# Unpaired Caricature-Visual Face Recognition via Feature Decomposition-Restoration-Decomposition

Yang Xu, Yan Yan, *Senior Member, IEEE,* Jing-Hao Xue, *Senior Member, IEEE,* Yang Hua, and Hanzi Wang, *Senior Member, IEEE*

***Abstract*—Existing caricature-visual face recognition methods train the models based on caricature-visual image pairs from the same identities. Unfortunately, in many real-world applications, facial caricatures and visual facial images are usually unpaired in the training set due to the difficulty of collecting facial caricatures drawn by artists. In this paper, we study caricature-visual face recognition under the practical setting that only unpaired facial caricature and visual facial images are available as training samples, and define this setting as unpaired caricature-visual face recognition. To this end, we develop a novel feature decomposition-restoration-decomposition method (FDRD), which mainly consists of a backbone network, an identity-oriented feature decomposition module, and a modality-oriented feature restoration module, to extract modality-irrelevant identity features. To effectively train FDRD in the case of limited facial caricature training samples, we develop a two-stage learning framework. In the first stage, we perform single-modality restoration, enabling the model to have the basic ability of feature decomposition and restoration for each modality. In the second stage, we perform cross-modality recognition by exchanging new modality features between the two modalities, facilitating the model to focus on the decoupling of identity features and modality features. Experimental results demonstrate that our method performs favorably against several state-of-the-art face recognition methods and cross-modality methods. Our code is available at https://github.com/Capricorn-Karma/FDRD.**

***Index Terms*—Cross-modality face recognition, Unpaired caricature-visual face recognition, Feature decomposition, Feature restoration.**

## I. Introduction

OVER the past few years, cross-modality face recognition has received significant attention due to the rapid growth of multi-modality data. Accordingly, a number of methods [1]–[4] have been developed and achieved promising performance. These efforts stem from the growing demand for advanced face recognition technologies that can operate across diverse visual representations (including sketch photos, near-infrared images, and caricatures), which are ubiquitous in many real-world applications.

Fig. 1: Examples of two facial images and their corresponding facial caricatures with diverse artistic styles. The images are from the CaVI dataset [5].

Unfortunately, cross-modality face recognition becomes extremely challenging when one of the modalities is a facial caricature. Unlike facial images captured in real-world scenarios, facial caricatures refer to facial artistic drawings that enhance certain facial instinctive features through extreme levels of distortions and exaggerations. Therefore, facial caricatures are not only significantly different from real-world facial images, but also show large intra-modality differences because of the diversity of artistic styles. In other words, both intra-class and inter-class variances of facial caricatures can be prominent. Some examples are given in Fig. 1. The large modality gap and intra-modality variations impose a huge challenge for cross-modality face recognition. In this paper, we study caricature-visual face recognition, which aims to match the facial images between caricature and visual (i.e., visible-light) modalities.

Existing caricature-visual face recognition methods [6]–[9] usually require caricature-visual image pairs from the same identities for model training. Note that facial caricatures are often created by artists, and thus they are relatively difficult to be collected. For example, existing facial caricature datasets (such as WebCaricature [10] and CaVI [5]) contain only celebrities, and the dataset scale is small. In contrast, real-world visual facial datasets (such as MS-Celeb-1M [11] and MS1MV3 [12]) are much larger in terms of dataset scale and identity number. Hence, facial caricatures and visual facial images are often unpaired (i.e., they do not belong to the same identities) in the training set for many applications. In this paper, we study an important but unexplored setting: unpaired caricature-visual face recognition (i.e., only unpaired facial

This work was in part supported by the National Natural Science Foundation of China under Grants 62372388, 62071404, and U21A20514. *(Corresponding author: Yan Yan.)*

Y. Xu, Y. Yan and H. Wang are with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China and the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: xuyang63250@stu.xmu.edu.cn; yanyan@xmu.edu.cn; hanzi.wang@xmu.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (e-mail: jinghao.xue@ucl.ac.uk).

Y. Hua is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK (e-mail: y.hua@qub.ac.uk).

caricature and visual facial images are available for training). Such a setting, which is practical in real-world scenarios, merits investigation.

Unpaired caricature-visual face recognition is important and applicable in various scenarios, such as social media, digital communication, digital entertainment, and gaming. Advancing this field contributes to the broader scope of face recognition technology, improving its accuracy and adaptability, especially in understanding artistic styles and cultural perceptions of beauty and representations.

A key issue of unpaired caricature-visual face recognition is how to extract modality-irrelevant identity features in the case of limited facial caricature training samples. A straightforward way is to generate caricature-visual pairs with the same identity and then leverage the contrastive-based loss for cross-modality matching. However, such a way cannot guarantee learning discriminative identity features since the contrastive-based loss easily suffers from overfitting when the training samples are not sufficient.

To address the above-mentioned issues, we tackle unpaired caricature-visual face recognition from the perspective of feature decomposition-restoration-decomposition. Generally, we first decompose the input caricature-visual images into identity features and modality features. Then, we restore new caricature and visual features by exchanging modality features between the two modalities. Thus, the caricature-visual feature pairs with the same identities can be generated. Next, instead of relying on the contrastive-based loss, we propose to further decompose the caricature-visual feature pairs, where we can impose consistency constraints on both identity and modality. As a result, we can successfully decouple the identity information and modality information. This greatly improves the performance when only limited facial caricature training samples are available.

Specifically, we propose a novel feature decomposition-restoration-decomposition (FDRD) method for unpaired caricature-visual face recognition. FDRD contains a backbone network, an identity-oriented feature decomposition (IFD) module, and a modality-oriented feature restoration (MFR) module. The IFD module decomposes the input features from the backbone into identity features and modality features, while the MFR module (containing a modality learning block and a feature fusion block) restores new caricature and visual features with the same identity, obtaining paired caricature-visual features. Such a way allows us to extract the modality-irrelevant identity information.

To effectively train FDRD, we develop a two-stage learning framework including a single-modality restoration stage and a cross-modality recognition stage for model training. The first stage pre-trains the model to enable the network to have the basic ability of feature decomposition and restoration for each modality. The second stage fine-tunes the pre-trained model to encourage the network to focus on the decoupling of identity features and modality features.

In summary, our contributions are summarized as follows:

- To the best of our knowledge, we are the first to study unpaired caricature-visual face recognition. We design a decomposition-restoration-decomposition structure to successfully extract identity features from unpaired caricature-visual images. Instead of using the contrastive-based loss, we perform decomposition on the generated caricature-visual feature pairs, avoiding overfitting given limited facial caricature training samples.
- We introduce a novel two-stage learning framework to effectively train our network. A pre-trained model is obtained by only performing single-modality restoration, and thus it can be easily fine-tuned to perform cross-modality recognition. In this way, the modality differences can be significantly alleviated.
- We extensively evaluate our method on several popular caricature-visual face recognition datasets and show its superiority over state-of-the-art methods. This clearly shows the potential of our decomposition-restoration-decomposition structure for addressing the unpaired cross-modality face recognition setting.

The remainder of this paper is organized as follows. First, we review the related work in Section II. Then, we elaborately describe our proposed method in Section III. Next, we perform extensive experiments on three caricature-visual face recognition datasets in Section IV. Finally, we draw the conclusion in Section V.

## II. Related Work

In this section, we review the methods closely related to our method. We first introduce cross-modality face recognition in Section II-A. Then, we briefly review caricature-visual face recognition in Section II-B.

### A. Cross-Modality Face Recognition

Cross-modality face recognition (or heterogeneous face recognition) methods can be roughly divided into modality-shared feature learning and modality-specific information compensation.

Modality-shared feature learning-based methods [13]–[17] either project the features from different modalities onto a common feature space or reduce the modality gap by extracting domain-independent feature representations. Wang *et al.* [13] propose a deep neural network-based method with canonical correlation analysis (CCA) and apply this method to heterogeneous face recognition. He *et al.* [14] map high-level facial feature representations into two orthogonal subspaces to encode domain-invariant identity information and domain-related spectrum information. Wu *et al.* [15] introduce coupled deep learning (CDL) by imposing a nuclear-norm constraint on a fully-connected layer to alleviate overfitting. Hu *et al.* [16] propose a new orthogonal modality disentanglement method with a joint modality-invariant loss and a deep representation alignment network to address the cross-modality face recognition problem. Hu *et al.* [17] develop a novel dual face alignment learning (DFAL) method to learn the potential domain-invariant neutral face representations from the cross-modality images.

Modality-specific information compensation-based methods [1], [4], [18] try to compensate for the missing modality-specific information in each modality. Yang *et al.* [4] use

a generative adversarial network to generate facial images, enriching the attribute diversity of synthetic images. DVG-Face [1] generates heterogeneous facial images with the same identity from noise. The identity consistency and diversity properties allow the model to use these generated images to extract domain-invariant features. Yang *et al.* [18] propose a novel neutral face learning and progressive fusion synthesis (NLPF) network to disentangle the latent attributes of heterogeneous faces and learn neutral face representations. Note that the above methods focus on paired cross-modality face recognition, where each identity involves images from different modalities. In contrast, we study unpaired caricature-visual face recognition, where only unpaired facial caricature and visual facial images are available for training. In addition, the above methods often rely on one-stage training for cross-modality matching. On the contrary, we develop a two-stage training framework specifically designed for unpaired caricature-visual face recognition. Such a way enables us to obtain an effective cross-modality recognition model.

Existing methods mainly work on visible-infrared face recognition. Different from real-world facial images, facial caricatures involve exaggerated and distorted transformations. This substantially increases the difficulty of preserving the identity during modality generation. In this paper, instead of generating new facial caricatures or visual facial images, we consider disentangling identity and modality information at the feature level and restoring new caricature and visual features. Such a way avoids the difficulty of generating new facial caricatures or visual facial images from another modality as well as potential ethical issues associated with caricature image generation. In addition, during the feature decomposition and restoration process, we learn the distributions of different modalities. Hence, we can randomly generate new modality features from the learned distributions, facilitating the extraction of modality-irrelevant identity information. Compared with existing methods, our method can effectively separate the identity information and modality information.

### B. Caricature-Visual Face Recognition

Most existing caricature-visual face recognition methods [5]–[9], [19]–[21] belong to modality-shared feature learning-based methods. These methods mainly focus on contrastive learning and feature decoupling.

Contrastive learning, which enhances the similarities between positive pairs and the differences between negative pairs, is widely used in face recognition. Li *et al*. [6] propose a unified feature representation and similarity learning framework for contrastive learning. Dai *et al*. [7] introduce the gating to fuse local and global features, and use a convolutional attention module to improve the discriminative ability of features. Huo *et al*. [19] evaluate the influence of unimodal and multimodal metrics on facial caricatures and visual facial images during feature matching. Mishra *et al*. [8] propose a nonlinear transformation method, which maps the features from different modalities into a common subspace. Wang *et al*. [20] design a novel large-margin cross-domain contrast (LCC) loss to stimulate intra-class densification and inter-class separability. Moreover, they develop a cross-batch semantic metric (CSM) mechanism to improve the performance of sketch-based image retrieval.

Feature decoupling-based methods decouple the image features into multiple independent factors. Garg *et al*. [5] decouple shared and modality-specific representations with an orthogonal constraint and classify the facial identity by using a combination of shared and modality-specific representations. Ming *et al*. [9] propose a dynamic multi-task learning framework to decouple identity-sharing features by dynamically adjusting the weights of each task.

Existing caricature-visual face recognition methods train the models based on image pairs with the same identity. Unlike these methods, we focus on unpaired caricature-visual face recognition and develop a novel modality-specific information compensation-based method for this setting. We first decompose the input basic features into identity features and modality features. Based on it, we restore the new caricature and visual features by exchanging the modality features between the two modalities. Therefore, we can construct feature pairs from the two modalities with the same identity. Based on feature pairs, we perform decomposition again with the parameter-shared IFD module to facilitate the extraction of modality-irrelevant identity information.

## III. Proposed Method

In this section, we introduce our proposed method for unpaired caricature-visual face recognition. First, we give the problem formulation in Section III-A. Then, we provide the overview of our method in Section III-B. Next, we give the two-stage learning framework in Section III-C. Next, we introduce the key components (including the IFD module and the MFR module) of our proposed method in Sections III-D and III-E, respectively. Finally, we describe the joint loss in Section III-F.

### A. Problem Definition

In this paper, we study caricature-visual face recognition under the setting that only unpaired facial caricature and visual facial images are given as training samples. We define this setting as unpaired caricature-visual face recognition. Such a setting is very important and practical since it is relatively difficult to obtain paired caricature-visual facial images for training in many applications (note that collecting facial caricatures drawn by artists is not a trivial task).

Formally, the whole training set consists of a facial caricature subset $\mathcal{D}_c = \{\mathbf{x}_{i,c}, y_{i,c}\}_{i=1}^{N_c}$ and a visual facial image subset $\mathcal{D}_v = \{\mathbf{x}_{j,v}, y_{j,v}\}_{j=1}^{N_v}$, where $\mathbf{x}_{i,c}$ and $y_{i,c}$ denote the $i$-th facial caricature and its corresponding identity label in $\mathcal{D}_c$, respectively; $\mathbf{x}_{j,v}$ and $y_{j,v}$ denote the $j$-th visual facial image and its corresponding identity label in $\mathcal{D}_v$, respectively; $N_c$ and $N_v$ represent the numbers of images in $\mathcal{D}_c$ and $\mathcal{D}_v$, respectively. The facial identities in $\mathcal{D}_c$ and $\mathcal{D}_v$ do not overlap. The test set $\mathcal{D}_t$ contains both facial caricatures and visual facial images. In this paper, we study close-set cross-modality face recognition (the identities of test images exist in the training set). Given a facial caricature or a visual facial image in $\mathcal{D}_t$,

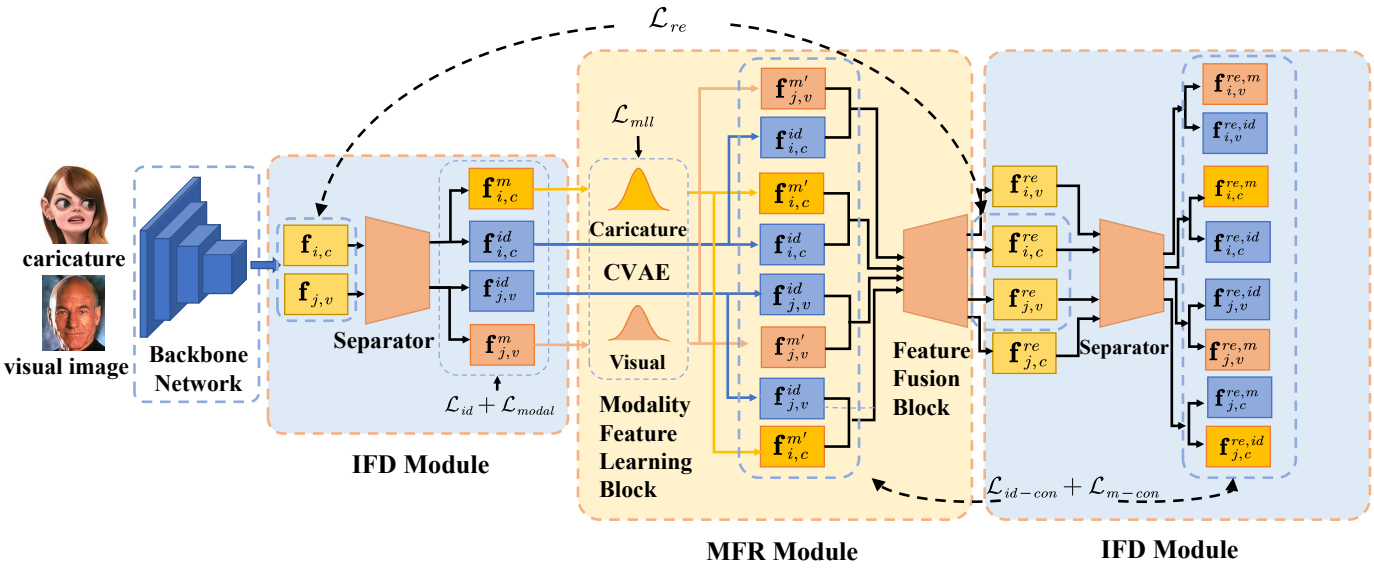



Fig. 2: Overview of our FDRD method. FDRD is composed of a backbone network, an identity-oriented feature decomposition (IFD) module, and a modality-oriented feature restoration (MFR) module. In the IFD module, the separator decomposes the input feature into the identity feature and modality feature. The modality feature learning block generates the new modality feature while the feature fusion block fuses the identity feature and modality feature to restore a new caricature or visual feature. Note that the IFD module is a parameter-shared module for feature decomposition.

we determine its facial identity based on the training images of another modality.

### B. Overview

In this paper, we develop a novel FDRD method, consisting of a backbone network, two IFD module, and an MFR module, for unpaired caricature-visual face recognition. The overview of our method is shown in Fig. 2.

More specifically, given unpaired caricature-visual facial images, we first pass them through a backbone network and obtain the basic features for each modality. In this paper, we use Inception-ResNet-v1 [22] as our backbone network. Then, we feed the basic features into the first IFD module to learn identity features and modality features for each modality. The modality features capture the modality-specific information which is irrelevant to the identity. Based on this, we further introduce an MFR module to restore new caricature and visual features with the same identity. Next, instead of relying on the contrastive-based loss (which is often hard to optimize and easily suffers from overfitting [23]), we feed paired features into the second IFD module for another decomposition, where the identity consistency loss and the modality consistency loss are introduced. Finally, the modality-irrelevant identity feature is extracted for each modality.

Note that the network parameters of the two IFD modules are shared. The inputs of the first IFD module and the second IFD module are the original basic features and the restored features, respectively. The purpose of designing the two IFD modules is to perform cyclic consistency learning during the decomposition-restoration-decomposition process. Such a way facilitates the decoupling process to focus on separating the identity information and modality information.

For model inference, only the backbone network and the first IFD module are required to perform feature extraction. In this way, the inference model is a lightweight network. For each image in the probe set, the identity feature is first extracted by the inference model, and then cosine similarity matching is performed between the identity feature of the probe image and those of the gallery set. Finally, the most similar identity in the gallery is selected as the query result. Note that the identity features of each identity in the gallery are extracted by the inference model and they are clustered according to the $k$-means algorithm ($k = 3$ in this paper). In this way, three feature centers are collected to represent different styles of caricatures/visual images for each identity.

### C. Two-Stage Learning Framework

To effectively train FDRD, we develop a two-stage learning framework. The framework contains a single-modality restoration stage and a cross-modality recognition stage. Technically, in the first stage, we pre-train the whole network without exchanging modality information, where only the first IFD module and an MFR module are used. In this stage, different modality feature distributions are learned. Based on it, the reconstruction loss between the original features and restored features is constructed to ensure that the model has the basic ability of feature decomposition and restoration for each modality. In the second stage, we introduce cross-modality reconstruction between identity features and different modality features to restore paired caricature-visual features. Restored features are re-decomposed by the second IFD module, and the entire network is fine-tuned according to identity-consistency and modality-consistency constraints. In this way, we focus on decoupling the identity information and modality information.

**Single-Modality Restoration Stage.** During each iteration, we first randomly select a batch of facial images from the training set, including facial caricatures and visual facial images with non-overlapping identities. Then, we feed the batch into the backbone to extract basic features, followed by the IFD module to extract the identity features and modality features. Next, these features are fed into the MFR module to generate new modality features and restore caricature features and visual features (the details of the IFD and MFR modules are introduced in Sections III-D and III-E).

For the restored caricature or visual features that have the same modality as the original one, we leverage the reconstruction loss to constrain the distances between the original features and the restored features, that is,

$$\mathcal{L}_{re} = \frac{1}{N_c}\sum_{i=1}^{N_c} L_1(\mathbf{f}_{i,c}, \mathbf{f}_{i,c}^{re}) + \frac{1}{N_v}\sum_{j=1}^{N_v} L_1(\mathbf{f}_{j,v}, \mathbf{f}_{j,v}^{re}), \quad (1)$$

where $L_1(\cdot,\cdot)$ denotes the $L_1$ distance. $N_c$ and $N_v$ denote the number of caricature facial and visual facial images, respectively. $\mathbf{f}_{i,c}$ and $\mathbf{f}_{j,v}$ denote the basic features from the backbone network, given the $i$-th facial caricature image $\mathbf{x}_{i,c}$ from $\mathcal{D}_c$ and the $j$-th visual facial image $\mathbf{x}_{j,v}$ from $\mathcal{D}_v$. $\mathbf{f}_{i,c}^{re}$ and $\mathbf{f}_{j,v}^{re}$ denote the restored caricature feature and visual feature, respectively.

**Cross-Modality Recognition Stage.** Based on the pre-trained model in the first stage, the cross-modality recognition stage encourages the model training to focus on extracting modality-irrelevant identity features and identity-irrelevant modality features.

On the one hand, we restore a new caricature-visual feature pair $\{\mathbf{f}_{i,c}^{re}, \mathbf{f}_{i,v}^{re}\}$ by combining the identity feature ($\mathbf{f}_{i,c}^{id}$) from the caricature modality and two modality features ($\mathbf{f}_{i,c}^{m'}$ and $\mathbf{f}_{j,v}^{m'}$). The feature pair is further fed into the feature separator of the IFD module to extract the identity features $\mathbf{f}_{i,c}^{re,id}$ and $\mathbf{f}_{i,v}^{re,id}$ (the details of feature decomposition and restoration are introduced in Sections III-D and III-E). Under such a decomposition-restoration-decomposition structure, we can impose the identity consistency loss on the identity features obtained from different modalities, enabling the model to focus on the extraction of identity information from the two modalities. The identity consistency loss for the caricature modality is defined as

$$\begin{aligned}\mathcal{L}_{id\text{-}con}^{c} =& \frac{1}{N_c}\sum_{i=1}^{N_c}\left(\left\|\mathbf{f}_{i,c}^{id} - \mathbf{f}_{i,c}^{re,id}\right\|_2^2 + \left\|\mathbf{f}_{i,c}^{id} - \mathbf{f}_{i,v}^{re,id}\right\|_2^2\right) \\ &- \frac{1}{N_c}\sum_{i=1}^{N_c}\sum_{c=1}^{C_c}\mathbb{I}_{[c=y_i^c]}\log\left(\mathcal{P}_{id}\left(\mathbf{f}_{i,c}^{re,id}\right)\right) \\ &- \frac{1}{N_c}\sum_{i=1}^{N_c}\sum_{c=1}^{C_c}\mathbb{I}_{[c=y_i^c]}\log\left(\mathcal{P}_{id}\left(\mathbf{f}_{i,v}^{re,id}\right)\right).\end{aligned} \quad (2)$$

Similarly, we can obtain the identity consistency loss for the visual modality as $\mathcal{L}_{id\text{-}con}^{v}$. Thus, the joint identity consistency loss is $\mathcal{L}_{id\text{-}con} = \mathcal{L}_{id\text{-}con}^{c} + \mathcal{L}_{id\text{-}con}^{v}$.

On the other hand, we also impose the modality consistency loss on the modality features from the two modalities. The new caricature-visual feature pair $\{\mathbf{f}_{i,c}^{re}, \mathbf{f}_{i,v}^{re}\}$ is fed into the feature separator of the IFD module to extract the modality features $\mathbf{f}_{i,c}^{re,m}$ and $\mathbf{f}_{i,v}^{re,m}$. The modality consistency loss for the caricature modality is defined as

$$\mathcal{L}_{m\text{-}con}^{c} = \mathcal{L}_{BCE}\left(\mathcal{P}_m\left(\mathbf{f}_{i,c}^{re,m}\right), m_i\right) + \mathcal{L}_{BCE}\left(\mathcal{P}_m\left(\mathbf{f}_{i,v}^{re,m}\right), m_i\right). \quad (3)$$

where $\mathcal{L}_{BCE}(\cdot,\cdot)$ denotes the binary cross-entropy.

Similarly, we can obtain the modality consistency loss for the visual modality as $\mathcal{L}_{m\text{-}con}^{v}$. Thus, the joint modality consistency loss is $\mathcal{L}_{m\text{-}con} = \mathcal{L}_{m\text{-}con}^{c} + \mathcal{L}_{m\text{-}con}^{v}$.

### *D. Identity-Oriented Feature Decomposition (IFD)*

The IFD module is designed to decompose the basic features into identity features and modality features. Technically, the IFD module consists of a $1 \times 1$ convolutional layer $\mathcal{D}(\cdot)$ and a separator $\mathcal{S}(\cdot)$ (including two convolutional blocks, each of which contains a $3 \times 3$ convolutional layer, a normalization layer, and an activation function). $\mathcal{D}(\cdot)$ plays the role of dimensionality reduction. $\mathcal{S}(\cdot)$ learns a non-linear projection function that decomposes the input features into identity features and modality features. Mathematically, the IFD module can be formulated as

$$\begin{aligned}\mathbf{f}_{i,c}^{id} &= \mathbf{w}_c\mathcal{S}\left(\mathcal{D}(\mathbf{f}_{i,c})\right),\ \mathbf{f}_{i,c}^{m} = \mathcal{D}(\mathbf{f}_{i,c}) - \mathbf{w}_c\mathcal{S}\left(\mathcal{D}(\mathbf{f}_{i,c})\right), \\ \mathbf{f}_{j,v}^{id} &= \mathbf{w}_v\mathcal{S}\left(\mathcal{D}(\mathbf{f}_{j,v})\right),\ \mathbf{f}_{j,v}^{m} = \mathcal{D}(\mathbf{f}_{j,v}) - \mathbf{w}_v\mathcal{S}\left(\mathcal{D}(\mathbf{f}_{j,v})\right),\end{aligned} \quad (4)$$

where $\mathbf{w}_c$ and $\mathbf{w}_v$ denote learnable parameters to normalize the spatial distribution of identity features, which can compensate for the distribution shift. $\mathbf{f}_{i,c}^{id}$ and $\mathbf{f}_{j,v}^{id}$ are the identity features for the caricature and visual modalities, respectively. $\mathbf{f}_{i,c}^{m}$ and $\mathbf{f}_{j,v}^{m}$ are the modality features for the caricature and visual modalities, respectively. The identity features are obtained by a non-linear mapping function in the separator, while the modality features are extracted by subtracting the identity features from the basic features.

To effectively extract identity features, we leverage the commonly used cross-entropy loss and the center loss [24] (which can enhance the compactness of the identity features from the same identities), given as follows:

$$\mathcal{L}_{id} = \frac{1}{2}\sum_{n=1}^{N}\left\|\mathbf{f}_n^{id} - \mathbf{c}_{y_n}\right\|_2^2 - \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\mathbb{I}_{[c=y_n]}\log\left(\mathcal{P}_{id}\left(\mathbf{f}_n^{id}\right)\right), \quad (5)$$

where $\mathbf{c}_{y_n}$ represents the feature center of the identity to which the current feature belongs, and it is iteratively updated along the direction of the mean of the sample feature vectors from the same class during the training process. $\mathbf{f}_n^{id}$ (which can be computed via Eq. (4)) and $y_n$ represent the identity feature obtained by the separator for the $n$-th image (a facial caricature or a visual facial image) and its corresponding label in the whole training set. $\mathcal{P}_{id}(\cdot)$ is the prediction function (an FC layer) for identity classification. $\mathbb{I}_{[c=y_n]}$ equals to 1 when $c = y_n$, and 0 otherwise. $N$=$N_c$+$N_v$ and $C$=$C_c$+$C_v$ denote the total number of images and the total number of facial identities in $\mathcal{D}$, respectively.

To extract modality features, we perform two-class classification, which distinguishes the modality features between

the two modalities. We employ a binary cross-entropy loss $\mathcal{L}_{BCE}(\cdot,\cdot)$, which is

$$\mathcal{L}_{modal} = \mathcal{L}_{BCE}\left(\mathcal{P}_m\left(\mathbf{f}_n^m\right), m_n\right), \tag{6}$$

where $m_n$ indicates that the current feature belongs to the caricature modality ($m_n = 1$) or the visual modality ($m_n = 0$). $\mathbf{f}_n^m$ (which can be computed via Eq. (4)) represents the modality feature obtained by the separator. $\mathcal{P}_m(\cdot)$ is the prediction function (an FC layer) for modality classification.

### *E. Modality-Oriented Feature Restoration (MFR)*

The IFD module decomposes the basic features into identity features and modality features for each modality. However, the distances between the identity features from the two modalities are large since they belong to different identities. Hence, the MFR module is designed to restore new caricature and visual features by exchanging modality features between the two modalities, and thus generate paired caricature-visual features with the same identity. In this way, we can impose identity constraints across the two modalities, facilitating the extraction of modality-irrelevant identity information. In particular, we randomly sample the modality features from modality feature distributions based on conditional variational autoencoders (CVAE) [25], [26]. Such a manner not only further filters out residual identity information from the modality information, but also enhances the diversity of modality information to prevent the model from overfitting.

The MFR module consists of a modality learning block and a feature fusion block. The modality learning block models the modality feature distributions for caricature and visual modalities. Thus, we can randomly generate new caricature and visual modality features. To distinguish modality features from the two modalities, we introduce the modality label information to CVAE. Technically, two modality feature distributions are learned from the facial caricature subset $\mathcal{D}_c$ and the visual facial image subset $\mathcal{D}_v$. Similar to VAE [25], assume that a latent vector $\mathbf{z}$ is generated from the prior distribution $p_\theta(\mathbf{z})$ and the modality feature $\mathbf{f}^m$ is generated by the generative distribution $p_\theta(\mathbf{f}^m|\mathbf{z})$ conditioned on $\mathbf{z}$: $\mathbf{z} \sim p_\theta(\mathbf{z}), \mathbf{f}^m \sim p_\theta(\mathbf{f}^m|\mathbf{z})$. In general, the posterior distribution is difficult to solve. Thus, an approximate posterior in the form of $q_\phi(\mathbf{z}|\mathbf{f}^m)$ is introduced to approximate the true posterior $p_\theta(\mathbf{z}|\mathbf{f}^m)$, where we assume that the posterior distribution is a multivariate Gaussian distribution with a diagonal covariance matrix. The parameters of the approximate posterior distribution can be fitted by the encoder $\mathcal{R}_{\text{encoder}}(\cdot,\cdot)$. $\mathcal{R}_{\text{encoder}}(\cdot,\cdot)$ is comprised of two paralleled FC layers $\mathcal{R}^{\mu}_{\text{encoder}}(\cdot,\cdot)$ and $\mathcal{R}^{\sigma}_{\text{encoder}}(\cdot,\cdot)$, which are used to fit the mean and variance, respectively.

To distinguish modality features from the two modalities, we introduce the modality label information to CVAE. The posterior distribution parameters fitted to a caricature modality feature $\mathbf{f}_{i,c}^m$ or a visual modality feature $\mathbf{f}_{j,v}^m$ are given as

$$\begin{aligned}
\mu_{i,c} &= \mathcal{R}^{\mu}_{\text{encoder}}\left(\mathbf{f}_{i,c}^m, m_c\right), \log\sigma_{i,c}^2 = \mathcal{R}^{\sigma}_{\text{encoder}}\left(\mathbf{f}_{i,c}^m, m_c\right), \\
\mu_{j,v} &= \mathcal{R}^{\mu}_{\text{encoder}}\left(\mathbf{f}_{j,v}^m, m_v\right), \log\sigma_{j,v}^2 = \mathcal{R}^{\sigma}_{\text{encoder}}\left(\mathbf{f}_{j,v}^m, m_v\right),
\end{aligned} \tag{7}$$

where $\mu_{i,c}$ and $\sigma_{i,c}^2$ represent the mean and variance vectors of the posterior distribution corresponding to $\mathbf{f}_{i,c}^m$, respectively. $\mu_{j,v}$ and $\sigma_{j,v}^2$ represent the mean and variance vectors of the posterior distribution corresponding to $\mathbf{f}_{j,v}^m$, respectively. $m_c$ and $m_v$ are the modality labels for the caricature and visual modalities, respectively. Modality labels are represented in a one-hot encoding way. Modality features are concatenated with the modality labels as the input of the encoder.

To ensure the differentiable of the reconstruction process, sampling is performed by using the reparameterization trick, and a total of $L$ times are sampled. Hence, a latent vector $\mathbf{z}_{i,c}^l = \mu_{i,c} + \varepsilon_1^l \otimes \sigma_{i,c}$ ($l = 1, \ldots, L$ and '$\otimes$' denotes the element-wise product) is randomly sampled from the posterior distribution corresponding to $\mathbf{f}_{i,c}^m$, where $\varepsilon_1^l \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a random vector. Similarly, a latent vector $\mathbf{z}_{j,v}^l = \mu_{j,v} + \varepsilon_2^l \otimes \sigma_{j,v}$ is randomly sampled from the posterior distribution corresponding to $\mathbf{f}_{j,v}^m$, where $\varepsilon_2^l \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a random vector. Hence, two new modality features can be obtained by passing $\mathbf{z}_{i,c}^l$ and $\mathbf{z}_{j,v}^l$ with the corresponding modality labels through the decoder $\mathcal{R}_{decoder}$ (an FC layer), that is,

$$\mathbf{f}_{i,c}^{m',l} = \mathcal{R}_{\text{decoder}}\left(\mathbf{z}_{i,c}^l, m_c\right), \ \mathbf{f}_{j,v}^{m',l} = \mathcal{R}_{\text{decoder}}\left(\mathbf{z}_{j,v}^l, m_v\right), \tag{8}$$

where $\mathbf{f}_{i,c}^{m',l}$ and $\mathbf{f}_{j,v}^{m',l}$ respectively represent the new caricature modality features and visual modality features generated from the latent variables $\mathbf{z}_{i,c}^l$ and $\mathbf{z}_{j,v}^l$. Latent variables are concatenated with the modality labels as the input of the modality decoder.

Based on the above, the new modality features $\mathbf{f}_{i,c}^{m'} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{f}_{i,c}^{m',l}$ and $\mathbf{f}_{j,v}^{m'} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{f}_{j,v}^{m',l}$ can be obtained for the caricature and visual modalities, respectively.

The modality learning loss consists of a reconstruction error term (i.e., the mean square error (MSE) term) and a regularization term (i.e., the Kullback-Leibler (KL) divergence term) [25], that is,

$$L_{mll}\left(\mathbf{f}_n^m\right) = MSE\left(\mathbf{f}_n^m, \mathbf{f}_n^{m'}\right) + KL\left(N\left(\mu_n, \sigma_n^2\right) \| N\left(\mathbf{0}, \mathbf{I}\right)\right), \tag{9}$$

where $\mathbf{f}_n^m$ and $\mathbf{f}_n^{m'}$ respectively denote the original and newly generated modality features from the $n$-th image (which can be a facial caricature or a visual facial image) in a batch. $\mu_n$ and $\sigma_n^2$ represent the mean and variance vectors of the multivariate Gaussian distribution $q_\phi(\mathbf{z}_n|\mathbf{f}_n^m)$, respectively. $\mathbf{z}_n$ is a hidden vector. The MSE term ensures that the input and the output of the modality learning block are similar, while the KL term aligns the modality distribution to the standard Gaussian distribution $N(\mathbf{0}, \mathbf{I})$. Such a way can prevent the variances from being 0 while new modality features can be generated by sampling from the Gaussian distribution.

The final modality learning loss term can be expressed as

$$\mathcal{L}_{mll} = \frac{1}{N}\sum_{n=1}^{N} L_{mll}\left(\mathbf{f}_n^m\right). \tag{10}$$

The feature fusion block fuses the identity feature and the newly generated modality feature to restore the caricature or visual features. In fact, feature fusion can be viewed as the inverse process of feature decomposition. We stack identity features and modality features according to a scaling factor $\lambda$. An inverse nonlinear projection function $\mathcal{S}'(\cdot)$, which consists

of two deconvolution layers and two $1\times1$ convolution layers, is used for feature fusion, that is,

$$\begin{aligned}\mathbf{f}_{i,c}^{re} &= \mathcal{S}'\left(\mathbf{f}_{i,c}^{id} + \lambda\mathbf{f}_{i,c}^{m'}\right), \mathbf{f}_{i,v}^{re} = \mathcal{S}'\left(\mathbf{f}_{i,c}^{id} + \lambda\mathbf{f}_{j,v}^{m'}\right), \\ \mathbf{f}_{j,v}^{re} &= \mathcal{S}'\left(\mathbf{f}_{j,v}^{id} + \lambda\mathbf{f}_{j,v}^{m'}\right), \mathbf{f}_{j,c}^{re} = \mathcal{S}'\left(\mathbf{f}_{j,v}^{id} + \lambda\mathbf{f}_{i,c}^{m'}\right),\end{aligned} \tag{11}$$

where $\mathbf{f}_{i,c}^{re}$ and $\mathbf{f}_{i,v}^{re}$ respectively denote the new caricature feature and the new visual feature obtained by combining the identity features extracted from the caricature facial image with the generated caricature modality feature and the generated visual modality feature. $\mathbf{f}_{j,v}^{re}$ and $\mathbf{f}_{j,c}^{re}$ respectively denote the new visual feature and the new caricature feature obtained by combining the identity features extracted from the visual facial image with the generated caricature modality feature and the generated visual modality feature.

### F. Joint Loss

For the single-modality restoration stage, we encourage the network to have the basic ability of feature decomposition and restoration. Hence, the joint loss in this stage can be defined as

$$\mathcal{L}_{srs} = \mathcal{L}_{id} + \mathcal{L}_{modal} + \mathcal{L}_{mll} + \mathcal{L}_{re}. \tag{12}$$

For the cross-modality recognition stage, we encourage the model to focus on extracting modality-irrelevant identity features and identity-irrelevant modality features. The joint loss in this stage can be defined as

$$\mathcal{L}_{crs} = \mathcal{L}_{id} + \mathcal{L}_{modal} + \mathcal{L}_{mll} + \mathcal{L}_{re} + \mathcal{L}_{id\text{-}con} + \mathcal{L}_{m\text{-}con}. \tag{13}$$

Note that experimental results show that our method can achieve superior performance without assigning balancing parameters between these loss items on different datasets. Therefore, we do not employ weighting factors for different loss terms.

## IV. Experiments

In this section, we first introduce the datasets in Section IV-A. Then, we present the implementation details of our method in Section IV-B. Next, we compare our method with several state-of-the-art methods in Section IV-C. Finally, we conduct ablation studies in Section IV-D and give some visualization results in Section IV-E.

### A. Datasets

In this paper, we introduce three widely-used caricature-visual face recognition datasets. The WebCaricature dataset [10] is a popular caricature-visual facial dataset consisting of 6,042 facial caricatures and 5,974 visual facial images from 252 identities collected from the web. The CaVI dataset [5] contains images of 205 identities. There are 5,091 facial caricatures ranging from 10 to 15 images per identity and 6,427 visual facial images ranging from 10 to 15 images per identity. The IIIT-CFW dataset [27] includes a total of 8,928 cartoon characters (including caricatures, sketches, and paintings) from 100 identities. Each identity also provides 10 visual facial images.

### B. Implementation Details

In our experiments, we use RetinaFace [28] to automatically detect all the facial caricatures and visual facial images in the datasets. Each image is then cropped, aligned, and finally resized to the size of $160\times160$. The backbone network is based on Inception-ResNet-v1, where we remove the last pooling layer and the FC layer. To enable the model to effectively extract the identity feature, Inception-ResNet-v1 is pre-trained on the CASIA-WebFace dataset [29] and fine-tuned on the unpaired caricature-visual face recognition dataset. Similar to [25], [26], the sampling time of latent variables $L$ is set to 1. In Eq. (11), the scaling factor $\lambda$ is set to 1. All the experiments are implemented by Pytorch and run on a single NVIDIA GTX3090. The model is trained for 100 epochs at the single-modality restoration stage and 100 epochs at the cross-modality recognition stage. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to $1\times e^{-3}$, and the learning rate is decayed every 10 epochs until it reaches $1\times\ e^{-5}$. The batch size is set to 128 for all the datasets (we randomly choose 64 caricature-visual image pairs of different identities).

For each dataset, we randomly select half of the facial caricatures and half of the visual facial images (with no overlapped identities) for training and the rest of the dataset is used for cross-modality testing (i.e., given a facial caricature/visual facial image, we determine its identity according to its nearest distance to the visual/caricature training images). The above process is repeated for ten rounds. The final Rank-1 accuracy is obtained by the average of the ten-round tests. The cosine similarity is used to match modality-irrelevant features from the training and test images. We evaluate our method with two modes: C-to-V (Caricature-to-Visual) and V-to-C (Visual-to-Caricature), where C-to-V represents that facial caricatures and visual facial images are used as the probe set and the gallery set, respectively, and the other way around for V-to-C.

### C. Comparison with State-of-the-Art Methods

We compare our method with several state-of-the-art methods, containing popular visual face recognition methods and cross-modality recognition methods (including caricature-visual face recognition methods). For visual face recognition methods, we train the models by using all the training images (facial caricatures and visual facial images with non-overlapping identities) and their identity labels. Hence, these methods only extract the identity information from different facial images without considering the modality information. For cross-modality recognition methods, we train models by simultaneously feeding the facial caricatures subset and the visual facial image subset into the network. The inputs of CaVINet [5], Multi-task [9], and SagNet [36] are the same as our FDRD (facial caricatures and visual facial image pairs with non-overlapping identities). Due to the lack of pairs of images with the same identities, we replace the triplet loss with the center loss in these methods. MMD-AAE [34] and DANN [35] use the same numbers of facial caricatures and visual facial images as the input while leveraging modality labels for domain discrimination. DDG [37] and CIRL [38] are

TABLE I: Performance comparisons in terms of average accuracy (%) under the unpaired caricature-visual face recognition setting on WebCaricature, CaVI, IIIT-CFW datasets. The best results are marked in bold.

| Methods | WebCaricature | | CaVI | | IIIT-CFW | |
|---|---|---|---|---|---|---|
| | C-to-V | V-to-C | C-to-V | V-to-C | C-to-V | V-to-C |
| Center Loss [24] | 36.75 | 43.41 | 23.86 | 35.98 | 20.53 | 37.42 |
| CosFace [30] | 36.48 | 50.64 | 35.72 | 47.24 | 30.23 | 39.59 |
| ArcFace [31] | 39.14 | 52.73 | 37.63 | 51.13 | 33.43 | 51.27 |
| MagFace [32] | 47.26 | 59.17 | 45.14 | 58.33 | 42.53 | 60.23 |
| AdaFace [33] | 50.35 | 61.72 | 48.25 | 60.48 | 45.34 | 62.12 |
| CaVINet [5] | 45.84 | 57.54 | 43.59 | 53.18 | 41.61 | 53.40 |
| Multi-task [9] | 46.57 | 58.27 | 44.38 | 54.26 | 42.59 | 54.13 |
| MMD-AAE [34] | 37.86 | 45.21 | 34.75 | 42.78 | 28.81 | 40.49 |
| DANN [35] | 49.72 | 64.18 | 47.13 | 61.85 | 45.46 | 62.57 |
| SagNet [36] | 50.32 | 60.85 | 47.86 | 54.50 | 46.58 | 58.66 |
| DDG [37] | 57.86 | 69.02 | 57.14 | 68.24 | 55.87 | 69.34 |
| CIRL [38] | 58.30 | 68.94 | 58.05 | 68.73 | 56.71 | 69.27 |
| FDRD (Ours) | **61.69** | **72.12** | **60.07** | **70.05** | **58.64** | **70.42** |

TABLE II: The average accuracy (%) obtained by different variants of our method in ablation studies. The best results are marked in bold.

| Methods | WebCaricature | | CaVI | | IIIT-CFW | |
|---|---|---|---|---|---|---|
| | C-to-V | V-to-C | C-to-V | V-to-C | C-to-V | V-to-C |
| BL(Baseline) | 49.72 | 64.18 | 47.13 | 61.85 | 45.46 | 62.57 |
| BL+CL | 53.83 | 67.89 | 51.27 | 63.64 | 49.13 | 64.74 |
| BL+CL+DR | 53.26 | 66.92 | 51.24 | 63.04 | 48.57 | 64.06 |
| BL+CL+DRD | 57.21 | 68.62 | 55.14 | 66.83 | 53.82 | 67.12 |
| BL+CL+DRD+CVAE | 60.34 | 71.42 | 58.73 | 69.53 | 57.10 | 69.84 |
| BL+CL+DRD+TS | 59.97 | 71.62 | 57.49 | 68.21 | 56.66 | 68.72 |
| BL+CL+DRD+CVAE+TS | **61.69** | **72.12** | **60.07** | **70.05** | **58.64** | **70.42** |

CL indicates that the method uses the center loss. DR indicates that the method uses the decomposition-restoration structure and the triplet loss. DRD indicates that the method adopts the decomposition-restoration-decomposition structure. CVAE indicates that the method uses CVAE to increase the diversity of training samples. TS indicates that the method adopts a two-stage learning framework.

domain generalization-based methods that use facial caricature images (or visual facial images) for training, and visual facial images (or facial caricatures) for testing. Table I shows the performance obtained by several competing under the unpaired caricature-visual face recognition setting.

Our proposed FDRD method outperforms the current caricature-visual face recognition methods (CaVINet and Multi-task and the visual face recognition methods (such as Center Loss, CosFace, ArcFace, MagFace, and AdaFace) on all three datasets. Due to the large modality gap, the visual face recognition methods give worse results. The caricature modality has exaggerated and diverse artistic expressions, which lead to difficulty in identifying caricature features under the same identity. The visual face recognition methods have good performance for visual facial images even with low quality such as blurred images, but they do not work well for facial caricatures with diverse styles. Conventional caricature face recognition methods often require caricature-visual image pairs with the same identity as the input. Thus, they cannot achieve good performance in our setting.

Our proposed method also outperforms representative cross-modality recognition methods (such as MMD-AAE, DANN, SagNet, DDG, and CIRL) on the three datasets. MMD and DANN reduce the modality differences by confusing the data of different modalities. SagNet extracts the content (identity) and style information for cross-modality recognition. DDG and CIRL focus on extracting domain-invariant semantic information for disentanglement. However, the facial identities of these methods are easily disturbed by the modality information (especially for facial caricatures with exaggerated facial morphology), which leads to a performance drop. Our proposed method addresses the modality discrepancy problem by designing a DRD structure to remove the modality information involved in the features and extract identity features. In addition, we design a two-stage training framework to ensure the quality of feature-level restoration while separating identity features from modality features. The above results show the superiority of our method.

### *D. Ablation Studies*

We evaluate the baseline and several variants of our method on three datasets. Table II gives the comparison performance on these datasets. We adopt the DANN method [35] with Inception-ResNet-v1 as the baseline.

**Influence of Center Loss.** From Table II, BL+CL outperforms the Baseline method, validating the effectiveness of the center

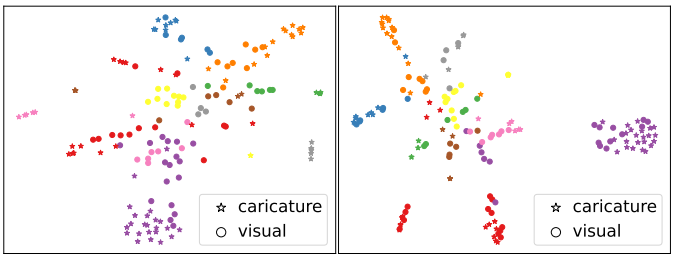
loss. This is because the center loss can effectively enhance the compactness of identity features.

**Influence of Decomposition-Restoration-Decomposition (DRD) Structure.** Compared with the BL+CL method, the variants that adopt the DRD structure achieve significant performance improvements in terms of the average recognition accuracy on three datasets. This is because the DRD structure is helpful to extract modality-irrelevant identity features by suppressing the modality information, greatly reducing the modality gap. On the contrary, BL+CL+DR achieves slightly lower performance than BL+CL. This indicates that the triplet loss is easier to suffer from overfitting compared with the classification loss. This is because the triplet loss requires a large number of triplets and is more difficult to be optimized. The above results show the importance of the DRD structure.

**Influence of CVAE.** BL+CL+DRD+CVAE models the modality features using CVAE and trains the model with only the cross-modality recognition stage. BL+CL+DRD+TS is similar to our proposed method except that the model is trained without CVAE. BL+CL+DRD+CVAE+TS is our proposed FDRD method which is trained by the two-stage learning framework.

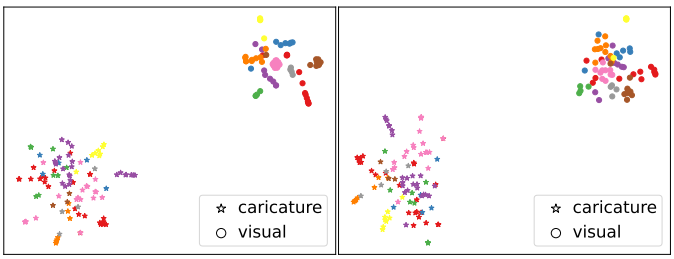
From Table II, we can see that BL+CL+DRD+CVAE+TS outperforms BL+CL+DRD+TS by a large margin and BL+CL+DRD +CVAE achieves better performance than BL+CL+DRD. This shows the importance of CVAE, which can generate new modality features, increasing the diversity of the training features. Such a way reduces overfitting and improves the final accuracy.

**Influence of Our Two-Stage Learning Framework.** From Table II, compared with BL+CL+DRD+CVAE, BL+CL+DRD+CVAE+TS achieves 1.06%, 0.96%, and 1.09% improvements on the Caricature, CaVI, and IIIT-CFW datasets, respectively. BL+CL+DRD+TS also outperforms BL+CL+DRD. This shows the necessity of the two-stage learning framework. That is, two-stage learning is beneficial for our unpaired caricature-visual face recognition setting. By performing single-modality restoration in the first stage, we can obtain a good pre-trained model, which can then be easily fine-tuned to achieve good performance in the case of limited training samples.



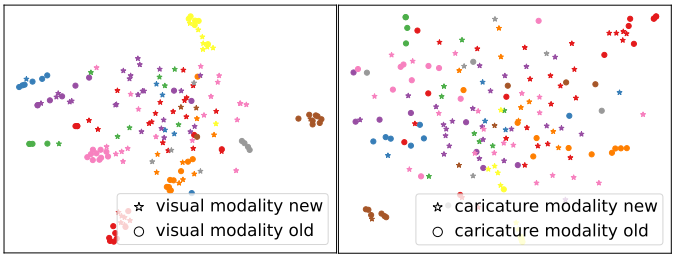

(a) (b)

Fig. 3: t-SNE visualization of identity features extracted by (a) the Baseline and (b) FDRD. The different colors represent different facial identities.




(a) (b)

Fig. 4: t-SNE visualization of (a) the original modality features and (b) the new modality features generated by CAVE.




(a) (b)

Fig. 5: t-SNE visualization of (a) the visual modality features and (b) the caricature modality features. Compare the original modality feature distribution with the newly generated modality feature distribution by CVAE

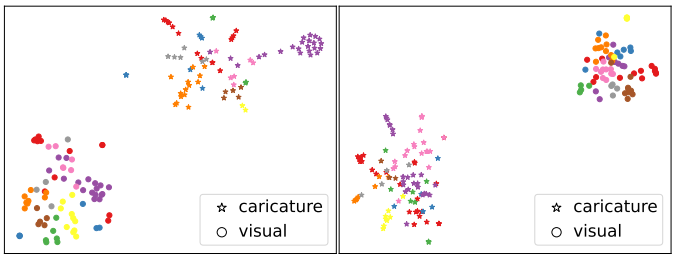
### *E. Visualization Results*

To further illustrate that the DFD structure effectively disentangles the identity features and the modality features, we randomly select facial caricatures of 10 identities in the training set and visual facial images of the same 10 identities in the test set, and then apply t-SNE [39] to visualize the identity features and the modality features. The results are shown in Figs. 3 and 4.

Fig. 3(a) shows the distribution of identity features extracted by the Baseline method. Fig. 3(b) shows the distribution of identity features given by our method (i.e., BL+CL+DFD+CVAE+TS). Due to the interference of the modality information, the identity features (from the two modalities) extracted by Baseline show a large modality gap. In contrast, the identity features learned by our method are distributed closely between the two modalities, indicating that our method significantly mitigates the modality gap. Therefore, the DFD structure can ensure that the extracted identity features from the two modalities are projected onto a common modality-irrelevant feature space.

Fig. 4 shows the distributions of the new modality features generated by CVAE and the original modality features given by IFD in our method. We can see that the distributions of the modality features extracted from facial caricature and visual images are different (clustering into two clusters). Fig. 5 visualizes the original and generated modality features in a single plot. We can see that the original modality features and the new modality features generated by CVAE belong to the same distribution, while the newly generated modality features show a diverse feature distribution.

Ideally, the modality features should be compactly dis-




(a) (b)

Fig. 6: t-SNE visualization of the modality features extracted after (a) the single-modality restoration stage and (b) the cross-modality recognition stage.


Fig. 7: t-SNE visualization of the basic features obtained by the backbone network and the restored features after the first learning stage by our method.

tributed for different facial identities. Fig. 6(a) shows the modality features obtained in the first learning stage. The modality features with the same identity tend to cluster together at this stage, indicating that the identity information and the modality information are not completely disentangled at this stage. Fig. 6(b) shows the modality features given by the IFD module after the two-stage learning. The modality features from the two modalities are more compactly distributed in two clusters, and there is no obvious identity gap. This indicates that the modality information is irrelevant to the identity information and is only dependent on each modality.

Fig. 7 shows the basic features extracted from the backbone and the new caricature and visual features restored by the MFR module after the first learning stage. The restored features and the basic features are close to each other. Thus, the model has the basic ability to decompose and restore after the first training stage. According to Figs. 6 and 7, based on the two-stage learning framework, the model effectively decouples modality-independent identity features and identity-independent modality features.

## V. Conclusion

In this paper, we propose a novel FDRD method for unpaired caricature-visual face recognition by designing a feature decomposition-restoration-decomposition structure. The proposed FDRD mainly consists of an IFD module and an MFR module to perform feature decomposition and cross-modality restoration. The IFD module decomposes the basic features from the backbone network into the identity features and the modality features, while the MRF module restores new caricature and visual features with the same identity, obtaining caricature-visual feature pairs. To train FDRD in the case of limited facial caricature training samples, we develop a two-stage learning framework. Extensive experiments show the effectiveness of our method on several popular caricature-visual face datasets. Currently, our method performs unpaired visual-caricature face recognition in a single domain. In future work, we plan to investigate cross-domain unpaired visual-caricature face recognition.

## References

[1] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "DVG-Face: Dual variational generation for heterogeneous face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2938–2952, 2021.

[2] H. Liu, X. Zhu, Z. Lei, D. Cao, and S. Z. Li, "Fast adapting without forgetting for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3093–3104, 2020.

[3] W. Hu and H. Hu, "Domain-private factor detachment network for NIR-VIS face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1435–1449, 2022.

[4] Z. Yang, J. Liang, C. Fu, M. Luo, and X.-Y. Zhang, "Heterogeneous face recognition via face synthesis with identity-attribute disentanglement," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1344–1358, 2022.

[5] J. Garg, S. V. Peri, H. Tolani, and N. C. Krishnan, "Deep cross modal learning for caricature verification and identification (CaVINet)," in *Proceedings of the ACM International Conference on Multimedia*, 2018, pp. 1101–1109.

[6] W. Li, J. Huo, Y. Shi, Y. Gao, L. Wang, and J. Luo, "A joint local and global deep metric learning method for caricature recognition," in *Proceedings of the Asian Conference on Computer Vision*, 2019, pp. 240–256.

[7] L. Dai, F. Gao, R. Li, J. Yu, X. Shen, H. Xiong, and W. Wu, "Gated fusion of discriminant features for caricature recognition," in *Proceedings of the International Conference on Intelligent Science and Big Data Engineering*, 2019, pp. 563–573.

[8] A. Mishra, "DHFML: Deep heterogeneous feature metric learning for matching photograph and caricature pairs," *International Journal of Multimedia Information Retrieval*, vol. 8, no. 3, pp. 135–142, 2019.

[9] Z. Ming, J.-C. Burie, and M. M. Luqman, "Dynamic deep multi-task learning for caricature-visual face cecognition," in *Proceedings of the International Conference on Document Analysis and Recognition Workshops*, 2019, pp. 92–97.

[10] J. Huo, W. Li, Y. Shi, Y. Gao, and H. Yin, "WebCaricature: A benchmark for caricature recognition," *arXiv preprint arXiv:1703.03230*, 2017.

[11] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 87–102.

[12] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi, "Lightweight face recognition challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 2638–2646.

[13] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 1083–1092.

[14] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for NIR-VIS face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 2000–2006.

[15] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 1679–1686.

[16] W. Hu and H. Hu, "Orthogonal modality disentanglement and representation alignment network for NIR-VIS face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3630–3643, 2021.

[17] W. Hu, W. Yan, and H. Hu, "Dual face alignment learning network for NIR-VIS face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2411–2424, 2021.

[18] Y. Yang, W. Hu, and H. Hu, "Neutral face learning and progressive fusion synthesis network for NIR-VIS face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[19] J. Huo, Y. Gao, Y. Shi, and H. Yin, "Variation robust cross-modal metric learning for caricature recognition," in *Proceedings of the on Thematic Workshops of ACM Multimedia*, 2017, pp. 340–348.

[20] X. Wang, D. Peng, P. Hu, Y. Gong, and Y. Chen, "Cross-domain alignment for zero-shot sketch-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[21] W. Zheng, L. Yan, F.-Y. Wang, and C. Gou, "Learning from the past: Meta-continual learning with knowledge embedding for jointly sketch, cartoon, and caricature face recognition," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 736–743.

[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.

[23] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[24] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 499–515.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[26] K. Sohn, X. Yan, and H. Lee, "Learning structured output representation using deep conditional generative models," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2015, pp. 3483–3491.

[27] A. Mishra, S. N. Rai, A. Mishra, and C. Jawahar, "IIIT-CFW: A benchmark database of cartoon faces in the wild," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 35–47.

[28] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[29] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[30] 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," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[31] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[32] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 225–14 234.

[33] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 750–18 759.

[34] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.

[35] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 1180–1189.

[36] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8690–8699.

[37] H. Zhang, Y.-F. Zhang, W. Liu, A. Weller, B. Schölkopf, and E. P. Xing, "Towards principled disentanglement for domain generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8024–8034.

[38] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8046–8056.

[39] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.


**Yang Xu** received the B.E. degree from the School of Computer Science and Engineering, Northeastern University, China, in 2021. He is currently pursuing the master's degree with the Institute of Artificial Intelligence, Xiamen University, China. His main research interests include computer vision and cross-modality recognition.

**Yan Yan** (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, China, in 2009. He worked as a Research Engineer with the Nokia Japan Research and Development Center from 2009 to 2010. He worked as a Project Leader with the Panasonic Singapore Laboratory in 2011. He is currently a Full Professor with the School of Informatics, Xiamen University, China. He has published around 100 papers in the international journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, CVPR, ICCV, ECCV, AAAI, and ACM MM. His research interests include computer vision and pattern recognition.

**Jing-Hao Xue** (Senior Member, IEEE) received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is currently a Professor with the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He received the Best Associate Editor Award of 2021 from the IEEE Transactions on Circuits and Systems for Video Technology, and the Outstanding Associate Editor Award of 2022 from the IEEE Transactions on Neural Networks and Learning Systems.

**Yang Hua** received the PhD degree from Université Grenoble Alpes/Inria Grenoble Rhone-Alpes, France, funded by Microsoft Research's Inria Joint Center. He is presently a lecturer with Queens University Belfast, United Kingdom. He has won four titles of prestigious international competitions in the field of computer vision and machine learning. His research interests include artificial intelligence and computer vision.

**Hanzi Wang** (Senior Member, IEEE) received the Ph.D. degree in computer vision from Monash University. He is currently a Distinguished Professor of Minjiang Scholars, Fujian, and the Founding Director of the Center for Pattern Analysis and Machine Intelligence (CPAMI), Xiamen University, China. He has published more than 100 academic papers in the international journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, ICCV, ECCV, CVPR, NIPS, and AAAI. His research interests include computer vision and related fields.