for appraising the general face recognition capabilities of the models. On the other hand, AgeDB-30 deliberately introduced age disparities within its intra-class sample pairs, aiming to assess the models' ability for cross-age facial recognition. The experimental results are summarized in Table I.

The experimental results reveal that the performance disparity between ArcFace and MobiFace on the LFW dataset is minimal, with a mere 0.17% difference. However, on the AgeDB-30 dataset, the performance gap between the two models reaches 1.77%, which is ten times more pronounced than

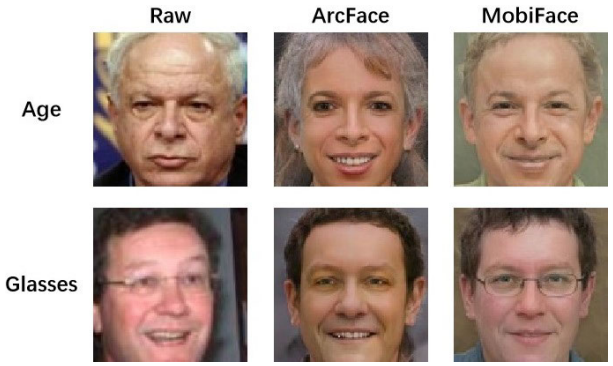

Fig. 8. Recovered facial images by ArcFace and MobiFace. In the left column are the original images. In the middle column are the images recovered from the deep face representations of ArcFace. In the right column are the images recovered from the deep face representations of MobiFace. Notably, the glasses are exhibited in the recovered images of MobiFace, and more details about age are encompassesed.

TABLE I

VALIDATION EXPERIMENT ON AGE ATTRIBUTE. THE PERFORMANCE GAP BETWEEN THE TWO MODELS ON AGEDB-30 IS TEN TIMES MORE PRONOUNCED THAN THAT ON THE LFW DATASET. THIS INDICATES THAT MOBIFACE TENDS TO RETAIN A GREATER AMOUNT OF AGE-RELATED INFORMATION DURING THE FEATURE EXTRACTION PROCESS

	ArcFace [3]	MobiFace [51]	Diff.
LFW [47]	99.73%	99.56%	**0.17%**
AgeDB-30 [52]	98.02%	96.23%	**1.77%**

that on the LFW dataset. This indicates that the introduction of age disparities within intra-class sample pairs significantly impacts the facial feature representation of MobiFace, compared to ArcFace. In other words, MobiFace tends to retain a greater amount of age-related information during the feature extraction process. These findings validate the quantitative analysis results of the proposed approach regarding the age attribute of these two models.

C. Sample-Specific Explanations

In contrast to model-specific explanations, sample-specific explanations focus on individual samples. This means that when recognizing a specific sample, sample-specific explanations assess the impact of its attributes on the recognition result.

1) Qualitative Explanations: By recovering the recognized faces and their corresponding reference faces, we can visually observe the similarities and differences as perceived by the recognition model. To demonstrate this, we have selected and analyzed two pairs of face samples that are incorrectly recognized by the model, as shown in Fig9.

Both pairs in the image are positive pairs, but the model assigned a low similarity score, resulting in misrecognition. Recovering the deep face representations reveals that for the pair on the left, due to the obstruction caused by glasses, the model "perceives" differences in the eyes and nose of the two individuals, while the degree of nasolabial folds also appears distinct. In the case of the pair on the right, the model's misrecognition is attributed to a difference in the

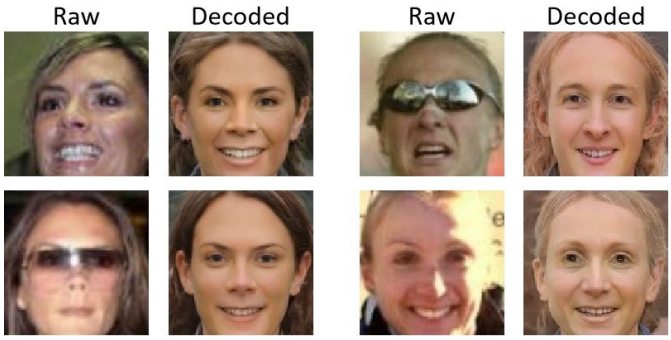

Fig. 9. Two pairs of face images that are incorrectly recognized by ArcFace. By recovering the recognized faces and their corresponding reference faces, we can visually observe the similarities and differences as perceived by the recognition model.

overall aspect ratio that it "perceives" between the two faces. Additionally, the recognition model "thinks" that there are distinct differences in the shape of the eyes and eyebrows, which contributed to the recognition error.

The aforementioned analysis provides an illustrative example of offering sample-specific, qualitative explanations for the recognition results. **When the model makes an error in recognizing a particular sample, we are no longer at loose ends, but can clearly indicate where the problem lies**. This enables us to gain a deeper understanding of the model's limitations and boundaries, while assisting us in identifying targeted improvements.

2) Quantitative Explanations: The proposed sample-specific quantitative metrics offers a more detailed and comprehensive analysis of the recognition results. For a facial image and its reference image, as shown in Fig 10, an adversarial sample is generated as the counterfactual samples for quantitative explanations. PGD [53], which is a state-of-the-art adversarial attacking method, is adopted to generate the adversarial sample. In this case, the adversarial sample alters the deep face representation with minimal modifications, resulting in an increased distance in the deep face representation space compared to the reference image, thereby altering the recognition result. Such modifications are initially challenging for humans to comprehend, but the proposed framework not only visually presents these changes in an intuitive manner but also allows for quantitative measurements from an attribute perspective.

Any facial attribute that can be precisely defined can be quantitatively analyzed through the proposed method. To demonstrate the scalability of this method, we have customized 24 attributes related to face shape and color:

- Attributes related to facial shape, such as the ratio of eye width to eyebrow width, the ratio of nose width to mouth width, and others.
- Color attributes encompassing skin, hair, and iris.

For further information on all attributes, please refer to the appendix.

The significance of the shape attributes and color attributes are shown in Fig 11. The 10 most significant attributes to recognize this facial image pair are: ① the ratio of the nose width to the mouth width, ② the ratio of the distance between

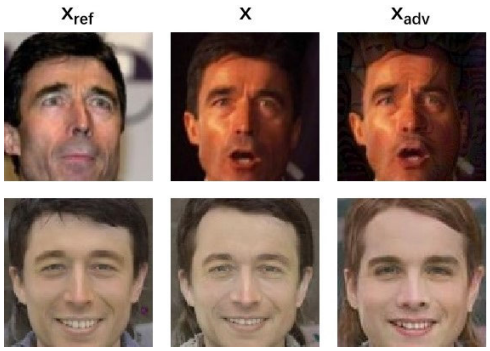

Fig. 10. x denotes the recognized facial image, x_{ref} denotes the reference facial image, x_{adv} denotes the adversarial sample. The corresponding recovered images are displayed in the second row. The adversarial sample alters the deep facial representation with minimal modifications, thereby altering the recognition result. The proposed framework visually presents these modifications.

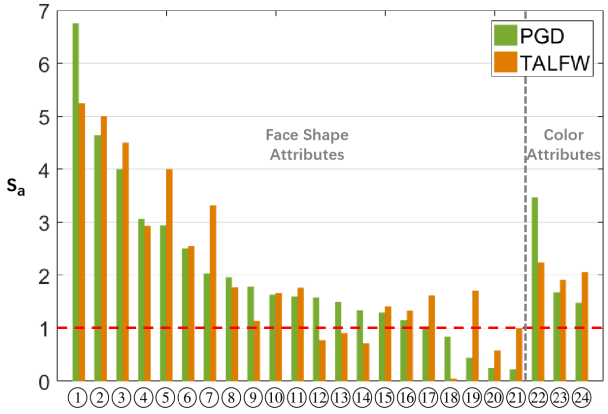

Fig. 11. These attributes serve as quantitative, image-specific explanations for the outcome. The proposed methodology ensures consistent image-specific explanations across various types of adversarial examples. For a comprehensive definition of the attributes, please refer to the appendix.

two eyes to the mouth width, ③ the ratio of the distance between the left eye and face edge to the mouth width, ㉒ the color of hair, ④ the ratio of the nose length to the distance between the nose and underjaw, ⑤ the ratio of the distance between the right eye and face edge to the mouth width, ⑥ the ratio of the eyebrow width to the nose width, ⑦ the ratio of the distance between mouth and underjaw to the distance between eyebrow and mouth, ⑧ the ratio of the distance between eye and nose to the distance between nose and underjaw, ⑨ the ratio of the distance between two eyes to the face width. The definitions of all attributes can be found in appendix.

Furthermore, it is even feasible to analyze the local texture of the face as a unique form of "attribute". The facial images are partitioned into patches of size 7×7, and the LBP feature is utilized to depict the local texture. The magnitudes of importance pertaining to the local texture are depicted in Fig. 12. Among these, the highest significance measures up to 2.57. The visualized figure illustrate that the texture found along the facial perimeter and the mouth region play a more crucial role in recognizing this particular face image pair compared to other regions.

The aforementioned sample-specific explanation based on counterfactual samples serves as an exemplar, verifying that

Fig. 12. The visualization presents the significance of local texture in relation to the recognition outcome. The left is the significance generated by PGD. The right is the significance generated by TALFW. In both scenarios, it is evident that the texture encompassing the facial perimeter and the mouth region assumes a notably more pivotal role in the recognition of this particular pair of facial images.

the proposed method can provide comprehensive explanations of a specific recognition result from the perspective of attributes. **This approach signifies a paradigm shift in our understanding of deep face recognition models, allowing a complex, parameter-rich model obtained through gradient backpropagation to "communicate" with us through attributes, which are an interface that humans can comprehend**.

D. Ablation Study

This subsection presents the ablation study conducted on the proposed method, examining it from three perspectives: Firstly, the effects of different adversarial attack methods employed during sample-specific explanations are explored. Secondly, an experimental analysis is carried out to investigate the impact of variations in the facial image decoder's capabilities. Finally, the role of the auxiliary encoder is demonstrated through experiments.

1) Impact of Different Kinds of Adversarial Examples: Adversarial examples offer a versatile approach to generating counterfactual samples. Nonetheless, it is worth considering whether different adversarial examples yield distinct explanations. To address this inquiry, we utilize TALFW [54] as a counterpart to PGD. TALFW is a transfer-based black-box adversarial attack, which significantly differs from PGD in both its attack principle and implementation methods.

The comparative analysis of sample-specific quantitative analysis generated by PGD and TALFW is presented in Fig. 11 and Fig. 12. From the illustration, it is evident that the significances of the two kinds of adversarial examples exhibits striking similarities. Among the top 10 most significant shape and color attributes, 9 of them are shared. Furthermore, in terms of texture attributes, both kinds of adversarial examples focus on the facial contour and the area around the mouth.

This experimental result indicates that although different types of adversarial examples may introduce certain variations in sample-specific quantitative explanations, they generally maintain consistency for the majority of attributes, particularly those that primarily influence the recognition outcome.

2) Impact of Different Facial Image Decoder: Within the proposed framework, the facial image decoder plays a crucial role as it directly affects the expressive capabilities of facial

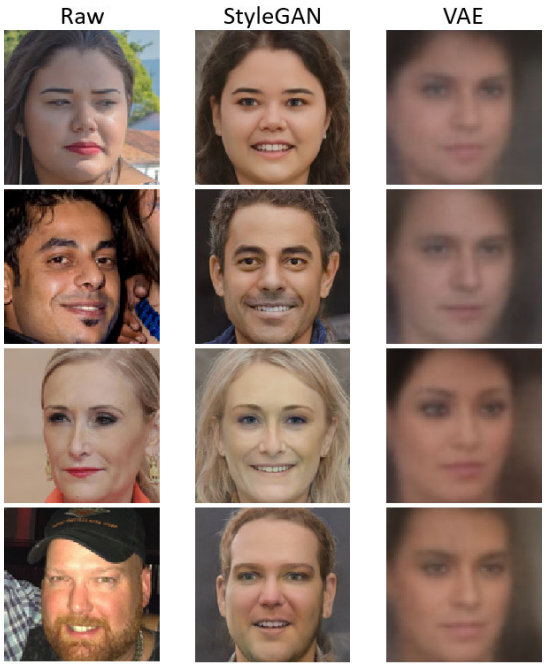

Fig. 13. Examples recovered by StyleGAN2-based and VAE-based facial image decoder. The facial images recovered by the StyleGAN2-based facial image decoder capture a greater amount of information from the original image. In contrast, the facial images recovered by the VAE-based facial image decoder appear blurred and only retain the general facial shape, losing many details.

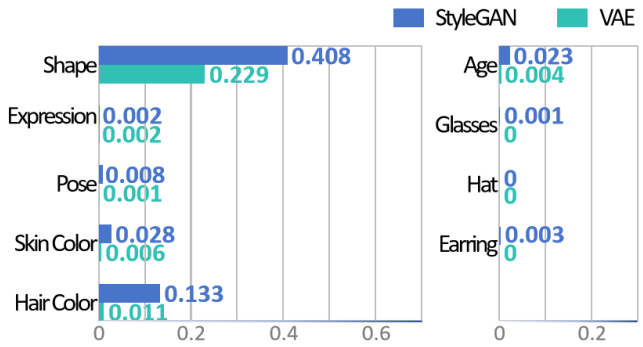

Fig. 14. Quantitative comparative analysis of StyleGAN2-based and VAE-based facial image decoder. The VAE-based facial image decoder struggles to recover attributes beyond facial shape.

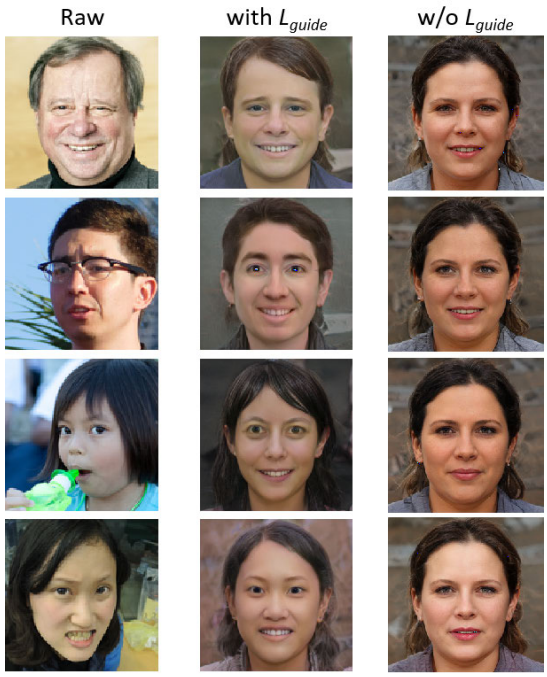

Fig. 15. Ablation study on $\mathcal{L}_{guide}$. The recovering capability of the proposed framework significantly diminishes without the direct supervision of $\mathcal{L}_{guide}$ in the encoding space of the facial image decoder. The recovered images of different faces exhibit only subtle variations, collapsing towards the average face.

attributes and overall performance of the proposed framework. To explore the impact of different facial image decoders, apart from the StyleGAN2-based facial image decoder, we train a VAE-based facial image decoder for comparison purposes. The training dataset used is identical to that of the StyleGAN2-based decoder. Furthermore, the corresponding latent space transformer was also retrained, with the training settings remaining consistent with those described in Sec. IV-A.

The facial image recovering results from both the StyleGAN2-based and VAE-based decoders are illustrated in Fig. 13. From the images, it is evident that the facial images recovered by the StyleGAN2-based facial image decoder capture a greater amount of information from the original image. In contrast, the facial images recovered by the VAE-based facial image decoder appear blurred and only retain the general facial shape, losing many details. Quantitative comparisons further validate this observation, when estimating the mutual information between deep face representation and the nine attributes, as shown in Fig. 14. The mutual information obtained from the VAE-based facial image decoder is significantly lower than that from the StyleGAN2-based facial image decoder. Additionally, the VAE-based facial image decoder struggles to recover attributes beyond facial shape. This is attributed to the proposed framework, which estimates the mutual information between facial attributes and deep face representation using a compact lower boundary. The difference between this lower bound and the actual mutual information, $I(a; z|\hat{a})$, is amplified when the facial image decoder has limited capabilities, resulting in inaccurate estimates. Hence, employing a weaker model does not provide a strong foundation for attribute-based interpretations of face recognition models.

3) Ablation Study on $\mathcal{L}_{guide}$: During the training process of the latent space transformer, in order to provide direct supervision signals in the encoding space of the facial image decoder, we introduce an auxiliary encoder and employ the $\mathcal{L}_{guide}$ loss function for optimization. To evaluate the efficacy of $\mathcal{L}_{guide}$, we conducted an ablation experiment where the model was trained without incorporating $\mathcal{L}_{guide}$.

The recovered facial images obtained from training without $\mathcal{L}_{guide}$ are shown in Fig. 15. From the figure, it is evident that the recovering capability of the proposed framework significantly diminishes without the direct supervision of $\mathcal{L}_{guide}$ in the encoding space of the facial image decoder. The recovered images of different faces exhibit only subtle variations, collapsing towards the average face. Consequently, it becomes

difficult to further interpret and analyze the facial recognition model. This experiment confirms the indispensability of $\mathcal{L}_{guide}$ during the training process of the latent space transformer. Moreover, it also demonstrates the significance of the latent space transformer in terms of its recovering capability within the proposed framework.

V. Conclusion and Future Work

To address the inadequacy of interpretability in face recognition models and the subsequent challenges it poses, this paper proposes the use of facial attributes to understand deep face representations and introduces a framework for quantitatively measuring the relationship between facial attributes and deep face representations. Within this framework, we propose an attribute importance measurement method based on mutual information and incorporate adversarial attack techniques as an effective tool. This enables us to quantitatively interpret deep face recognition models from both a model-specific and sample-specific perspective. It allows a black-box deep learning model to "communicate" with us through perceptible attributes, thereby providing an interface that is intelligible to human understanding. Additionally, the proposed framework facilitates a profound comprehension of how the recognition model "understands" the concept of identity, thus granting us a deeper understanding of its limitations and boundaries. The proposed framework may bring about a paradigm shift in the comprehension of deep face recognition models.

The proposed framework needs to be trained for diverse face recognition models to achieve facial attribute recovery. This requirement arises from the diverse distributions of deep facial representations extracted by different recognition models. These differences necessitate distinct parameters for the Latent Space Transformer. To enhance the practical applicability of the proposed framework, future efforts should concentrate on improving its generalizability across different face recognition models. This would enable effective facial attribute recovery and interpretation across various models. To achieve this, it is first necessary to investigate the deep face representation spaces of different recognition models. If the deep face representation spaces of various models are isomorphic or have similar structures, targeted transformations of these spaces may suffice. However, if there are significant differences between the deep face representation spaces of different models, methodologies from the field of domain generalization, such as domain-invariant representation learning and meta-learning strategies, could be employed. These approaches aim to increase the adaptability to diverse distributions of deep facial representations, thereby transforming the proposed framework into a plug-and-play module.

Acknowledgment

The authors would like to thank the associate editor and the reviewers for their valuable comments and advices.

References

[1] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.

[2] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[3] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," 2018, *arXiv:1801.07698*.

[4] B. Yin, L. Tran, H. Li, X. Shen, and X. Liu, "Towards interpretable face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9347–9356.

[5] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," 2018, *arXiv:1810.11580*.

[6] A. Stylianou, R. Souvenir, and R. Pless, "Visualizing deep similarity networks," 2019, *arXiv:1901.00536*.

[7] T. Zee, G. Gali, and I. Nwogu, "Enhancing human face recognition with an interpretable neural network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.

[8] J. R. Williford, B. B. May, and J. Byrne, "Explainable face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 248–263.

[9] M. Ren, Y. Wang, Z. Sun, and T. Tan, "Dynamic graph representation for occlusion handling in biometrics," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 11940–11947, Apr. 2020.

[10] M. Ren, Y. Wang, Y. Zhu, K. Zhang, and Z. Sun, "Multiscale dynamic graph representation for biometric recognition with occlusions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15120–15136, Dec. 2023.

[11] O. Lang et al., "Explaining in style: Training a GAN to explain a classifier in stylespace," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 693–702.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[13] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.

[14] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3449–3457.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[16] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[18] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–10.

[19] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[20] M. Böhle, M. Fritz, and B. Schiele, "Convolutional dynamic alignment networks for interpretable classifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10024–10033.

[21] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[22] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4829–4837.

[23] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, p. 841, 2017.

[24] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 607–617.

[25] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2376–2384.

[26] A. K. Jain, D. Deb, and J. J. Engelsma, "Biometrics: Trust, but verify," 2021, *arXiv:2105.06625*.

[27] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.

[28] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196.

[29] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015, *arXiv:1506.06579*.

[30] O. Chris, M. Alexander, and S. Ludwig, "Feature visualization," *Distill*, vol. 2, pp. 1–7, Nov. 2017.

[31] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[32] W. Brendel and M. Bethge, "Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.

[33] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 2169–2178.

[34] P. Dhar, A. Bansal, C. D. Castillo, J. Gleason, P. J. Phillips, and R. Chellappa, "How are attributes expressed in face DCNNs?" in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 85–92.

[35] N. Abudarham, I. Grosbard, and G. Yovel, "Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition," *Cogn. Sci.*, vol. 45, no. 9, pp. 1–20, Sep. 2021.

[36] Y. Shi and A. Jain, "Probabilistic face embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6901–6910.

[37] S. Li, J. Xu, X. Xu, P. Shen, S. Li, and B. Hooi, "Spherical confidence learning for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15629–15637.

[38] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[40] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[41] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4832–4842.

[42] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[43] J. Guo, X. Zhu, and Z. Lei. (2018). *3DDFA*. [Online]. Available: https://github.com/cleardusk/3DDFA

[44] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 152–168.

[45] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, Jan. 2019.

[46] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn.*, 1999, pp. 187–194.

[47] G. B. Huang, M. Mattar, T. Berg, and L.-M. Eric, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images Detection, Alignment, Recognit.*, 2008, pp. 1–11.

[48] J. Sun, Q. Deng, Q. Li, M. Sun, M. Ren, and Z. Sun, "AnyFace: Free-style text-to-face synthesis and manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18666–18675.

[49] W. Zhong et al., "Identity-preserving talking face generation with landmark and appearance priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9729–9738.

[50] R. Huang, P. Lai, Y. Qin, and G. Li, "Parametric implicit face representation for audio-driven facial reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12759–12768.

[51] C. N. Duong, K. G. Quach, I. Jalata, N. Le, and K. Luu, "MobiFace: A lightweight deep learning face recognition on mobile devices," in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2019, pp. 1–6.

[52] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1997–2005.

[53] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[54] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1452–1466, 2021.