Heterogeneous Face Recognition via Face Synthesis With Identity-Attribute Disentanglement

Ziming Yang[ID], Jian Liang[ID], Chaoyou Fu[ID], Mandi Luo[ID], *Member, IEEE,*
and Xiao-Yu Zhang[ID], *Senior Member, IEEE*

Abstract—**Heterogeneous Face Recognition (HFR) aims to match faces across different domains (*e.g.*, visible to near-infrared images), which has been widely applied in authentication and forensics scenarios. However, HFR is a challenging problem because of the large cross-domain discrepancy, limited heterogeneous data pairs, and large variation of facial attributes. To address these challenges, we propose a new HFR method from the perspective of heterogeneous data augmentation, named Face Synthesis with Identity-Attribute Disentanglement (FSIAD). Firstly, the identity-attribute disentanglement (IAD) decouples face images into identity-related representations and identity-unrelated representations (called attributes), and then decreases the correlation between identities and attributes. Secondly, we devise a face synthesis module (FSM) to generate a large number of images with stochastic combinations of disentangled identities and attributes for enriching the attribute diversity of synthetic images. Both the original images and the synthetic ones are utilized to train the HFR network for tackling the challenges and improving the performance of HFR. Extensive experiments on five HFR databases validate that FSIAD obtains superior performance than previous HFR approaches. Particularly, FSIAD obtains 4.8% improvement over state of the art in terms of VR@FAR=0.01% on LAMP-HQ, the largest HFR database so far.**

Index Terms—**Heterogeneous face recognition, cross-domain, face augmentation, face disentanglement.**

I. INTRODUCTION

IN RECENT years, face recognition has made significant progress with deep convolution neural networks (CNNs) and has been widely applied in real-world scenarios like surveillance and payment [1]–[6]. Face recognition methods

Ziming Yang and Xiao-Yu Zhang are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: yangziming@iie.ac.cn; zhangxiaoyu@iie.ac.cn).

Jian Liang, Chaoyou Fu, and Mandi Luo are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing and the Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: liangjian92@gmail.com; chaoyou.fu@nlpr.ia.ac.cn; luomandi2019@ia.ac.cn).

always assume that face images are captured with visible imaging (VIS) devices [7]. However, this assumption does not hold in many realistic scenarios where face images are captured by different sensors. For instance, near-infrared (NIR) sensors are universally adopted in authentication systems and video surveillance cameras. The large discrepancy between different domains degrades face recognition performance. It arises the demand for heterogeneous face recognition (HFR) that refers to identifying faces across different domains, such as NIR-VIS, Sketch-Photo, and Thermal-VIS. Generally, HFR is confronted with three major challenges: (i) Large cross-domain discrepancy. The large discrepancy between faces in different domains enlarges intra-class distance and thus worsens the performance of HFR [8]. (ii) Lack of heterogeneous face data. It is time-consuming and expensive to collect large-scale heterogeneous face images. The limited number of subjects and small-scale HFR dataset are prone to result in the over-fitting problem [9]. (iii) Large variation on facial attributes. Face images have various facial attributes including pose, complexion, expression, and illumination, which further increase intra-class distance and make it difficult for face matching [10].

Over the last decade, HFR has attracted considerable attention of researchers to developing effective approaches [11], [12]. Existing HFR methods can be divided into three main categories: domain-invariant feature based methods, common subspace learning based methods, and image synthesis based methods. The domain-invariant feature based methods seek to extract the discriminative feature of each subject that is invariant across heterogeneous domains [13]–[15]. The common subspace learning based methods map heterogeneous face images into a subspace for face recognition [8], [16], [17]. The image synthesis based methods refer to the transformation of images from one modality into the others to recognize faces within the same modality [18]–[20]. Previous HFR methods mainly focus on reducing the domain discrepancy and have achieved promising performance. Nevertheless, they have neglected the variations of facial attributes in real-world applications.

To further tackle this issue, it is desirable to utilize image synthesis based methods to synthesize large numbers of images with diverse facial attributes. These synthetic images are used to train HFR models for promoting the generalization of face recognition [21]. Recently, [22] employs data augmentation to generate images with various face deformations for
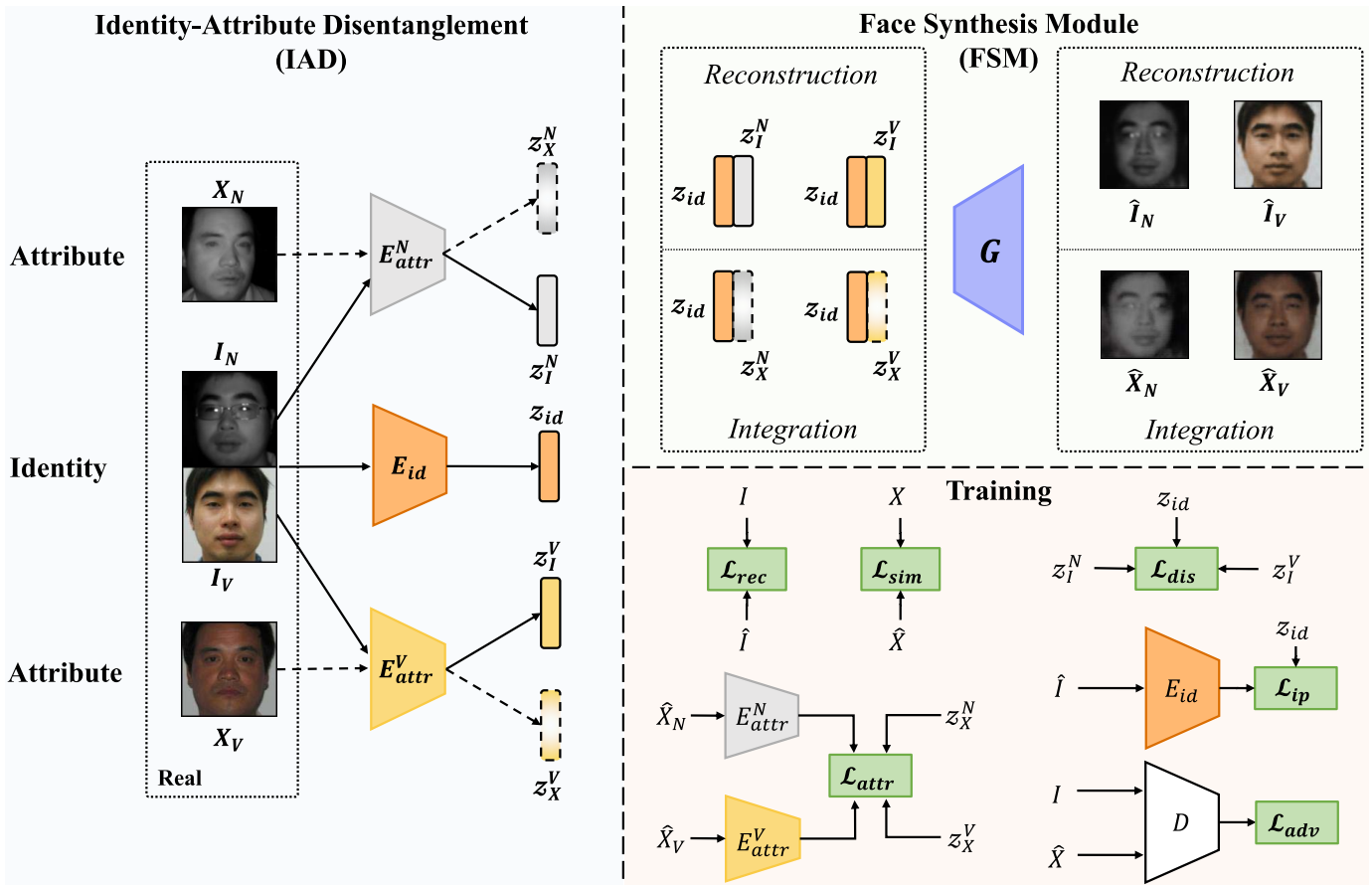
Fig. 1. The pipeline of FSIAD. FSIAD consists of two main components, Identity-attribute disentanglement (IAD) and Face synthesis module (FSM). IAD: Given paired images $I = \{I_N, I_V\}$ and $X = \{X_N, X_V\}$, an identity encoder E_{id} extracts identity representations z_{id} from I. Meanwhile, attribute encoders E_{attr}^N and E_{attr}^V learn facial attribute representations from I and X. Solid and dotted arrows indicate the processing flows of I and X, respectively. FSM: A generator G reconstructs and synthesizes images with different combinations of representations. In the training procedure, a discriminator D supervises G to generate high-fidelity images. The disentanglement loss $\mathcal{L}_{dis}$ makes identities uncorrelated with attributes. The reconstruction loss $\mathcal{L}_{Rec}$ reduces the difference between reconstructed images and the corresponding input. The identity preserving loss $\mathcal{L}_{ip}$ keeps identities of $\hat{I}$ consistent with those of I. The attribute similarity loss $\mathcal{L}_{attr}$ and $\mathcal{L}_{sim}$ minimize the distance between attribute representations and structural features, respectively.

alleviating the negative impacts of facial attribute variations on face recognition. Inspired by [22], we propose a novel method called Face Synthesis with Identity-Attribute Disentangle (FSIAD) to synthesize abundant heterogeneous face images with diverse facial attributes. Rather than randomly sampling from the distributions of heterogeneous data [23], FSIAD disentangles facial identities and attributes, and then generates a large number of images with stochastic combinations of disentangled identities and attributes to augment the raw HFR databases. As depicted in Fig. 1, the scheme of FSIAD contains two components: Identity-Attribute Disentanglement (IAD) and Face Synthesis Module (FSM). First, IAD introduces an identity encoder and attribute encoders to extract the identity-related representations and identity-unrelated representations from images, respectively. These two representations are denoted as facial identities and attributes. To disentangle attributes and identities, we regularize facial attributes to be orthogonal to identities. It decreases the correlation between facial attributes and identities and facilitates attribute representation learning. Second, FSIAD utilizes FSM to synthesize images with the combinations of disentangled identities and attributes. Massive combinations of identities

and attributes significantly increase the diversity in facial attributes of synthetic images. We propose an identity preserving constraint to ensure that the identities of synthetic images are consistent with those of original images. In addition, a discriminator is applied to promote FSIAD to generate high-fidelity images by discriminating between synthetic images and real ones. The large-scale heterogeneous faces synthesized by FSIAD are used for training HFR models to fundamentally supply sufficient heterogeneous faces, learn various facial attributes and reduce the cross-domain discrepancy.

To sum up, the main contributions of this work are as follows:

- To improve the performance of HFR, we propose a Face Synthesis with Identity-Attribute Disentanglement (FSIAD) framework that naturally tackles three major challenges of HFR through data augmentation.
- We devise an Identity-Attribute Disentanglement (IAD) module to decouple facial attributes and identities from face images. The facial attributes are constrained to be orthogonal to identities for decreasing their correlation.
- We introduce a Face Synthesis Module (FSM) that generates large-scale images with the integration of

disentangled facial identities and attributes to augment the HFR databases and enrich the diversity of facial attributes.

II. RELATED WORK

Heterogeneous face recognition methods can be mainly divided into three categories, including domain-invariant feature learning, common subspace learning, and image synthesis. In this section, we review previous HFR methods and the representative generative models for image synthesis.

A. Heterogeneous Face Recognition

1) Domain-Invariant Feature Learning Methods: Domain-invariant feature learning methods aim to extract the identity-related features that are invariant across spectral domains. Conventional methods are mainly supported by hand-crafted features, such as Local Binary Patterns (LBP) [24], Difference-of-Gaussian (DoG) [25], Histograms of Oriented Gradients (HoG) [13], and Local Binary Pattern Histogram (LBPH) [26].

Deep learning has achieved great success in feature learning. Many efforts are devoted to extracting domain-invariant features with deep learning algorithms. The center loss [27] and triple loss [28] are utilized to reduce NIR-VIS discrepancy. [9] proposes a Wasserstein CNN (W-CNN) to capture invariant deep features by minimizing the Wasserstein distance between NIR and VIS features. Based on [9], [29] designs a Disentangled Variational Representation (DVR) framework to disentangle the identity information and within-person variations from heterogeneous faces. [14] and [30] learn global relationships between local features of heterogeneous faces to represent the domain-invariant identity information.

Domain-invariant feature learning methods provide solutions to extract representations that are rarely related to facial identities. However, they are prone to suffer from the overfitting problem due to the small-scale heterogeneous face datasets [31].

2) Common Subspace Learning Methods: Common subspace learning methods project faces of different domains into a compact latent space, where the distance between faces of the same subject is short. Many approaches of dimensionality reduction are used to map heterogeneous features to a low-dimensional space, including Linear Discriminant Analysis (LDA) [32], [33], Principal Component Analysis (PCA) [34], Canonical Correlation Analysis (CCA) [18], [35], and Partial Least Squares (PLS) [36]. The emergence of deep learning has significantly promoted subspace learning. Invariant Deep Representation (IDR) [37] and Coupled Deep Learning (CDL) [38] introduce orthogonal constraint and relevance constraint to learn a shared feature space, respectively. Reference [17] disentangles identity-related, modality-related, and residual features from heterogeneous faces for reducing variations of modality and irrelevance (*e.g.* pose and expression). Extended from [17] and [39] further improves disentanglement with an orthogonal constraint and then learns residual-invariant representations by aligning high-level features of the non-neutral face and neutral face.

The common subspace learning methods intuitively project heterogeneous faces to a shared space through increasing intra-class compactness and inter-class separability. Unfortunately, since these methods require pairs of faces in different modalities for training, their performances are limited by the lack of paired heterogeneous faces.

3) Image Synthesis Methods: Image synthesis methods mainly include conditional and unconditional approaches. For conditional approaches, images are transformed from one modality to another modality for matching faces in the same modality. Face photo-sketch synthesis [40], [41] offers the first insight into face recognition via generation. References [42]–[44] perform cross-spectral image reconstruction with coupled or joint dictionary learning. The advanced deep generative models [45], [46] have made rapid progress in image synthesis. References [47] and [48] extend CycleGAN [49] to handle heterogeneous image transformation. Later, [50] improves [48] by decomposing synthesis into texture inpainting and pose correction. Pose-preserving Cross-spectral Face Hallucination (PCFH) [51] and Pose Aligned Cross-spectral Hallucination (PACH) [52] align the poses and expressions of NIR faces to those of VIS faces for producing paired NIR-VIS faces.

Synthetic images produced from unconditional approaches are allowed to be inconsistent with original images. References [23] and [31] propose a Dual Variational Generation (DVG) framework to learn a joint distribution of paired heterogeneous faces and then generate diverse heterogeneous images from noise. References [53] and [54] employ knowledge distillation to use teacher networks pre-trained on abundant VIS images to train generators with insufficient paired heterogeneous faces. Heterogeneous Face Interpretable Disentangled Representation (HFIDR) [55] explicitly interprets semantic information of face representations to synthesize cross-modality face images. References [56] and [57] synthesize multimodal faces from the predetermined visual descriptions of facial attributes.

Unlike DVR [29] that only disentangles the identity information and within-person variations, we perform disentanglement and image synthesis in a unified framework. Contrary to DVG [23] and DVG-Face [31] that generate the entire faces from noise and introduce external identity information from large-scale VIS faces to enrich the identity diversity of generated faces, FSIAD enrich the diversity in facial attributes of generated faces with the combinations of disentangled representations instead of external information and simultaneously tackle the challenges of HFR. Compared with HFIDR [55] that only interprets disentangled representations of identity and modality, FSIAD can separate representations of identity, modality, and facial attributes. Different from [56] and [57] that synthesize images from limited descriptions of facial attributes, FSIAD automatically extracts facial attributes without manual descriptions for synthesis. Rather than decreasing variations in domains and facial attributes to learn discriminative features [39], we increase variation in facial attributes of synthetic faces to improve the generalization of HFR models to diverse attributes.

In contrast to other categories, the largest advantage of image synthesis based methods is yielding sufficient heterogeneous faces. These methods intuitively solve the lack of paired

HFR data. But there are two challenges of HFR still remain: large cross-domain discrepancy and large variation of facial attributes. The performances of these methods are strongly related to the quality of synthetic images. Meanwhile, face synthesis is an ill-posed problem that there exist multiple solutions for each input [10]. Our method tackles this problem by introducing comprehensive constraints to control the identities and facial attributes of synthetic faces, thereby avoiding the uncertainty of face synthesis.

B. Face Generation

Powered by deep generative models including Generative Adversarial Networks (GANs) [45] and Variational Anto-Encoders (VAEs) [46], many works have achieved outstanding performance on face generation. MUNIT [58] learns content features and style features from images for multimodal image generation with unsupervised learning. Nirkin *et al.* [59] use 3D Morphable face Models (3DMM) for face segmentation and manipulation. Reference [60] introduces a FaceShifter algorithm for face swapping and tackles the occlusion challenge to generate high-fidelity images. References [61]–[63] propose StyleGANs to automatically separate facial styles and stochastic effects, for improving the controllability of face synthesis. Based on [62], Nitzan *et al.* [64] propose a novel face identity disentanglement framework to disentangle identity and attribute with weakly supervised learning. Zhu *et al.* [65] formulate face swapping as an optimal transfer problem, and propose an Appearance Optimal Transfer (AOT) algorithm to synthesize realistic face images with large appearance discrepancies. Daniel and Tamar [66] introduce a Soft-IntroVAE to synthesize high-resolution face images through introspective variational inference. Zhang *et al.* [67] design a multi-identity face reenactment network FReeNet to transfer expressions among different subjects. InfoSwap [68] utilizes the information bottleneck to extract representations of identity and perceptual features for subject-agnostic face swapping.

III. METHOD

In this section, we describe the proposed method – FSIAD in detail. As shown in Fig. 1, FSIAD contains two components: Identity-Attribute Disentanglement (IAD) and Face Synthesis Module (FSM). First of all, IAD is designed to disentangle identity representations and facial attribute representations from face images. The identity representations are supposed to be independent of facial attribute representations. Secondly, FSM aims to produce large-scale heterogeneous faces with disentangled representations for improving the performance of HFR.

Our method takes two pairs of heterogeneous face images $I = \{I_N, I_V\}$ and $X = \{X_N, X_V\}$ as input data, where N and V denote two different spectral domains. It is applicable to NIR-VIS, Thermal-VIS, and Sketch-Photo heterogeneous faces synthesis. On the one hand, the paired images I_N and I_V have the same identity, which is expected to be preserved during face synthesis. On the other hand, images $\{X_N, X_V\}$ are randomly sampled from HFR datasets and provide external attribute features for enriching the diversity in attributes of generated faces.

TABLE I
THE STRUCTURE OF THE ATTRIBUTE ENCODERS E_{attr}^N AND E_{attr}^V

Input	Layer	Kernel/Stride/Padding	Output	Output size
image	Conv1	5 / 1 / 2	x	$32 \times 128 \times 128$
x	Conv1	3 / 2 / 1	x	$64 \times 64 \times 64$
x	Conv1	3 / 2 / 1	x	$128 \times 32 \times 32$
x	Conv1	3 / 2 / 1	x	$256 \times 16 \times 16$
x	Conv1	3 / 2 / 1	x	$512 \times 8 \times 8$
x	Conv1	3 / 2 / 1	x	$512 \times 4 \times 4$
x	FC	-	μ, σ	256, 256

A. Identity-Attribute Disentanglement (IAD)

The key idea of IAD is to decouple identities and attributes from faces. To extract the identity features from face images, we employ a pre-trained face recognition model LightCNN [69] as the identity encoder E_{id}. Latent vector $z_{id} = \frac{1}{2}(E_{id}(I_N) + E_{id}(I_V))$ contains the information merely relevant to facial identity, which is a L2-normalized feature embedding extracted by E_{id}. In addition, we design facial attribute encoders E_{attr}^N and E_{attr}^V based on VAEs as illustrated in Table I. The Conv1 layer comprises a convolution layer, an instance normalization layer, and a Leaky ReLU activation layer. The FC layer denotes a fully connected layer. The parameters of E_{attr}^N and E_{attr}^V are denoted as q_{ϕ_N} and q_{ϕ_V}, respectively. Taking NIR image I_N as input data, attribute encoder E_{attr}^N computes the mean μ_I^N and the standard deviation σ_I^N for approximating the posterior distribution $q_{\phi_N}(z_I^N|I_N) = \mathcal{N}(z_I^N; \mu_I^N, \sigma_I^{N2})$. The attribute representation z_I^N is sampled via the re-parameterization trick, *i.e.*, $z_I^N = \mu_I^N + \sigma_I^N \cdot \epsilon$, where ϵ denotes the random noise that $\epsilon \sim \mathcal{N}(0, I)$. Similarly, z_I^V is sampled from the posterior distribution $q_{\phi_V}(z_I^V|I_V) = \mathcal{N}(z_I^V; \mu_I^V, \sigma_I^{V2})$: $z_I^V = \mu_I^V + \sigma_I^V \cdot \epsilon$.

The training of IAD is divided into two steps, *i.e.*, feature disentanglement and distribution learning. Feature disentanglement aims to decrease the correlation between identity and attribute representations. Hence we introduce a disentanglement objective function $\mathcal{L}_{dis}$ between identity representation z_{id} and attribute representations $\{z_I^N, z_I^V\}$:

$$\mathcal{L}_{dis} = \cos(z_{id}, z_I^N) + \cos(z_{id}, z_I^V), \quad (1)$$

where $\cos(\cdot, \cdot)$ is the cosine similarity function. The identity representation z_{id} is orthogonal to attribute representations z_I^N and z_I^V when $\mathcal{L}_{dis}$ descents to 0. In other words, the learned facial attributes become independent to identity.

The second step is designed for attribute distribution learning. Motivated by VAEs [46] that are optimized through maximizing the evidence lower bound objective (ELBO), we adopt its formulation for our method as follows:

$$\log p_\theta(I_N) \geq \mathbb{E}_{q_{\phi_N}(z_I^N|I_N)} \log p_\theta(I_N|z_I^N)$$
$$- D_{KL}(q_{\phi_N}(z_I^N|I_N)||p(z_I^N)), \quad (2)$$

where the first term denotes the reconstruction objective of I_N, and the second term is a Kullback-Leibler divergence between the approximated posterior distribution $q_{\phi_N}(z_I^N|I_N)$ and the prior distribution $p(z_I^N)$. The ELBO of $p_\theta(I_V)$ can

TABLE II

THE STRUCTURE OF THE GENERATOR G

Input	Layer	Kernel/Stride/Padding	Output	Output size
z_{id}, z_X	FC	-	x	8192
x	TransConv	4 / 2 / 1	x	$256 \times 8 \times 8$
x	TransConv	4 / 2 / 1	x	$128 \times 16 \times 16$
x	TransConv	4 / 2 / 1	x	$64 \times 32 \times 32$
x	TransConv	4 / 2 / 1	x	$32 \times 64 \times 64$
x	TransConv	4 / 2 / 1	x	$32 \times 128 \times 128$
x	Conv1	3 / 1 / 1	x	$32 \times 128 \times 128$
x	Conv1	3 / 1 / 1	x	$3 \times 128 \times 128$
x	Tanh	-	image	$3 \times 128 \times 128$

TABLE III

THE NETWORK ARCHITECTURE OF THE DISCRIMINATOR D

(a) The structure of D.

Input	Layer	Kernel/Stride/Padding	Output	Output size
image	RefPad	-	x	$3 \times 134 \times 134$
x	Conv2	7 / 1 / 0	x	$64 \times 128 \times 128$
x	Conv2	3 / 2 / 1	x	$128 \times 64 \times 64$
x	Conv2	3 / 2 / 1	x	$256 \times 32 \times 32$
x	ResBlock	-	x	$256 \times 32 \times 32$
x	ResBlock	-	x	$256 \times 32 \times 32$
x	ResBlock	-	x	$256 \times 32 \times 32$
x	Sigmoid	-	y	1

(b) The structure of ResBlock.

Input	Layer	Kernel/Stride/Padding	Output	Output size
x	RefPad	-	τ	$256 \times 34 \times 34$
τ	Conv2	3 / 1 / 0	τ	$256 \times 32 \times 32$
τ	RefPad	-	τ	$256 \times 34 \times 34$
τ	Conv2	3 / 1 / 0	τ	$256 \times 32 \times 32$
τ	Skip Connection	-	x	$256 \times 32 \times 32$

be formed by simply replacing N to V of Eq. (2). Therefore, we define the distribution learning objective $\mathcal{L}_{kl}$ to minimize the Kullback-Leibler divergence:

$$\mathcal{L}_{kl} = D_{KL}(q_{\phi_N}(z_I^N|I_N)||p(z_I^N))$$
$$+ D_{KL}(q_{\phi_V}(z_I^V|I_V)||p(z_I^V)), \quad (3)$$

where the prior distributions $p(z_I^N)$ and $p(z_I^V)$ are assumed to obey the multivariate normal distribution $\mathcal{N}(0, I)$.

In brief, the loss function of IAD is formulated as:

$$\mathcal{L}_{IAD} = \lambda_{dis}\mathcal{L}_{dis} + \mathcal{L}_{kl}, \quad (4)$$

where λ_{dis} is a trade-off parameter.

B. Face Synthesis Module (FSM)

As depicted in the previous section, the disentangled identities and attributes are obtained from IAD. We design an FSM to generate images with combinations of disentangled identities and attributes. Apart from the encoders E_{id}, E_{attr}^N and E_{attr}^V shared with IAD, FSM also contains a generator G and a discriminator D. The structures of G and D are shown in TABLE II and III, respectively. The TransConv layer contains a transposed convolution layer, an Adaptive Instance Normalization (AdaIN) [70] layer, a Leaky ReLU activation layer, and a residual block. The Tanh denotes the Tanh activation layer. The RefPad is the reflection padding operation. The Conv2 layer is comprised of a convolution layer, a batch normalization layer, and a ReLU activation layer. The Skip Connection layer adds the result τ to the input x.

The training procedure of FSM is simultaneously conducted through two branches: reconstruction and integration. The first branch aims to reconstruct images with identities and attributes that are derived from the same subjects, $\hat{I}_N = G(z_{id}, z_I^N)$, $\hat{I}_V = G(z_{id}, z_I^V)$. The second branch integrates identities of I and attributes of X to synthesize images $\hat{X}_N = G(z_{id}, z_X^N)$, $\hat{X}_V = G(z_{id}, z_X^V)$, where z_X^N and z_X^V are attribute representations of reference images X_N and X_V, respectively.

1) Reconstruction: The reconstructed images $\{\hat{I}_N, \hat{I}_V\}$ are supposed to be identical to original images $\{I_N, I_V\}$. We use the reconstruction loss $\mathcal{L}_{Rec}$ to reduce the squared Euclidean distances between original images and reconstructed ones:

$$\mathcal{L}_{rec} = \|I_N - \hat{I}_N\|_2^2 + \|I_V - \hat{I}_V\|_2^2. \quad (5)$$

2) Integration: In order to preserve identities of synthetic images, we extract the identity representations from synthetic images $\{\hat{I}_N, \hat{I}_V\}$ and $\{\hat{X}_N, \hat{X}_V\}$. Then we utilize identity preserving objective $\mathcal{L}_{ip}$ to increase the similarity between identity representations. $\mathcal{L}_{ip}$ is formulates as:

$$\mathcal{L}_{ip} = \|z_{id} - E_{id}(\hat{I}_N)\|_2^2 + \|z_{id} - E_{id}(\hat{I}_V)\|_2^2$$
$$+ \|z_{id} - E_{id}(\hat{X}_N)\|_2^2 + \|z_{id} - E_{id}(\hat{X}_V)\|_2^2. \quad (6)$$

In addition, the facial attributes of synthesized images $\{\hat{X}_N, \hat{X}_V\}$ are expected to be similar to the reference images $\{X_N, X_V\}$. We take both latent representations and image contents into consideration in terms of similarity. From the perspective of latent representations, the distances between attribute representations of $\{X_N, X_V\}$ and those of $\{\hat{X}_N, \hat{X}_V\}$ ought to be short. Similar to $\mathcal{L}_{ip}$, we define the attribute similarity objective $\mathcal{L}_{attr}$ as:

$$\mathcal{L}_{attr} = \|z_X^N - \hat{z}_X^N\|_2^2 + \|z_X^V - \hat{z}_X^V\|_2^2, \quad (7)$$

where $\hat{z}_X^N$ and $\hat{z}_X^V$ are attribute representations of $\hat{X}_N$ and $\hat{X}_V$.

Attribute-related contents including pose, complexion, and illumination, are transferred from input images into the synthesized ones. The similarity of contents can be measured from two aspects: error metric [71] and structural similarity. Error metrics are prevailing approaches to evaluate the difference between images, which are calculated between images pixel by pixel. Owing to the advantage of color and illumination consistency, L_1-norm loss function is a predominant error metric to regularize networks for image generation. However, networks tend to generate a complete copy of the targeted image when only constrained by error metrics. Besides, error metrics are not sensitive to the pixel dependencies that contain critical structural features of objects in the visual scene [72]. It results in the low perceptual quality of synthetic images. Therefore, we employ the Multi-Scale Structural Similarity index (MS-SSIM) [72] as structural similarity loss to increase the structural similarity between input images and synthesized

images. Specifically, MS-SSIM aims to measure the difference between structural information of images, which is related to the attributes of objects in the view and independent to extrinsic factors such as illumination, contrast, and noise [71]. MS-SSIM is highly adapted for human visual perception and is generally used for image synthesis. The higher MS-SSIM means the smaller difference between objects in different images. We propose a similarity objective $\mathcal{L}_{sim}$ with an combination of a L_1-norm loss and MS-SSIM[1]:

$$\mathcal{L}_{sim} = (1 - \alpha) \cdot \|X - \hat{X}\|_1 + \alpha \cdot \mathcal{L}_{MS}(X, \hat{X}), \qquad (8)$$

where $\mathcal{L}_{MS}(\cdot, \cdot) = (1 - MS\text{-}SSIM(\cdot, \cdot))$, $\hat{X}$ means the concatenation of synthesized images, *i.e.*, $\hat{X} = [\hat{X}_N, \hat{X}_V]$. α is a trade-off hyperparameter for tuning the ratio of $\mathcal{L}_{MS}$ to L_1-norm loss, which is set to 0.84 according to the empirical investigation [73].

According to Eq. (6-8), we summarize the loss function of integration $\mathcal{L}_{Int}$ as follows:

$$\mathcal{L}_{int} = \mathcal{L}_{ip} + \mathcal{L}_{attr} + \mathcal{L}_{sim}. \qquad (9)$$

For the sake of further improvement on image synthesis, generator G is optimized in an adversarial manner. We impose a discriminator D to differentiate real images from synthetic images that generated by G. Meanwhile, a generator G seeks to generate photo-realistic images for confusing D. The adversarial loss $\mathcal{L}_{adv}$ is formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_{I \sim \mathbb{P}_{data}(I)}[\log D(I)] + \mathbb{E}_{\hat{X} \sim \mathbb{P}_G(\hat{X})}[\log(1 - D(\hat{X}))]. \qquad (10)$$

Hereby, we define loss function of FSM as $\mathcal{L}_{FSM}$:

$$\mathcal{L}_{FSM} = \mathcal{L}_{rec} + \lambda_{int}\mathcal{L}_{int} + \lambda_{adv}\mathcal{L}_{adv}, \qquad (11)$$

where λ_{int} and λ_{adv} are trade-off parameters. According to Eq. (4, 11), the overall loss of FSIAD is summarized as:

$$\mathcal{L}_{all} = \mathcal{L}_{IAD} + \mathcal{L}_{FSM}. \qquad (12)$$

C. Heterogeneous Face Recognition

We utilize FSIAD to synthesize a large number of heterogeneous faces for data augmentation. The HFR network F learns abundant features of facial attributes from synthetic faces and alleviates the degradation caused by variations in pose, expression, and other extrinsic factors. Both real images and augmented images are jointly used to train the HFR network.

Basically, HFR network is trained on a pair of real images $\{I_N, I_V\}$. We define HFR objective $\mathcal{L}_{ce}$ with the cross-entropy loss to measure the classification error:

$$\mathcal{L}_{ce} = -\sum_i^n \sum_{M \in \{N, V\}} y_i \log(\text{softmax}(F(I_M^i))), \qquad (13)$$

where n is the number of paired real images, y_i is the identity label of i-th paired real images $\{I_N^i, I_V^i\}$.

[1]https://github.com/VainF/pytorch-msssim

In addition, the HFR network is supposed to diminish the intra-class discrepancies and extract discriminative representations. The augmented images are applied to reduce discrepancies in domains and attributes. Since paired augmented images $\{\tilde{X}_N, \tilde{X}_V\}$ are generated from the same subject, the identities of $\{\tilde{X}_N, \tilde{X}_V\}$ are expected to be closed to each other [23]. To shorten the intra-class distance, we formulate the intra-class loss $\mathcal{L}_{in}$ as:

$$\mathcal{L}_{in} = \|F(\tilde{X}_N) - F(\tilde{X}_V)\|_2^2. \qquad (14)$$

On the whole, the loss function for training HFR network is defined as:

$$\mathcal{L}_{HFR} = \mathcal{L}_{ce} + \gamma \mathcal{L}_{in}, \qquad (15)$$

where γ is a trade-off parameter.

For comprehensive elaboration of our work, we introduce the generic training strategy in algorithm 1.

Algorithm 1 Training Strategy of FSIAD

Input: Source images $I = \{I_N, I_V\}$. Reference images $X = \{X_N, X_V\}$. A pre-trained identity encoder E_{id}. Attribute encoders E_{attr}^N, E_{attr}^V. A generator G. A discriminator D. A heterogeneous face recognition network F. The number of synthesized images n.

Output: The parameters of E_{attr}^N, E_{attr}^V, G, D and F: ϕ_N, ϕ_V, Ψ_G, Ψ_D and Θ.

1: **for** $i = 1$ to T **do**
2:   // IAD component.
3:   $z_{id} = \frac{1}{2}(E_{id}(I_N) + E_{id}(I_V))$.
4:   $z_I^N = E_{attr}^N(I_N)$; $z_I^V = E_{attr}^V(I_V)$.
5:   $z_X^N = E_{attr}^N(X_N)$; $z_X^V = E_{attr}^V(X_V)$.
6:   Update ϕ_N, ϕ_V by Eq. (4).
7:   // FSM component.
8:   Reconstruction: $\hat{I}_N = G(z_{id}, z_I^N)$; $\hat{I}_V = G(z_{id}, z_I^V)$.
9:   Integration: $\hat{X}_N = G(z_{id}, z_X^N)$; $\hat{X}_V = G(z_{id}, z_X^V)$.
10:   Fix ϕ_N, ϕ_V, Ψ_G:
11:     Update Ψ_D by Eq. (10).
12:   Fix Ψ_D:
13:     Update ϕ_N, ϕ_V, Ψ_G by Eq. (9).
14: **end for**
15: // HFR network.
16: Initialize Θ by a pre-trained model.
17: **for** $iteration = 1$ to n **do**
18:   Randomly sample $I = \{I_N, I_V\}$ and $X = \{X_N, X_V\}$.
19:   $z_{id} = \frac{1}{2}(E_{id}(I_N) + E_{id}(I_V))$.
20:   $z_X^N = E_{attr}^N(X_N; \phi_N)$; $z_X^V = E_{attr}^V(X_V; \phi_V)$.
21:   $\tilde{X}_N = G(z_{id}, z_X^N; \Psi_G)$; $\tilde{X}_V = G(z_{id}, z_X^V; \Psi_G)$.
22:   Update Θ by Eq. (15).
23: **end for**
24: **return** Θ, ϕ_N, ϕ_V, Ψ_G, Ψ_D.

IV. EXPERIMENTS

In this section, we perform extensive experiments to evaluate our proposed FSIAD qualitatively and quantitatively

on five heterogeneous face recognition databases, including CASIA NIR-VIS 2.0 [74], BUAA-VisNir [75], Oulu-CASIA NIR-VIS [76], Tufts Face [77], and LAMP-HQ [10]. To demonstrate the superior advantage of our proposed FSIAD, we make comparisons with state-of-the-art methods. Finally, ablation studies on the effectiveness of different loss functions are conducted.

A. Databases and Protocols

1) CASIA NIR-VIS 2.0: the most challenging public HFR database. It contains 725 subjects and each subject has 5~50 NIR and 1~22 VIS images with large within-class variations including pose, expression, and illumination. We conduct experiments with 10-fold cross-validation. For each fold, the training set consists of about 6,100 NIR and 2,500 VIS images from 360 identities. For evaluation, the testing set is composed of over 6,000 NIR and 358 VIS images from 358 identities who are excluded from those in the training set.

2) BUAA-VisNir: a standard HFR database. It contains 150 subjects and each subject has 9 pairs of NIR-VIS images, including one frontal view, four different other views, and four expressions. Since the NIR and VIS images are captured simultaneously with a multi-spectral sensor, paired NIR-VIS images are identical except for the spectral domain. The training set consists of 900 images from 50 subjects and the testing set consists of 1,800 images from the remaining 100 subjects.

3) Oulu-CASIA NIR-VIS: contains 80 subjects. The paired NIR-VIS images of each subject are captured with six different expressions (anger, disgust, fear, happiness, sadness, and surprise) and three different illuminations (normal indoor, weak, and dark). Following the protocol introduced in [78], we select 40 subjects and each subject has 48 paired NIR-VIS images as the training set and testing set. Both the training set and the testing set comprise 20 subjects.

4) Tufts Face: a Thermal-VIS HFR database. It contains 1,582 paired Thermal-VIS images of 113 subjects. Each subject has 14 pairs of Thermal-VIS images with 9 different yaw angles and 5 different expressions. Since this database has not designed a protocol for evaluation, we divide it into a training set with 85 subjects and a testing set with the rest 28 subjects. The training set is composed of 1,190 pairs of Thermal-VIS images, while the testing set has 28 VIS gallery images and 275 thermal probe images.

5) LAMP-HQ: the latest and largest HFR database. It is featured with large-scale, high-resolution, and wide-diversity (*e.g.*, age, race, and accessories), which contains 56,788 NIR and 16,828 VIS images from 573 subjects. Each subject has 66 paired NIR-VIS images that are captured with three distinct expressions and three yaw angles in five illumination scenes. The evaluation is conducted by 10-fold experiments. For each fold, the training dataset consists of approximately 29,000 NIR and 8,800 VIS images from about 300 subjects. For testing, the gallery set contains the remaining 273 individuals and each individual has one VIS image, while the probe set has about 27,000 NIR images from the same individuals.

B. Experimental Settings

Implementation Details: We firstly transform the size of input images to 128×128. A face recognition model LightCNN [69] pre-trained on the MS-Celeb-1M dataset [79] is adopted as the identity encoder E_{id} to extract identity features from face images. The L_2 normalized latent vector output from the penultimate fully connected layer in LightCNN is utilized as identity feature representation. The attribute encoders E_{attr_N}, E_{attr_V}, generator (or decoder) G, and discriminator D are built with the architectures that are described in TABLE I, II and III, respectively. Our proposed network is implemented with the deep learning framework Pytorch. One Nvidia RTX GPU is used for acceleration. For optimization on training, we employ an Adam [80] optimizer that is configured with a learning rate of 2e-4, coefficients $\beta_1=0.5$ and $\beta_2=0.99$ that are used to compute the running averages of gradient and its square. The trade-off parameters λ_{dis}, λ_{int}, λ_{adv}, and γ in Eq.(4,11,15) are set to 2, 5, 1, and 0.001, respectively.

For fair comparisons, we follow the protocols of [23], [29], [52], [55], and [9] to use LightCNN [69] as the HFR model. The pre-trained LightCNN model is imported to initialize the parameters of the network. In the experiments, we utilize FSIAD to synthesize 100,000 pairs of heterogeneous faces as augmented data. Then we fine-tune the LightCNN network with the combination of the original HFR dataset and augmented data. For optimization of fine-tuning, we use the Stochastic Gradient Descent algorithm (SGD) with the following configurations: momentum factor=0.9, initial learning rate=1e-3, and weight decay=1e-4. To tune the hyperparameters of our proposed loss functions, we split the training data of each HFR database into a training set and a validation set by a ratio of 9:1.

C. Qualitative Analyses

In order to analyze the effectiveness of face augmentation, we conduct qualitative experiments on the CASIA NIR-VIS 2.0 database [74] to evaluate our proposed FSIAD with the comparison of state-of-the-art methods MUNIT [58] and FaceShifter [60]. Both the disentanglement and generation of face augmentation are evaluated in the qualitative experiments. We firstly decouple face images into identity representations and attribute representations. Then we produce images of 128×128 resolution with stochastic combinations of identities and attributes. The results generated by the aforementioned methods are displayed in Fig. 2. Intuitively, although MUNIT produces clear results, it fails to disentangle identities from attributes. The synthetic images of MUNIT are almost equal to the source images and have low similarity in facial attributes compared with the reference images. Additionally, we can observe that FaceShifter can swap identity features from the subject of the source image to that of the reference image. However, the results of FaceShifter are mixed with artifacts and lack the smoothness of facial textures. Although FaceShifter [60] has gained outstanding performance of face generation on the large-scale VIS databases, the insufficient HFR data hinders it from generating high-quality

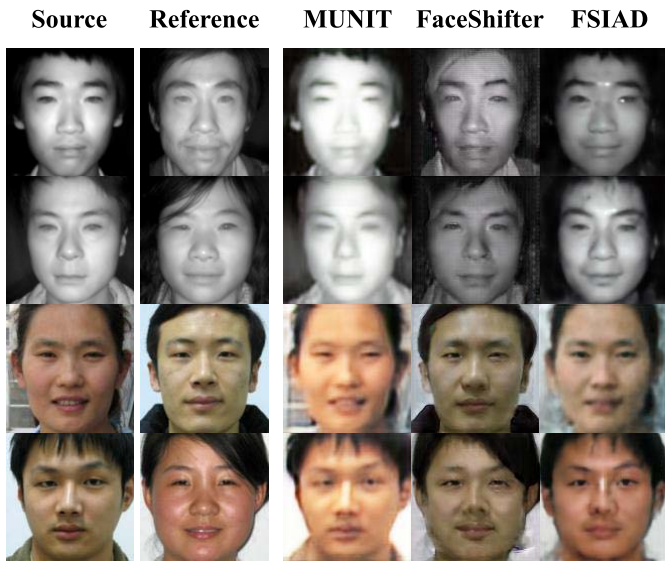| Source	Reference	MUNIT	FaceShifter	FSIAD

Fig. 2. Experimental results on feature disentanglement and generation with FSIAD and state-of-the-art methods MUNIT and FaceShifter.

Fig. 3. An example of face reconstruction on the CASIA NIR-VIS 2.0 database.

heterogeneous faces. In contrast to these methods, FSIAD achieves better performance in feature disentanglement and generation. For one thing, the identities of source images are disentangled from other facial attributes, which are well preserved in the synthetic images. For another, our method can well blend the identities with facial attributes and prevent the emergence of artifacts and rough textures. Moreover, facial attributes such as expression, pose, and complexion are transferred from reference images to the generated images.

For further analysis, we visualize the results of reconstruction and integration to demonstrate the ability of our method. Fig. 3 illustrates that the reconstructed images are high-quality and realistic. Besides, we synthesize images with the integration of identity representations and attribute representations. As shown in Fig. 4, the synthesized results maintain the identities of source images and have similar facial attributes with reference images. In the meantime, we explore cross-domain synthesis that source and reference images are derived from different spectral domains. Fig. 5 reveals that FSIAD is also applicable to cross-domain synthesis and has outstanding abilities in feature disentanglement and face generation. Note that the source images are different from reference images in both spectral domains and subjects, which raises the difficulty in face synthesis.
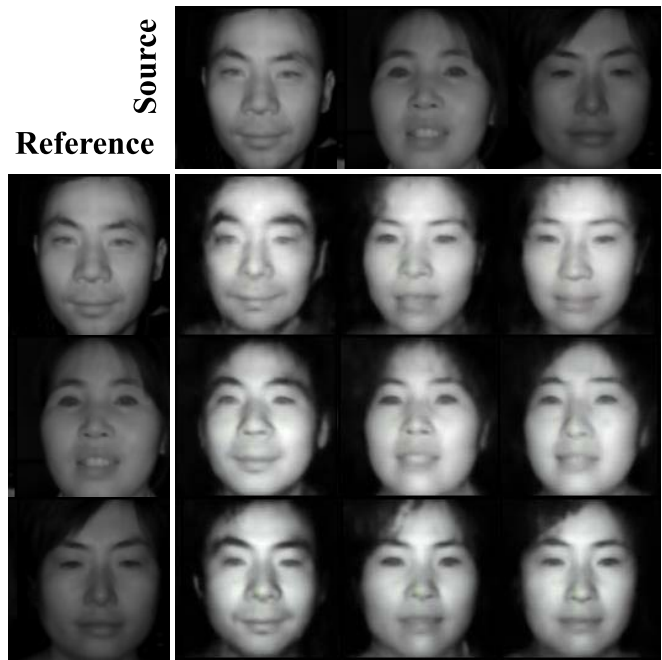

Fig. 4. A grid of synthetic faces. The images in the top row and left most column are source images and reference images, respectively. The rest images are synthesized by FSIAD with identities of source images and attributes of reference images.
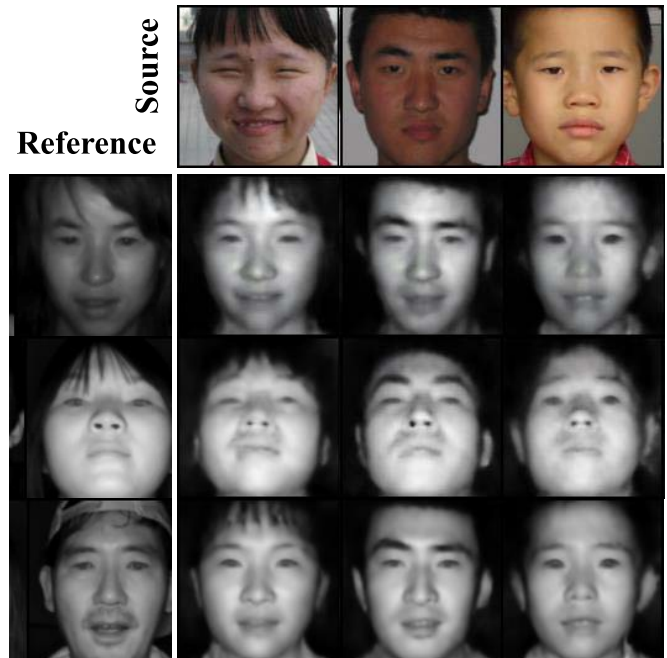

Fig. 5. The visual results of cross-domain synthesis. In contrast to Fig. 4, the source images and reference images are not only derived from different subjects, but also belong to different spectral domains.

D. Quantitative Analyses

Quantitative analyses are carried out to evaluate the performance of our proposed method with comparison to the state-of-the-art methods. In this section, quantitative analyses are divided into three parts, including quantitative evaluations on

TABLE IV

THE FRÉCHET INCEPTION DISTANCES BETWEEN DISTRIBUTIONS OF
REAL IMAGES ON THE CASIA NIR-VIS 2.0 DATASET AND THE
SYNTHETIC IMAGES GENERATED BY DIFFERENT METHODS

Method	FID↓	
	NIR	VIS
MUNIT[58]	17.4735	17.5492
FaceShifter[60]	11.5227	9.0366
FSIAD	**3.2995**	**2.9839**

TABLE V

THE QUANTITATIVE RESULTS OF ATTRIBUTE SIMILARITY ON THE
CASIA NIR-VIS 2.0 DATASET. SSIM DENOTES THE
STRUCTURAL SIMILARITY INDEX

Method	SSIM↑	
	NIR	VIS
MUNIT[58]	0.2572	0.1310
FaceShifter[60]	0.2294	0.1623
FSIAD	**0.4139**	**0.3617**

TABLE VI

THE INFERENCE SPEEDS OF DIFFERENT METHODS. THE
NUMBER OF SYNTHESIZED IMAGES IS 12,480

Method	Time(s)↓	Speed(FPS)↑
MUNIT[58]	65.4009	190.823
FaceShifter[60]	415.9640	30.002
FSIAD	**36.7743**	**339.367**

image synthesis, heterogeneous face recognition experiments and ablation studies.

1) Quantitative Evaluations on Image Synthesis: We conduct quantitative evaluations to compare the effectiveness and efficiency of MUNIT [58], FaceShifter [60] and our proposed FSIAD. Three metrics are employed for evaluations: Fréchet Inception Distance (FID) [81], Structural Similarity Index (SSIM) [71] and inference speed. FID is a metric to measure the distance between the distribution of real images and that of synthetic images. The lower FID reveals that the model achieves better performance to synthesize high-fidelity images like real images. In the experiments, 10,000 pairs of NIR-VIS images are randomly sampled from the CASIA NIR-VIS 2.0 dataset. Each pair contains source images $\{I_N, I_V\}$ and reference images $\{X_N, X_V\}$. Then 10,000 pairs of NIR-VIS synthetic images $\{\hat{X}_N, \hat{X}_V\}$ are generated by the aforementioned methods. Furthermore, we utilize a pre-trained Inception V3 network [82] to map the real images and the synthetic images into 192-dimensional feature vectors for computing FID [83]. As illustrated in TABLE IV, our proposed method achieves the best FID values compared with MUNIT and FaceShifter. In other words, our method is preferable to generate realistic and high-quality images for face augmentation.

Apart from FID, we quantitatively evaluate the performance of feature disentanglement. To assess the similarity of facial attributes between references and synthetic images, we employ the SSIM to measure similarity in structural information between images. Structural information is defined as the attributes of objects in the scene, which is highly adapted for human visual perception [71]. We measure the SSIM score between references and synthetic images. The higher SSIM score suggests that facial attributes of synthetic images are more similar to those of references. As depicted in TABLE V, FSIAD performs better than state-of-the-art methods and achieves the best ability in feature disentanglement.

To further assess the efficiency of the aforementioned methods, we measure the time consumed by synthesizing 12,480 pairs of images with these methods. Inference speed is measured by the frames per second (FPS) score that equals a ratio of the number of synthesized images to the time spent on synthesis. As shown in TABLE VI, the proposed FSIAD is the most efficient method and particularly outperforms FaceShifter by 11 times.

2) Heterogeneous Face Recognition Experiments: In order to validate the effectiveness of FSIAD on heterogeneous

face recognition, we conduct extensive experiments on five HFR databases, including CASIA NIR-VIS 2.0 [74], BUAA-VisNir [75], Oulu-CASIA NIR-VIS [76], Tufts Face [77] and LAMP-HQ [10]. Our method is compared with the state-of-the-art methods TRIVET [28], IDR [37], ADFL [48], W-CNN [9], PACH [52], DVR [29], DVG [23], HFIDR [55], and OMDRA [39]. Since our method utilizes LightCNN [69] as HFR model, the LightCNN only trained with original HFR dataset is selected as a baseline method. Interestingly, we also train LightCNN [69] with the synthetic face images generated by MUNIT [58] and FaceShifter [60] to explore the effectiveness of these two methods. We denote LightCNN models trained with synthetic faces generated by these two methods as LightCNN+MUNIT and LightCNN+FS, respectively. The experiments are conducted by following the protocols as described in Section IV-A. The specific experimental analyses are provided as follows.

a) CASIA NIR-VIS 2.0 database: We conduct 10-fold cross validation on the CASIA NIR-VIS 2.0 database. The Rank-1 accuracy, Verification Rate(VR)@False Accept Rate(FAR)=0.1% and VR@FAR=0.01% are employed for testing. TABLE VII and Fig. 6(a) show that most methods gain satisfactory performance of HFR and achieve Rank-1 accuracy over 90%. Although HFIDR(LightCNN-9) [55] has the lowest Rank-1 accuracy, it can dramatically improve by simply replacing the backbone from LightCNN-9 to LightCNN-29. Compared with TRIVET [28], IDR [37], LightCNN [69], and ADFL [48], advanced methods including W-CNN [9], DVR [29], HFIDR [55], OMDRA [39], PACH [52], and DVG [23] effectively address the over-fitting problem and obtain better performance. Our proposed FSIAD outperforms the state-of-the-art methods and demonstrates its superiority. FSIAD surpasses PACH by 1.6% of VR@FAR=0.1%, which reveals unconditional face generation provides sufficient heterogeneous faces and fundamentally deals with a key obstacle
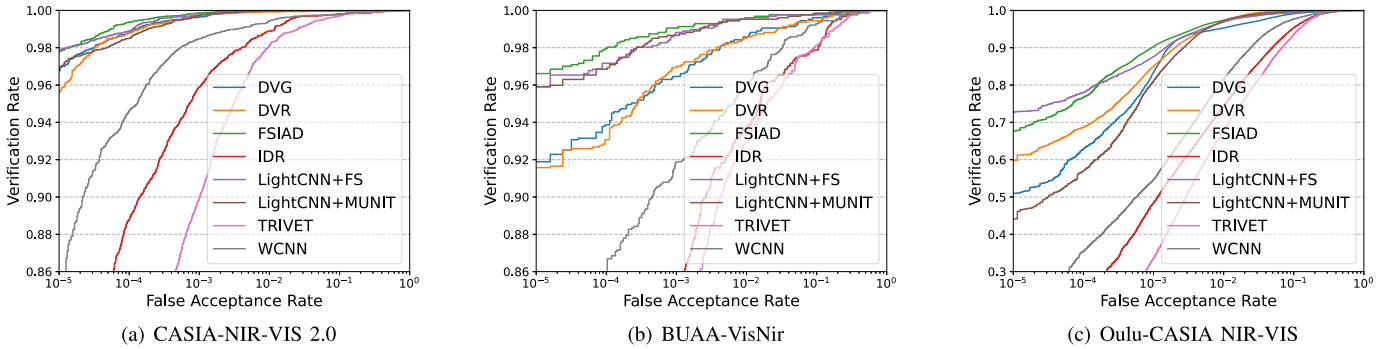
Fig. 6. The ROC curves of different methods on the CASIA-NIR-VIS 2.0, BUAA-VisNir, and Oulu-CASIA NIR-VIS datasets.

TABLE VII
THE 10-FOLD EXPERIMENTAL RESULTS OF RECOGNITION
ON THE CASIA NIR-VIS 2.0 DATABASE

Method	Rank-1 (%)	VR@FAR =0.1%(%)	VR@FAR =0.01%(%)
TRIVET[28]	95.7±0.5	91.0±1.3	74.5±0.7
IDR[37]	97.3±0.4	95.7±0.7	-
ADFL[48]	98.2±0.3	97.2±0.5	-
W-CNN[9]	98.7±0.3	98.4±0.4	94.3±0.4
PACH[52]	98.9±0.2	98.3±0.2	-
DVR[29]	99.7±0.1	99.6±0.3	98.6±0.3
DVG[23]	99.8±0.1	99.8±0.1	98.8±0.2
HFIDR(LightCNN-9)[55]	87.5±0.0	-	-
HFIDR(LightCNN-29)[55]	98.6±0.0	-	-
OMDRA[39]	99.6±0.1	99.4±0.2	-
LightCNN[69]	96.7±0.2	94.8±0.4	88.5±0.2
LightCNN+MUNIT[58]	99.6±0.0	99.5±0.0	98.1±0.1
LightCNN+FS[60]	99.7±0.0	99.6±0.0	98.1±0.2
FSIAD	**99.9±0.0**	**99.9±0.0**	**99.2±0.1**

TABLE VIII
THE RESULTS OF RECOGNITION ON THE BUAA-VISNIR DATABASE

Method	Rank-1 (%)	VR@FAR =1%(%)	VR@FAR =0.1%(%)
TRIVET[28]	93.9	93.0	80.9
IDR[37]	94.3	93.4	84.7
ADFL[48]	95.2	95.3	88.0
W-CNN[9]	97.4	96.0	91.9
PACH[52]	98.6	98.0	93.5
DVR[29]	99.2	98.5	96.9
DVG[23]	99.3	98.5	97.3
OMDRA[39]	**100.0**	**99.9**	**99.7**
LightCNN[69]	96.5	95.4	86.7
LightCNN+MUNIT[58]	99.4	99.5	98.7
LightCNN+FS[60]	99.5	99.5	98.8
FSIAD	99.8	99.7	99.1

of HFR. Furthermore, FSIAD outperforms DVG [23] by 0.4% in terms of VR@FAR=0.01%. It shows that enriching the diversity in facial attributes is conducive to improving HFR. We also find that synthetic images generated by MUNIT [58] and FaceShifter [60] facilitate the performance of LightCNN and exceed state-of-the-art methods OMDRA [39] and HFIDR [55]. However, the results of LightCNN+MUNIT and LightCNN+FS fail to surpass DVG due to the poor capabilities of face generation and feature disentanglement. Moreover, FSIAD has the smallest standard deviation in the performance of HFR among state-of-the-art methods. Experimental results substantiate that our method effectively tackles the challenges of HFR and facilitates the performance of HFR.

b) BUAA-VisNir database: The Rank-1 accuracy, VR@FAR=1%, and VR@FAR=0.1% are used for testing on the BUAA-VisNir database. As illustrated in TABLE VIII and Fig. 6(b), we observe the results of most state-of-the-art methods including TRIVET [28], IDR [37], ADFL [48], W-CNN [9], PACH [52], DVR [29], and DVG [23] are lower than 99% in terms of VR@FAR=1%. The possible reason is that BUAA-VisNir is a small-scale HFR dataset and contains images of limited subjects. These methods except DVG suffer from the over-fitting problem and

get unsatisfactory performance. Although DVG generates sufficient heterogeneous faces as auxiliary training data, it does not take various facial appearances into consideration and leads to degradation due to variations in the pose and expression of faces on this dataset. LightCNN+MUNIT [58], LightCNN+FS [60] and the proposed FSIAD enrich the diversity of facial attributes in generated images and achieve outstanding performance, whose results are higher than 99.5% in terms of VR@FAR=1%. OMDRA [39] learns residual-independent identity representations and obtains the best results. The results of FSIAD are close to those of OMDRA. FSIAD reaches 99.7% in terms of VR@FAR=1%, which is second to OMDRA (99.9%).

c) Oulu-CASIA NIR-VIS database: The Rank-1 accuracy, VR@FAR=1%, and VR@FAR=0.1% are used for testing. TABLE IX and Fig. 6(c) report that all methods confront degradation of performance on the Oulu-CASIA NIR-VIS dataset compared with CASIA NIR-VIS 2.0 and BUA-VisNir datasets. Since this dataset contains more types of expressions and illuminations but has fewer subjects than those of the former two datasets, it makes HFR more difficult. The results of TRIVET [28], IDR [37], ADFL [48], W-CNN [9] are lower than 80% in terms of VR@FAR=0.1%. These methods are weak in generalization for HFR and fail to

TABLE IX

THE RESULTS OF RECOGNITION ON THE OULU-CASIA NIR-VIS DATABASE

Method	Rank-1 (%)	VR@FAR =1%(%)	VR@FAR =0.1%(%)
TRIVET[28]	92.2	67.9	33.6
IDR[37]	94.3	73.4	46.2
ADFL[48]	95.5	83.0	60.7
W-CNN[9]	98.0	81.5	54.6
DVR[29]	100.0	97.2	84.9
PACH[52]	100.0	97.9	88.2
DVG[23]	100.0	97.5	90.6
HFIDF(LightCNN-9)[55]	100.0	-	-
HFIDF(LightCNN-29)[55]	100.0	-	-
OMDRA[39]	100.0	**98.5**	**92.2**
LightCNN[69]	96.7	92.4	65.1
LightCNN+MUNIT[58]	100.0	97.2	81.0
LightCNN+FS[60]	100.0	97.0	87.6
FSIAD	**100.0**	98.1	92.0

TABLE X

THE RESULTS OF RECOGNITION ON THE TUFTS FACE DATABASE

Method	Rank-1(%)	VR@FAR=1%(%)
LightCNN[69]	15.3	6.1
DVG[23]	51.6	31.6
FSIAD	**57.1**	**35.7**

TABLE XI

THE 1-FOLD EXPERIMENTAL RESULTS ON THE LAMP-HQ DATABASE. RESULTS OF ADFL [48] AND PACH [52] ARE CITED FROM [10]

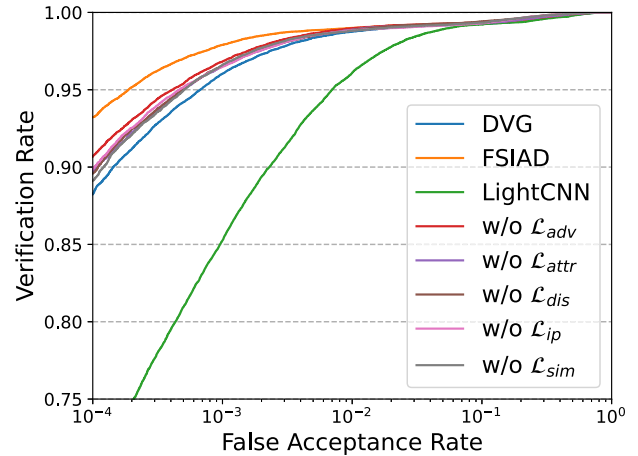| Method	Rank-1 (%)	VR@FAR =1%(%)	VR@FAR =0.1%(%)	VR@FAR =0.01%(%)
LightCNN[69]	96.2	96.1	85.3	69.3
ADFL[48]	95.8	91.5	71.0	-
PACH[52]	96.9	93.9	78.7	-
DVG[23]	98.3	98.8	96.0	88.2
FSIAD	**98.8**	**99.1**	**97.9**	**93.0**

Fig. 7. The ROC curves of LightCNN, DVG, the proposed FSIAD and its five variants on the LAMP-HQ database.

solve the difficulty of insufficient subjects on this dataset. Besides, DVR [29], PACH [52], and DVG [23] have not considered feature disentanglement and can not adapt to variations in facial attributes for better performance. HFIDF [55], OMDRA [39], LightCNN+MUNIT [58], LightCNN+FS [60] and FSIAD achieve prominent results and gain 100% in terms of Rank-1 accuracy. It is notable that LightCNN+MUNIT and LightCNN+FS improve LightCNN [69] from 65.1 to 81.0 and 87.6 in terms of VR@FAR=0.1%, respectively. This significant improvement indicates that the diverse facial attributes of augmented heterogeneous faces exert a positive influence on HFR. OMDRA achieves the best performance among other methods. FSIAD obtains comparable results to those of OMDRA, where the difference of VR@FAR=0.1% is 0.2%.

d) Tufts face database: We compare our method with LightCNN and DVG on the Tufts Face database. TABLE X demonstrates that it is challenging to match faces between thermal and VIS domains due to the lack of facial textures and geometries and the large cross-domain discrepancy. Owing to the small-scale training data, LightCNN [69] gets poor results on the Tufts Face dataset and reaches 15.3% in terms of Rank-1 accuracy. Benefited from synthetic faces, DVG [23] increases the Rank-1 accuracy to 51.6%. It is noted that there are variations in expressions, angles, and accessories such as sunglasses. FSIAD further improves the performance of HFR and reaches 57.1% in terms of Rank-1 accuracy. The results

reveal that augmentation of facial attributes is conducive to Thermal-VIS face recognition.

e) LAMP-HQ: We conduct 1-fold and 10-fold experiments on the LAMP-HQ database to validate the effectiveness of our method. Four metrics are used for evaluations, including Rank-1 accuracy, VR@FAR=1%, VR@FAR=0.1% and VR@FAR=0.01%. The results of the 1-fold experiment are reported in TABLE XI and Fig. 7. Conditional synthesis based methods ADFL [48] and PACH [52] have poor generalization abilities to deal with various facial attributes and get results close to those of baseline LightCNN [69]. DVG [23] takes the advantage of unconditional face generation to synthesize sufficient heterogeneous faces as training data. Unfortunately, since DVG is not able to generate faces with diverse facial attributes, large variations in facial attributes hinder the enhancement of its performance. DVG exceeds LightCNN by 2.1% and 2.7% in terms of Rank-1 accuracy and VR@FAR=1%, respectively. FSIAD has made significant progress and outperforms all the state-of-the-art methods. Especially, FSIAD improves Rank-1 accuracy and VR@FAR=0.01% from 98.3 to 98.8 and from 88.2 to 93.0, respectively. These results prove that the diversity in facial attributes of training data has a critical impact on the improvement of HFR.

The results of 10-fold experiments are presented in TABLE XII and Fig. 8. We observe that conditional synthesis based methods ADFL [48] and PACH [52] have similar performance with LightCNN [69]. It reveals the limited

TABLE XII

THE 10-FOLD EXPERIMENTAL RESULTS ON THE LAMP-HQ DATABASE. RESULTS OF ADFL [48] AND PACH [52] ARE CITED FROM [10]

Method	Rank-1 (%)	VR@FAR =1%(%)	VR@FAR =0.1%(%)	VR@FAR =0.01%(%)
LightCNN[69]	95.8±0.1	95.5±0.3	82.4±2.3	62.5±10.4
ADFL[48]	95.1±0.5	92.1±0.9	73.3±2.2	-
PACH[52]	95.4±0.5	93.1±0.4	75.3±1.7	-
DVG[23]	98.3±0.1	99.0±0.1	96.4±0.2	88.6±1.5
FSIAD	**98.7±0.1**	**99.2±0.1**	**97.3±0.2**	**92.6±1.2**
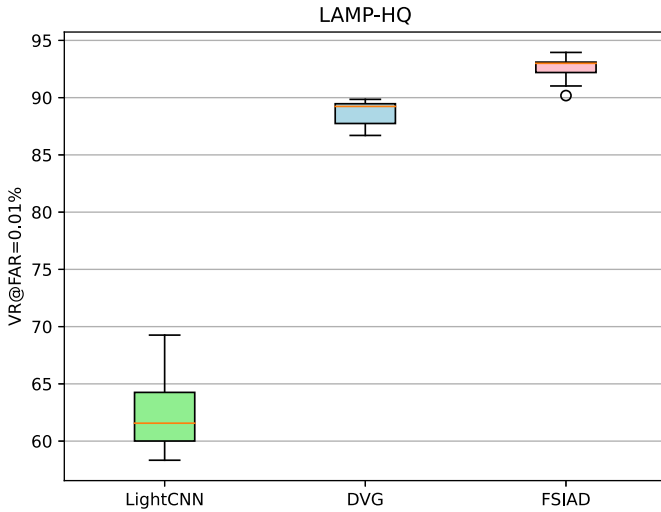
Fig. 8. The box plots of LightCNN, DVG and the proposed FSIAD on the LAMP-HQ database.

TABLE XIII

THE QUANTITATIVE RESULTS OF ABLATION STUDIES ON THE LAMP-HQ DATABASE

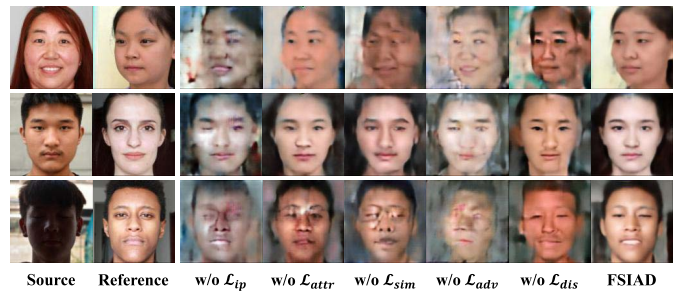| Method	Rank-1 (%)	VR@FAR =1%(%)	VR@FAR =0.1%(%)	VR@FAR =0.01%(%)
BL	94.5	92.5	77.3	60.3
LightCNN[69]	96.2	96.1	85.3	69.3
DVG[23]	98.3	98.8	96.0	88.2
w/o $\mathcal{L}_{in}$	96.2	96.1	85.3	69.3
w/o $\mathcal{L}_{ip}$	98.5	98.8	96.5	89.9
w/o $\mathcal{L}_{dis}$	98.5	98.9	96.6	89.6
w/o $\mathcal{L}_{sim}$	98.5	98.9	96.6	89.0
w/o $\mathcal{L}_{attr}$	98.6	98.9	97.0	90.9
w/o $\mathcal{L}_{adv}$	98.6	99.0	96.8	90.7
FSIAD	**98.8**	**99.1**	**97.9**	**93.0**

Fig. 9. The qualitative results of ablation studies on the LAMP-HQ database. The first and second columns are source images and reference images. The images in the rightmost column are synthesized by FSIAD, while the rest columns are synthesized by the ablation versions without loss functions $\mathcal{L}_{ip}$, $\mathcal{L}_{attr}$, $\mathcal{L}_{sim}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{dis}$, respectively.

paired heterogeneous faces hinder the improvement of HFR. Compared with ADFL and PACH, DVG gains progress and improves VR@FAR=1% from 95.5% to 99.0%. The results of DVG suggest that increasing heterogeneous faces by unconditional synthesis boosts the performance of HFR. FSIAD exceeds state-of-the-art methods, whose Rank-1 accuracy, VR@FAR=1%, VR@FAR=0.1%, and VR@FAR=0.01% are further improved by 0.4%, 0.2%, 0.9, and 4.0%, respectively. Moreover, Fig. 8 illustrates FSIAD also has stable performance and achieves the minimal standard deviation of results. Experimental results indicate it is fundamental to enrich the attribute diversity of synthetic heterogeneous faces to facilitate HFR. FSIAD produces sufficient heterogeneous faces with abundant facial attributes, which improve the generalization ability of the HFR network and substantially alleviate the degradation of performance caused by large variations in attributes.

E. Ablation Studies

To investigate the effectiveness of the proposed loss functions in FSIAD, we conduct ablation studies on the LAMP-HQ database. Since we follow the experimental settings of [23], [29], [52], [55], and [9] to adopt LightCNN [69] as the HFR network, we initialize its parameters with a model pre-trained on the MS-Celeb-1M dataset [79] as a baseline (BL). Then the baseline model is fine-tuned on the LAMP-HQ dataset and denoted as `LightCNN` in TABLE XIII.

For ablation studies, we construct five variants and each variant is trained without a specific loss function. ' Qualitative and quantitative evaluations are conducted. We firstly train variants on the LAMP-HQ dataset and then synthesize faces from disentangled representations of identities and attributes. Fig. 9 illustrates the synthetic faces generated by five variants. Obviously, the faces of FSIAD w/o $\mathcal{L}_{ip}$ are worst and lose most of the information of identity. Without the identity preserving loss $\mathcal{L}_{ip}$, it is difficult for FSIAD to synthesize faces whose identities are consistent with those of source faces. Hence the results of inferior quality indicate that identity preserving loss plays an elementary role in face synthesis. As for facial attributes, we find that the facial attributes are dissimilar with those of reference images when the attribute loss $\mathcal{L}_{attr}$ is removed. For example, we discover that complexions of synthetic faces in the first and second rows are different from those of their corresponding reference faces. Therefore, the attribute loss $\mathcal{L}_{attr}$ can supervise FSIAD to align facial attributes of synthetic faces to those of references. However, only relying on $\mathcal{L}_{attr}$ is not enough for attribute similarity constraint. We observe that the facial textures are blurry and ragged when FSIAD is trained without the structural similarity loss $\mathcal{L}_{sim}$. As we can see in Eq. (8), $\mathcal{L}_{sim}$ is composed of a L_1-norm loss and a multi-scale structural similarity loss $\mathcal{L}_{MS}$. On the one hand, the L_1-norm loss contributes to pixel consistency for face synthesis, and makes global

features including color and illumination of synthetic faces similar to those of reference faces. On the other hand, the $\mathcal{L}_{MS}$ loss further regularizes the local features such as eye, mouth, texture, and contour to close with those of references. Consequently, similarities of either attribute representations or structural information can not be dispensed with. The faces generated by FSIAD w/o $\mathcal{L}_{adv}$ are prototypes of synthetic results. But these faces are full of artifacts since the generator lacks the supervision of adversarial discriminator. It indicates that $\mathcal{L}_{adv}$ optimizes FSIAD to refine synthetic images and improve fidelity. Compared with the previous results, FSIAD w/o $\mathcal{L}_{dis}$ produces images of better quality. However, these synthetic faces exist blending boundaries and incoherent textures, especially the first and second images. This phenomenon is caused by the lack of feature disentanglement. Without $\mathcal{L}_{dis}$, the attribute representations learned from attribute encoders are mixed up with the identity features of reference faces. It is hard for a generator to produce faces with confused representations and thus result in fuzzy synthetic faces. Compared with variants, FSIAD takes the advantage of our proposed loss functions and successfully produces high-quality faces with the combinations of identity representations and facial attribute representations. The synthetic faces of FSIAD have smooth textures, coherent contours but no noise or artifact. Qualitative ablation studies reveal that all the proposed loss functions are essential for FSIAD to produce high-quality faces and implement heterogeneous face augmentation.

Apart from qualitative analyses, we also conduct quantitative ablation studies on the HFR performance of these variants. As demonstrated in TABLE XIII and Fig. 7, the baseline (BL) of ablation studies performs worst on the LAMP-HQ dataset because it is only pre-trained on the VIS faces. After BL is fine-tuned on LAMP-HQ, it achieves better performance and improves the VR@FAR=1% from 92.5% to 96.1%. To reveal the effectiveness of the synthetic faces, we construct a variant FSIAD w/o $\mathcal{L}_{in}$ by setting the γ to 0 in Eq. (15). It is equivalent to LightCNN, hence its results are identical to those of LightCNN. Owing to the enrichment of facial attributes, all variants of FSIAD outperform LightCNN and DVG. However, FSIAD w/o $\mathcal{L}_{ip}$ neglects the identity preservation and causes inconsistency of identity between source faces and synthetic faces. Hence, it directly damages face recognition and slightly improves VR@FAR=0.1% to 96.5%. Due to lacking feature disentanglement of identity and facial attributes, FSIAD w/o $\mathcal{L}_{dis}$ has a flaw in synthesis and causes incoherent textures in the generated faces. It hinders FSIAD from facilitating the performance of HFR and its results are close to those of FSIAD w/o $\mathcal{L}_{ip}$. Without $\mathcal{L}_{sim}$, FSIAD lacks the structural information of face from the reference image and synthesizes irregular faces. It results in the unsatisfactory performance of HFR. Compared with FSIAD w/o $\mathcal{L}_{sim}$, FSIAD w/o $\mathcal{L}_{attr}$ has better ability in face synthesis and improves VR@FAR=0.1% to 97.0%. FSIAD w/o $\mathcal{L}_{adv}$ can basically accomplish the goal of FSIAD that produces faces from identities of source and attributes of reference, and so achieves the best VR@FAR=1% among the aforementioned variants. But its synthetic faces are of low quality and filled with artifacts and thus lead to the medium performance of HFR.

FSIAD outperforms the variants and remarkably improves VR@FAR=0.01% from 90.9% to 93.0%. Only when all the proposed loss functions are activated can FSIAD exploit its optimal performance to achieve the best results. In summary, the quantitative and qualitative analyses indicate that all the proposed loss functions in FSIAD are indispensable, and they jointly supervise FSIAD to produce high-quality heterogeneous faces and facilitate the performance of HFR.

V. Conclusion

In this paper, we propose a novel method Face Synthesis with Identity-Attribute Disentanglement (FSIAD) to augment heterogeneous face images for facilitating the performance of HFR. FSIAD consists of Identity-Attribute Disentanglement (IAD) and Face Synthesis Module (FSM) components. The IAD component is designed to decouple faces into representations of identities and attributes, where attribute representations are regularized to be uncorrelated with identities. Then FSM synthesizes abundant heterogeneous faces from the combinations of disentangled representations of identities and attributes, which augment the insufficient training data of HFR and increase diversity in facial attributes. We train the HFR network with both original HFR data and synthetic heterogeneous faces for improving its performance on face recognition. Through disentanglement and synthesis, FSIAD naturally tackles three major challenges of HFR. Extensive experiments on five HFR databases reveal that our method is superior to previous methods and yields state-of-the-art performance.

References

[1] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 471–478.

[2] Y.-H. Tsai, H.-M. Hsu, C.-A. Hou, and Y.-C.-F. Wang, "Person-specific domain adaptation with applications to heterogeneous face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 338–342.

[3] B. S. Riggan *et al.*, "ICME grand challenge results on heterogeneous face recognition: Polarimetric thermal-to-visible matching," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2018, pp. 1–4.

[4] A. Kantarcı and H. K. Ekenel, "Thermal to visible face recognition using deep autoencoders," in *Proc. Int. Conf. Biometrics Special Interest Group*, 2019, pp. 1–5.

[5] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, "Heterogeneous face recognition by margin-based cross-modality metric learning," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1814–1826, Jun. 2018.

[6] C. Peng, N. Wang, J. Li, and X. Gao, "Re-ranking high-dimensional deep local representation for NIR-VIS face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4553–4565, Sep. 2019.

[7] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102805.

[8] T. de Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 7, pp. 1803–1816, Jul. 2018.

[9] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, Jul. 2019.

[10] A. Yu, H. Wu, H. Huang, Z. Lei, and R. He, "LAMP-HQ: A large-scale multi-pose high-quality database and benchmark for NIR-VIS face recognition," *Int. J. Comput. Vis.*, vol. 129, pp. 1–17, Feb. 2021.

[11] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, C. C. Loy, and X. Wang, "A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution," *Image Vis. Comput.*, vol. 56, pp. 28–48, Dec. 2016.

[12] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015.

[13] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Sep. 2011.

[14] M. Cho, T. Kim, I.-J. Kim, K. Lee, and S. Lee, "Relational deep feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 376–388, 2021.

[15] S. Yang, K. Fu, X. Yang, Y. Lin, J. Zhang, and C. Peng, "Learning domain-invariant discriminative features for heterogeneous face recognition," *IEEE Access*, vol. 8, pp. 209790–209801, 2020.

[16] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2012.

[17] W. Hu and H. Hu, "Dual adversarial disentanglement and deep representation decorrelation for NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 70–85, 2021.

[18] Z. Lei, Q. Bai, R. He, and S. Z. Li, "Face shape recovery from a single image using CCA mapping between tensor spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[19] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6628–6637.

[20] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 845–862, Jun. 2019.

[21] A. Kortylewski, A. Schneider, T. Gerig, B. Egger, A. Morel-Forster, and T. Vetter, "Training deep face recognition systems with synthetic data," 2018, *arXiv:1802.05891*.

[22] M. Luo, J. Cao, X. Ma, X. Zhang, and R. He, "FA-GAN: Face augmentation GAN for deformation-invariant face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2341–2355, 2021.

[23] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "Dual variational generation for low shot heterogeneous face recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2674–2683.

[24] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[25] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Proc. Adv. Biometrics*, 2009, pp. 209–218.

[26] D. Goswami, C. H. Chan, D. Windridge, and J. Kittler, "Evaluation of face recognition system in heterogeneous environments (visible vs NIR)," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2160–2167.

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

[28] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for NIR-VIS heterogeneous face recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[29] X. Wu, H. Huang, V. M. Patel, R. He, and Z. Sun, "Disentangled variational representation for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, no. 1, 2019, pp. 9005–9012.

[30] M. Cho, T.-Y. Chung, T. Kim, and S. Lee, "NIR-to-VIS face recognition via embedding relations and coordinates of the pairwise features," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–8.

[31] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "DVG-face: Dual variational generation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 18, 2021, doi: 10.1109/TPAMI.2021.3052549.

[32] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li, "Face matching between near infrared and visible light images," in *Proc. Int. Conf. Biometrics*, 2007, pp. 523–530.

[33] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.

[34] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit. (FG)*, May 2015, pp. 1–7.

[35] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 251–260.

[36] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 593–600.

[37] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for NIR-VIS face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2000–2006.

[38] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, no. 1, 2018, pp. 1–8.

[39] W. Hu and H. Hu, "Orthogonal modality disentanglement and representation alignment network for NIR-VIS face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 16, 2021, doi: 10.1109/TCSVT.2021.3105411.

[40] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 687–694.

[41] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[42] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2216–2223.

[43] D.-A. Huang and Y.-C.-F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2496–2503.

[44] F. Juefei-Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 141–150.

[45] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[46] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[47] H. B. Bae, T. Jeon, Y. Lee, S. Jang, and S. Lee, "Non-visual to visual translation for cross-domain face recognition," *IEEE Access*, vol. 8, pp. 50452–50464, 2020.

[48] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell.*, no. 1, 2018, pp. 7355–7363.

[49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[50] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025–1037, May 2020.

[51] J. Yu, J. Cao, Y. Li, X. Jia, and R. He, "Pose-preserving cross spectral face hallucination," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1018–1024.

[52] B. Duan, C. Fu, Y. Li, X. Song, and R. He, "Cross-spectral face hallucination via disentangling independent factors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7927–7935.

[53] M. Zhu, J. Li, N. Wang, and X. Gao, "Knowledge distillation for face photo-sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 27, 2020, doi: 10.1109/TNNLS.2020.3030536.

[54] W. Hu, W. Yan, and H. Hua, "Dual face alignment learning network for NIR-VIS face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 18, 2021, doi: 10.1109/TCSVT.2021.3081514.

[55] D. Liu, X. Gao, C. Peng, N. Wang, and J. Li, "Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 16, 2021, doi: 10.1109/TNNLS.2021.3071119.

[56] X. Di and V. M. Patel, "Multimodal face synthesis from visual attributes," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 427–439, Jul. 2021.

[57] X. Di and V. M. Patel, "Facial synthesis from visual attributes via sketch using multiscale generators," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 1, pp. 55–67, Jan. 2020.

[58] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.

[59] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2018, pp. 98–105.

[60] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5074–5083.

[61] 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. 4396–4405.

[62] 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.

[63] T. Karras *et al.*, "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–12.

[64] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or, "Face identity disentanglement via latent space mapping," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–14, Dec. 2020.

[65] H. Zhu, C. Fu, Q. Wu, W. Wu, C. Qian, and R. He, "AOT: Appearance optimal transport based identity swapping for forgery detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21699–21712.

[66] T. Daniel and A. Tamar, "Soft-IntroVAE: Analyzing and improving the introspective variational autoencoder," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4391–4400.

[67] J. Zhang *et al.*, "FReeNet: Multi-identity face reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5326–5335.

[68] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, "Information bottleneck disentanglement for identity swapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3404–3413.

[69] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[70] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[71] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[72] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

[73] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[74] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 348–353.

[75] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001, 2012, vol. 6.

[76] J. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikainen, "Learning mappings for face synthesis from near infrared to visual light images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 156–163.

[77] K. Panetta *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, Mar. 2020.

[78] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 451–463, Feb. 2016.

[79] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 87–102.

[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[81] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.

[82] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[83] T. Salimans *et al.*, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.
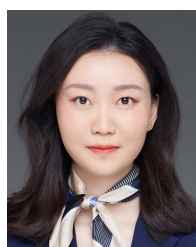
Ziming Yang received the B.E. degree in computer science and technology from Jinan University, Guangzhou, China, in 2020. He is currently pursuing the M.S. degree in computer application technology with the School of Cyber Security, University of Chinese Academy of Sciences, and the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. His research interests include biometrics, pattern recognition, and computer vision.

Jian Liang received the B.E. degree in electronic information and technology from Xi'an Jiaotong University and the Ph.D. degree in pattern recognition and intelligent systems from NLPR, CASIA in July 2013 and January 2019, respectively. He was a Research Fellow at the National University of Singapore from June 2019 to April 2021. Currently, he is an Associate Professor with NLPR. His research interests include transfer learning, pattern recognition, and computer vision.

Chaoyou Fu received the B.E. degree in automation from Anhui University, Hefei, China, in 2017. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition and computer vision.

Mandi Luo (Member, IEEE) received the B.E. degree in automation engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the B.Sc. and M.Sc. degrees in electronic engineering from Katholieke Universiteit te Leuven, Leuven, Belgium, in 2017 and 2018, respectively. She is currently pursuing the Ph.D. degree in computer application technology with the University of Chinese Academy of Sciences, Beijing, China. Her research interests include biometrics, pattern recognition, and computer vision.

Xiao-Yu Zhang (Senior Member, IEEE) received the B.S. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He has authored or coauthored more than 60 refereed publications in international journals and conferences. His research interests include artificial intelligence, data mining, and computer vision. He is a Senior Member of the ACM, CCF, and CSIG. His awards and honors include the Silver Prize of Microsoft Cup of the IEEE China Student Paper Contest in 2009, the Second Prize of Wu Wen-Jun AI Science and Technology Innovation Award in 2016, the CCCV Best Paper Nominate Award in 2017, the Third Prize of BAST Beijing Excellent S and T Paper Award in 2018, and the Second Prize of CSIG Science and Technology Award in 2019.