arXiv:1903.10203v1 [cs.CV] 25 Mar 2019

Dual Variational Generation for Low-Shot Heterogeneous Face Recognition

Chaoyou Fu[1,2,3*], **Xiang Wu**[1,2*], **Yibo Hu**[1,2], **Huaibo Huang**[1,2,3], **Ran He**[1,2,3†]

[1]Center for Research on Intelligent Perception and Computing, CASIA
[2]National Laboratory of Pattern Recognition, CASIA
[3]University of Chinese Academy of Sciences

{chaoyou.fu, rhe}@nlpr.ia.ac.cn, alfredxiangwu@gmail.com, {yibo.hu, huaibo.huang}@cripac.ia.ac.cn

Abstract

Heterogeneous Face Recognition (HFR) is a challenging issue because of the large domain discrepancy and a lack of heterogeneous data. This paper considers HFR as a dual generation problem, and proposes a new Dual Variational Generation (DVG) framework. It generates large-scale paired heterogeneous images with the same identity from noise, for the sake of reducing the domain gap of HFR, which provides a new insight into the two challenging issues in HFR. Specifically, we first introduce a dual variational autoencoder to represent a joint distribution of paired heterogeneous images. Then, we impose a distribution alignment loss in the latent space and a pairwise identity preserving loss in the image space. These ensure that DVG can generate diverse paired heterogeneous images of the same identity. Moreover, a pairwise distance loss between the generated paired heterogeneous images contributes to the optimization of the HFR network, aiming at reducing the domain discrepancy. Significant recognition improvements are observed on four HFR databases, paving a new way to address the low-shot HFR problems.

1 Introduction

With the development of deep learning, face recognition has made significant progress [Wu *et al.*, 2018a] in recent years. However, in many real-world applications, such as video surveillance, facial authentication on mobile devices and computer forensics, it is still a great challenge to match heterogeneous face images in different modalities, including sketch images [Zhang *et al.*, 2011], near infrared images [Li *et al.*, 2013] and polarimetric thermal images [Zhang *et al.*, 2019]. Therefore, heterogeneous face recognition (HFR) has attracted much attention in the face recognition community. Due to the domain gap, one challenge is that the face recognition model trained on VIS data often degrades significantly for HFR. Therefore, lots of cross domain feature matching methods [He *et al.*, 2017; Wu *et al.*, 2018b] are introduced to reduce the large domain gap between heterogeneous face images. However, since it is expensive and time-consuming to collect a large number of heterogeneous face images, there is no public large-scale heterogeneous face database. With the limited training data, CNNs trained for HFR often tend to be overfitting.

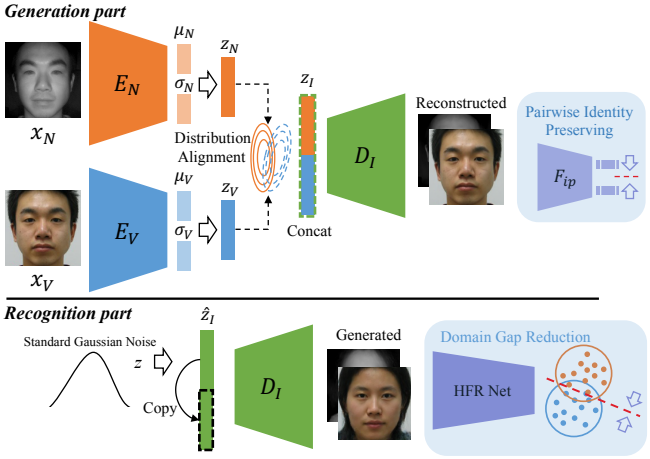

Figure 1: The framework of Dual Variational Generation (DVG). It contains generation and recognition parts. For generation, given a pair of heterogeneous images x_N and x_V from the same identity, a dual variational autoencoder represents a joint distribution $z_I = [z_N\ z_V] \in \mathbb{R}^{2d}$, where the paired representations $z_N \in \mathbb{R}^d$ and $z_V \in \mathbb{R}^d$ are obtained from two separate encoders E_N and E_V, respectively. The distribution alignment between z_N and z_V in the latent space and the pairwise identity preserving between the generated paired images in the image space are imposed into the dual variational autoencoder to guarantee the identity consistency of generated paired images. For recognition, given $\hat{z}_I = [z\ z] \in \mathbb{R}^{2d}$, where $z \in \mathbb{R}^d$ is sampled from the standard Gaussian noise, DVG can generate diverse pairs of new heterogeneous images, which contributes to the optimization of HFR network, aiming at reducing the domain discrepancy for low-shot HFR.

Recently, the great progress of high-quality face synthesis [Huang *et al.*, 2017; Hu *et al.*, 2018] has made "recognition via generation" possible. TP-GAN [Huang *et al.*, 2017] and CAPG-GAN [Hu *et al.*, 2018] introduce face synthesis to improve the quantitative performance of large pose face recognition. For HFR, [Song *et al.*, 2018] proposes a

*Equal Contribution
†Corresponding Author

two-path model to synthesize VIS images from NIR images. [Zhang *et al.*, 2019] utilizes a GAN based multi-stream feature fusion technique to generate VIS images from Polarimetric Thermal faces. However, all these methods are based on image-to-image translation framework, leading to two potential challenges: 1) Diversity: Given one image, a generator only synthesizes one new image of the target domain [Song *et al.*, 2018], which means such image-to-image translation methods can only generate limited number and diversity of images. In particular, this challenge will be very prominent in the low-shot heterogeneous face recognition, i.e., learning from few heterogeneous data. 2) Consistency: When generating large-scale samples, it is challenging to guarantee that the synthesized face images belong to the same identity of the input images. Although identity preserving loss [Hu *et al.*, 2018] can constrain the distances between features of the input and synthesized images, it does not constraint the intra-class and inter-class distances of the embedding space.

To tackle the above two challenges, we propose a novel Dual Variational Generation (DVG) framework that contains a generation part and a recognition part, as shown in Fig. 1. The generation part focuses on generating diverse pairs of new heterogeneous images, and the recognition part aims at utilizing these generated paired images to improve the performance of low-shot HFR. Specifically, for the generation part, we introduce a dual variational autoencoder to learn a joint distribution of paired heterogeneous images. Inspired by [Wu *et al.*, 2019], we introduce both a distribution alignment loss in the latent space and a pairwise distance loss in the image space. These constraints avoid the identity consistency problem of previous methods, since DVG only pays attention to the identity consistency of the paired heterogeneous images rather than the identity whom the paired heterogeneous images belong to. For recognition part, considering that variational autoencoder has the property of generating diverse new data [Kingma and Welling, 2014], we utilize it to generate new pairs of images. That is, by sampling and copying a noise from a standard Gaussian distribution, DVG can generate diverse pairs of new heterogeneous images with the same identity, as shown in Fig. 2. Finally, these generated paired images are used to optimize the HFR network by a pairwise distance loss, aiming at reducing the domain discrepancy.

In summary, the main contributions are as follows:

- We provide a new insight into the problems of HFR. That is, we consider HFR as a dual generation problem, and propose a novel dual variational generation framework. This framework generates diverse paired heterogeneous images to reduce the domain gap of HFR.
- A distribution alignment loss and a pairwise identity preserving loss are proposed to guarantee the identity consistency of the generated paired heterogeneous images. These avoid the identity consistency problem of the previous methods.
- A pairwise distance loss between the paired images generated from noise is introduced in the HFR network. By generating large-scale diverse paired images, we can reduce the domain gap of HFR.
- Experiments on four HFR databases demonstrate that our method can generate photo-realistic paired heterogeneous images and significantly improve the performance of recognition, paving a new way to solve low-shot HFR problems.

2 Background and Related Work

2.1 Heterogeneous Face Recognition

Lots of researchers pay their attention to Heterogeneous Face Recognition (HFR). For the feature-level learning, [Klare *et al.*, 2011] employs HOG features with sparse representation for HFR. [Goswami *et al.*, 2011] utilizes LBP histogram with Linear Discriminant Analysis to obtain domain-invariant features. [He *et al.*, 2017] proposes Invariant Deep Representation (IDR) to disentangle representations into two orthogonal subspaces for NIR-VIS HFR. Further, [He *et al.*, 2018] extends IDR by introducing Wasserstein distance to obtain domain invariant features for HFR. Disentangled Variational Representation (DVR) [Wu *et al.*, 2019] is proposed to model the compact and discriminative disentangled latent variable spaces for heterogeneous images. For the image-level learning, the common idea is to transform heterogeneous face images from one modality into another one via image synthesis. [Juefei-Xu *et al.*, 2015] utilizes joint dictionary learning to reconstruct face images for boosting the performance of face matching. [Lezama *et al.*, 2017] proposes a cross-spectral hallucination and low-rank embedding to synthesize a heterogeneous image in a patch way.

2.2 Generative Models

Variational autoencoders (VAEs) [Kingma and Welling, 2014] and Generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] are the most prominent generative models. VAEs consist of an encoder network $q_\phi(z|x)$ and a decoder network $p_\theta(x|z)$. $q_\phi(z|x)$ maps input images x to the latent variables z that match to a prior $p(z)$, and $p_\theta(x|z)$ samples images x from the the latent variables z. The evidence lower bound objective (ELBO) of VAEs:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) - D_{\mathrm{KL}}(q_\phi(z|x)||p(z)). \tag{1}$$

The two components in ELBO are a reconstruction error and a Kullback-Leibler divergence, respectively.

Differently, GANs adopt a generator G and a discriminator D to play a min-max game. G generates images from a prior $p(z)$ to confuse D, and D is trained to distinguish between generated data and real data. This adversarial rule takes the form:

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}(x)} \left[\log D(x)\right] + \mathbb{E}_{z\sim p_z(z)} \left[\log(1 - D(G(z)))\right]. \tag{2}$$

They have achieved remarkable success in various applications, such as image generation [Huang *et al.*, 2018], image translation [Song *et al.*, 2018], face editing [Huang *et al.*, 2017]. According to [Huang *et al.*, 2018], VAEs have nice manifold representations, while GANs are better at generating sharper images.

Another work to address the similar problem of our method is CoGAN [Liu and Tuzel, 2016], which uses a weight-sharing manner to generate paired images in two different

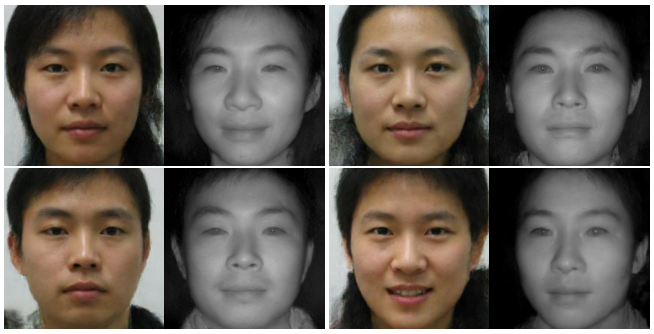

Figure 2: The dual generation results (256×256 resolution). For each pair, the left is VIS and the right is the paired NIR image.

modalities. However, CoGAN does not explicitly constrain the distributions of two modalities in the latent space, and the weight-sharing manner of CoGAN can not maintain the identity consistency of paired images. These two factors make it challenging for CoGAN to generate paired images with the same identity, as showin in Fig. 3. Differently, we explicitly constrain the distributions of two modalities in the latent space and the identity consistency in the image space.

2.3 Low-Shot Learning

Low-shot learning is a fundamental problem in machine learning. Obviously, lacking labeled images is a serious problem in heterogeneous face recognition [Wu *et al.*, 2019]. For example, the CUHK Face Sketch FERET database [Zhang *et al.*, 2011] contains images of 1194 subjects, but each subject only has 2 images. Learning from such few examples is a huge challenge. There are three main approaches have been proposed to tackle this problem [Wang *et al.*, 2018], including generative models, representation learning and meta-learning. Considering generative models for low-shot learning, it aims at synthesizing additional images of different categories. Through an additional corpus with attribute annotations, [Dixit *et al.*, 2017] performs feature augmentation by varying attributes. [Schwartz *et al.*, 2018] transfers intra-class deformations of one class to another class to synthesize samples. Different from the previous generative models, we generate new paired images with the same identity, instead of adding samples for existing classes.

3 Proposed Method

Our method mainly consists of two parts, i.e., a generation part for generating new paired heterogeneous images and a recognition part for learning domain-invariant features. In this section, we will introduce these two parts in details. Note that we specifically discuss the NIR-VIS images for better expression. Other heterogeneous images are also applicable.

3.1 Dual Variational Generation

As shown in Fig. 1, the generation part consists of a feature extractor F_{ip}, and a dual variational autoencoder: two encoder networks and a decoder network, all of which play the same roles as VAEs [Kingma and Welling, 2014]. Specifically, F_{ip} extracts the semantic information of the generated images to preserve the identity information. The encoder network $q_{\phi_N}(z_N|x_N)$ maps NIR images x_N to a latent space z_N by a reparameterization trick: $z_N = u_N + \sigma_N \odot \epsilon$, where u_N and σ_N denote mean and standard deviation, respectively. In addition, ϵ is sampled from a multi-variate standard Gaussian and $\odot$ denotes the Hadamard product. The encoder network $q_{\phi_V}(z_V|x_V)$ has the same manner as $q_{\phi_N}(z_N|x_N)$, which is for VIS images x_V. After obtaining the two independent distributions, we concatenate z_N and z_V to get the joint distribution z_I.

Distribution Learning We utilize VAEs to learn the joint distribution of the paired NIR-VIS images. Given a pair of NIR-VIS images $\{x_N, x_V\}$, we constrain the posterior distribution $q_{\phi_N}(z_N|x_N)$ and $q_{\phi_V}(z_V|x_V)$ by the Kullback-Leibler divergence:

$$\begin{aligned}\mathcal{L}_{\text{kl}} = & D_{\text{KL}}(q_{\phi_N}(z_N|x_N)||p(z_N)) \\ & + D_{\text{KL}}(q_{\phi_V}(z_V|x_V)||p(z_V)),\end{aligned} \tag{3}$$

where the prior distributions $p(z_N)$ and $p(z_V)$ are both the multivariate standard Gaussian distributions. Like the original VAEs, we require the decoder network $p_\theta(x_N, x_V|z_I)$ to be able to reconstruct the input images x_N and x_V from the learned distribution:

$$\mathcal{L}_{\text{rec}} = -\mathbb{E}_{q_{\phi_N}(z_N|x_N) \cup q_{\phi_V}(z_V|x_V)} \log p_\theta(x_N, x_V|z_I), \tag{4}$$

Distribution Alignment We expect a pair of NIR-VIS images $\{x_N, x_V\}$ to be projected into a common latent space by the encoders $E_N(x_N)$ and $E_V(x_V)$, i.e., the NIR distribution $p(z_N^{(i)})$ is the same as the VIS distribution $p(z_V^{(i)})$, where i denotes the identity information. That means we maintain the identity consistency of the generated paired images in the latent space. Explicitly, we align the NIR and VIS distributions by minimizing the Wasserstein distance between the two distributions. Given two Gaussian distributions $p(z_N^{(i)}) = N(u_N^{(i)}, {\sigma_N^{(i)}}^2)$ and $p(z_V^{(i)}) = N(u_V^{(i)}, {\sigma_V^{(i)}}^2)$, the 2-Wasserstein distance between $p(z_N^{(i)})$ and $p(z_V^{(i)})$ is defined as:

$$\begin{aligned}W(p(z_N^{(i)}), p(z_V^{(i)})) = & ||u_N^{(i)} - u_V^{(i)}||_2^2 \\ & + \text{Tr}(\sigma_N^{(i)} + \sigma_V^{(i)} - 2({\sigma_V^{(i)}}^{\frac{1}{2}} \sigma_N^{(i)} {\sigma_V^{(i)}}^{\frac{1}{2}})^{\frac{1}{2}}).\end{aligned} \tag{5}$$

According to [He *et al.*, 2017], Eq. (5) can be simplified as

$$W(p(z_N^{(i)}), p(z_V^{(i)})) = \frac{1}{2}\left[||u_N^{(i)} - u_V^{(i)}||_2^2 + ||\sigma_N^{(i)} - \sigma_V^{(i)}||_2^2\right] \tag{6}$$

We minimize the above Wasserstain distance with total M identities:

$$\mathcal{L}_{\text{dist}} = \sum_i^M W(p(z_N^{(i)}), p(z_V^{(i)})). \tag{7}$$

Pairwise Identity Preserving In previous image-to-image translation works [Huang *et al.*, 2017; Hu *et al.*, 2018], identity preserving is usually introduced to maintain identity information. The traditional approach uses a pre-trained feature

extractor to enforce the features of the generated images to be close to the features of the target images. However, it is challenge to guarantee the synthesized images to belong to the same identity as the target images, because this manner does not constrain the intra-class and inter-class distances of the embedding space. In our method, since we generate a pair of heterogeneous images, we only need to consider the identity consistency of the paired images.

Specifically, we adopt Light CNN [Wu *et al.*, 2018a] as the feature extractor F_{ip} to constrain the features between the reconstructed paired images:

$$\mathcal{L}_{\text{ip-pair}} = ||F_{ip}(\hat{x}_N) - F_{ip}(\hat{x}_V)||_2^2, \tag{8}$$

we also use F_{ip} to make the features of the reconstructed images and the original input images close enough:

$$\mathcal{L}_{\text{ip-rec}} = ||F_{ip}(\hat{x}_N) - F_{ip}(x_N)||_2^2 + ||F_{ip}(\hat{x}_V) - F_{ip}(x_V)||_2^2, \tag{9}$$

where $\hat{x}_N$ and $\hat{x}_V$ denote the reconstructions of the input paired images x_N and x_V, respectively. All of these constraints can be formulated as:

$$\mathcal{L}_{\text{ip}} = \mathcal{L}_{\text{ip-rec}} + \mathcal{L}_{\text{ip-pair}}, \tag{10}$$

Diversity Constraint In order to further increase the diversity of the generated images, we also design a diversity loss. For each generated paired images from noise, we randomly assign them to an identity and optimize with cross entropy loss. Concretely, a pre-trained Light CNN has about 100K identities and we randomly choose one y_{random} as the identity of the generated image pair. Formally, the diversity loss is formulated as

$$\mathcal{L}_{\text{div}} = \sum_{i \in \{N,V\}} \text{softmax}(F_{ip}(\tilde{x}_i), y_{\text{random}}), \tag{11}$$

where $\tilde{x}_i (i \in \{N, V\})$ denotes the generated images from noise.

Overall Loss Moreover, in order to increase the sharpness of our generated images, we also adopt an adversarial loss as [Shu *et al.*, 2018]. Hence, the overall loss to optimize the generation network (dual variational autoencoder) can be formulated as

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{kl}} + \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{dist}} + \lambda_2 \mathcal{L}_{\text{ip}} + \lambda_3 \mathcal{L}_{\text{div}} + \lambda_4 \mathcal{L}_{\text{adv}}, \tag{12}$$

where λ_1, λ_2, λ_3 and λ_4 are the trade-off parameters.

3.2 Heterogeneous Face Recognition

For the recognition part, our training data contains the original limited labeled data $x_i (i \in \{N, V\})$ and the large-scale generated unlabeled paired NIR-VIS data $\tilde{x}_i (i \in \{N, V\})$. Here, we define a heterogeneous face recognition network F to extract features $f_i = F(x_i; \Theta)$, where $i \in \{N, V\}$ and Θ is the parameters of F. For the original labeled NIR and VIS images, we utilize a softmax loss:

$$\mathcal{L}_{\text{cls}} = \sum_{i \in \{N,V\}} \text{softmax}(F(x_i; \Theta), y), \tag{13}$$

where y is the label of identity.

For the generated paired heterogeneous images, since they are generated from noise, there are no specific class for the paired images. But as mentioned in section 3.1, DVG can ensure that the generated paired images belong to the same identity. Therefore, a pairwise distance loss between the paired heterogeneous samples is formulated as follows:

$$\mathcal{L}_{\text{pair}} = ||F(\tilde{x}_N; \Theta) - F(\tilde{x}_V; \Theta)||_2^2, \tag{14}$$

In this way, we can efficiently minimize the domain discrepancy by generating large-scale unlabeled paired heterogeneous images. As stated above, the final loss to optimize for the heterogeneous face recognition network can be written as

$$\mathcal{L}_{\text{hfr}} = \mathcal{L}_{\text{cls}} + \alpha_1 \mathcal{L}_{\text{pair}}, \tag{15}$$

where α_1 is the trade-off parameter.

4 Experiments

4.1 Databases and Protocols

Three NIR-VIS heterogeneous face recognition databases and one viewed sketch-photo database are used to evaluate our proposed method. For the NIR-VIS face recognition, following [Wu *et al.*, 2019], we report rank-1 accuracy and verification rate (VR)@ false accept rate (FAR) for the CASIA NIR-VIS 2.0 Face database [Li *et al.*, 2013], the Oulu-CASIA NIR-VIS database [Chen *et al.*, 2009] and the BUAA-VisNir Face database [Huang *et al.*, 2012]. Note that, for the Oulu-CASIA NIR-VIS database, there are only 20 subjects are selected as the training set. In addition, considering the viewed sketch-photo face recognition, the IIIT-D Viewed Sketch database [Bhatt *et al.*, 2012] are employed. Due to the few number of images in IIIT-D Viewed Sketch database, following the protocols of [Wu *et al.*, 2018b], we use CUHK Face Sketch FERET (CUFSF) [Zhang *et al.*, 2011] as the training set and report the rank-1 accuracy and VR@FAR=1% for comparisons.

4.2 Experimental Details

For the image generation part, the architecture of the encoder and decoder networks is the same as [Huang *et al.*, 2018] and the architecture of our discriminator is the same as [Shu *et al.*, 2018]. These networks are trained using Adam optimizer with a fixed rate of 0.0002. Other parameters λ_1, λ_2, λ_3 and λ_4 in Eq. (12) are set to 0.1, 0.2, 0.1 and 0.2, respectively. For the recognition part, we utilize both LightCNN-9 and LightCNN-29 [Wu *et al.*, 2018a] as the backbones for HFR. The models are pre-trained on the MS-Celeb-1M database [Guo *et al.*, 2016] and fine-tuned on the HFR training sets. All the face images are aligned to 144×144 and randomly cropped to 128×128 as the input for training. Stochastic gradient descent (SGD) is used as the optimizer, where the momentum is set to 0.9 and weight decay is set to $5e$-4. The learning rate is set to $1e$-3 initially and reduces to $5e$-4 gradually. The batch size is set to 64 and the dropout ratio is 0.5. The trade-off parameters α_1 in Eq. (15) is set to 0.01 during training.

Method	MD	FID	Rank-1
CoGAN	0.61	10.6	95.2
Baseline	0.50	7.8	96.8
DVG	**0.24**	**7.0**	**99.2**

(a)

Method	Rank-1
w/o $\mathcal{L}_{\text{dist}}$	95.5
w/o $\mathcal{L}_{\text{ip-pair}}$	96.3
w/o $\mathcal{L}_{\text{div}}$	98.6
DVG	**99.2**

(b)

Table 1: Experimental analyses on the CASIA NIR-VIS 2.0 Face database. The backbone is LightCNN-9. (a) The quantitative comparisons of different methods. MD means the mean feature distance between the generated paired NIR and VIS images. FID (lower is better) is measured based on the features of LightCNN-9, instead of the traditional Inception model. (b) The ablation study of DVG.

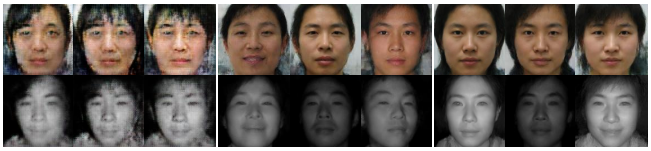

Figure 3: Visual comparisons of dual image generation results. The generated paired images of DVG are more similar than those of baseline and CoGAN.

4.3 Experimental Analyses

In this section, we analyze three metrics, including identity consistency, distribution consistency and visual quality, to demonstrate the effectiveness of DVG, compared with our baseline method and CoGAN [Liu and Tuzel, 2016]. The baseline method just has one encoder network, and the input is the concatenated NIR-VIS images. In other words, it directly learns the joint distribution, instead of aligning the distributions of two modalities in the latent space.

Identity Consistency. In order to analyze the identity consistency, we measure the feature distance between the generated paired images on the CASIA NIR-VIS 2.0 database. Specifically, we first use a pre-trained Light CNN-9 [Wu *et al.*, 2018a] to extract features and then measure the mean distance (MD) of the paired images. The results are reported in Table 1a. MD is computed from 50K generated image pairs and the MD value of the original database is 0.26. We can clearly see that the MD value of DVG is even smaller than the original database, which means that our method can effectively guarantee the identity consistency of the generated paired images. The recognition performance of different methods is also reported in Table 1a. We can see that DVG correspondingly achieves the best results.

Distribution Consistency. On the CASIA NIR-VIS 2.0 database, we take Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] to measure the Fréchet distance of two distributions in the feature space, reflecting the distribution consistency. We first measure the FID between the generated VIS images and the real VIS images, and the FID between the generated NIR images and the real NIR images, respectively. Then we calculate the mean FID as the final FID, which is reported in Table 1a. Considering that the face recognition network can better extract features of face images, we use a LightCNN-9 to extract features for calculating FID instead of acquiescent Inception model. Similarly, FID is computed from 50K generated image pairs. As shown in Table 1a, DVG achieves best results, demonstrating that DVG has really learned the distributions of two modalities.

Figure 4: The dual generation results on the CASIA NIR-VIS 2.0 Face database (first two rows) and the IIIT-D Viewed Sketch database (last two rows).

Visual Quality. In Fig. 3, we compare the dual generation results (128 $\times$ 128 resolution) of different methods on the CASIA NIR-VIS 2.0 Face database. Our visual results are obviously better than baseline and CoGAN. Moreover, we can observe that the generated paired images of baseline and CoGAN are not similar, which leads to worse rank-1 accuracy during optimizing HFR network (see Table 1a). More dual generation results of DVG are shown in Fig. 4.

Ablation Study. Table 1b presents the comparison results of our DVG and its three variants on the CASIA NIR-VIS 2.0 database. We observe that the recognition performance will decrease if one component is not adopted. Particularly, the accuracy drops significantly when the distribution alignment loss $\mathcal{L}_{\text{dist}}$ or the pairwise identity preserving loss $\mathcal{L}_{\text{ip-pair}}$ are not used. These results suggest that every component is crucial in our model.

Moreover, we analyze how the number of generated samples influence the HFR network on the Oulu-CASIA NIR-VIS database that only contains 20 identities about 1,000 images for training. We generate 1K, 5K, 10K and 50K pairs of heterogeneous images via DVG, and we obtain 80.2%, 84.8%, 89.3% and 89.2% on VR@FAR=0.1% by LightCNN-9, respectively. The results have been significantly improved with the increasing number of the generated pairs, suggesting that DVG can boost the performance of the low-shot heterogeneous face recognition.

4.4 Experimental Results

The recognition performance of our proposed DVG are demonstrated in this section on four heterogeneous face recognition databases. The performance of state-of-the-art methods, such as IDNet [Reale *et al.*, 2016], HFR-CNN [Saxena and Verbeek, 2016], Hallucination [Lezama *et al.*, 2017], DLFace [Peng *et al.*, 2019] TRIVET [Liu *et al.*, 2016], IDR [He *et al.*, 2017], CDL [Wu *et al.*, 2018b], W-CNN [He *et al.*, 2018], DVR [Wu *et al.*, 2019] and RCN [Deng *et al.*, 2019] are compared in Table 2.

For the most challenging CASIA NIR-VIS 2.0 database, it is obvious that DVG outperforms other state-of-the-art

Method	CASIA NIR-VIS 2.0		Oulu-CASIA NIR-VIS			BUAA-VisNir			IIIT-D Viewed Sketch	
	Rank-1	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%	Rank-1	FAR=1%
IDNet [Reale *et al.*, 2016]	87.1 ± 0.9	74.5	-	-	-	-	-	-	-	-
HFR-CNN [Saxena and Verbeek, 2016]	85.9 ± 0.9	78.0	-	-	-	-	-	-	-	-
Hallucination [Lezama *et al.*, 2017]	89.6 ± 0.9	-	-	-	-	-	-	-	-	-
DLFace [Peng *et al.*, 2019]	98.68	-	-	-	-	-	-	-	-	-
TRIVET [Liu *et al.*, 2016]	95.7 ± 0.5	91.0 ± 1.3	92.2	67.9	33.6	93.9	93.0	80.9	-	-
IDR [He *et al.*, 2017]	97.3 ± 0.4	95.7 ± 0.7	94.3	73.4	46.2	94.3	93.4	84.7	-	-
CDL [Wu *et al.*, 2018b]	98.6 ± 0.2	98.3 ± 0.1	94.3	81.6	53.9	96.9	95.9	90.1	85.35	82.52
W-CNN [He *et al.*, 2018]	98.7 ± 0.3	98.4 ± 0.4	98.0	81.5	54.6	97.4	96.0	91.9	-	-
DVR [Wu *et al.*, 2019]	99.7 ± 0.1	99.6 ± 0.3	100.0	97.2	84.9	**99.2**	**98.5**	96.9	-	-
RCN [Deng *et al.*, 2019]	99.3 ± 0.2	98.7 ± 0.2	-	-	-	-	-	-	90.34	-
LightCNN-9	97.1 ± 0.7	93.7 ± 0.8	93.8	80.4	43.8	94.8	94.3	83.5	84.07	75.30
LightCNN-9 + DVG	99.2 ± 0.2	98.8 ± 0.2	100.0	97.5	89.3	98.0	97.2	93.0	86.63	92.24
LightCNN-29	98.1 ± 0.4	97.4 ± 0.5	99.0	93.1	68.3	96.8	97.0	89.4	83.24	81.04
LightCNN-29 + DVG	$\mathbf{99.8 \pm 0.1}$	$\mathbf{99.8 \pm 0.1}$	**100.0**	**98.4**	**92.9**	**99.2**	**98.5**	**97.3**	**96.98**	**97.84**

Table 2: Comparisons with other state-of-the-art deep HFR methods on the CASIA NIR-VIS 2.0 database, the Oulu-CASIA NIR-VIS database, the BUAA-VisNir database and the IIIT-D Viewed Sketch database.

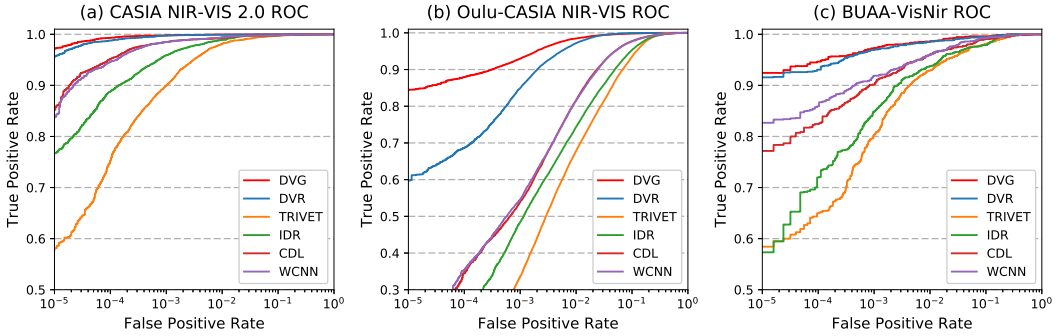

Figure 5: The ROC curves on the CASIA NIR-VIS 2.0, the Oulu-CASIA NIR-VIS and the BUAA-VisNir databases, respectively

methods. For fair comparisons with the previous works, including TRIVET, IDR, CDL, W-CNN, DVR and RCN, we employ LightCNN-9 as the backbone to perform DVG, which obtains $\mathbf{99.2}\%$ on Rank-1 accuracy and $\mathbf{98.8}\%$ on VR@FAR=0.1%. Further, when backbone changed to more powerful LightCNN-29, DVG also gains $\mathbf{0.1}\%$ on Rank-1 accuracy and $\mathbf{0.2}\%$ on VR@FAR=0.1%. Moreover, for BUAA-VisNir Face database, DVG also obtains $\mathbf{99.2}\%$ on Rank-1 accuracy and $\mathbf{97.3}\%$ on VR@FAR=0.1%, which outperforms other state-of-the-art methods.

To further analyze the effectiveness of the proposed DVG for low-shot heterogeneous face recognition, we evaluate DVG on Oulu-CASIA NIR-VIS and IIIT-D Viewed Sketch Face databases. As mentioned in section 4.1, there are fewer identities or images in these two databases. Table 2 presents the performance of DVG on these two challenging low-shot HFR databases. For Oulu-CASIA NIR-VIS database, we observe that DVG with LightCNN-29 significantly boosts the performance from 84.9% [Wu *et al.*, 2019] to $\mathbf{92.9}\%$ on VR@FAR=0.1%. Besides, for IIIT-D Viewed Sketch Face database, DVG also obtains $\mathbf{96.98}\%$ on Rank-1 accuracy and $\mathbf{97.84}\%$ on VR@FAR=1%, which outperforms state-of-the-art methods including CDL and RCN by a large margin.

Fig. 5 presents the ROC curves, including TRIVET, IDR, CDL, W-CNN, DVR and the proposed DVG. To better demonstrate the results, we only perform ROC curves of DVR and DVG trained on LightCNN-29. It is obvious that DVG outperforms other state-of-the-art methods, especially on the Oulu-CASIA NIR-VIS database.

5 Conclusion

This paper has developed a novel dual variational generation (DVG) framework for low-shot heterogeneous face recognition. It contains a generation part for generating diverse new paired heterogeneous images and a recognition part for using these generated paired images to reduce the domain gap of HFR, which provides a new insight into the problems of HFR. A dual variational autoencoder is first proposed to learn a joint distribution of paired heterogeneous images. Then, both distribution alignment loss in the latent space and pairwise distance loss in the image space are utilized to ensure the identity consistency of the generated image pairs. After that, DVG can generate diverse pairs of new heterogeneous images with the same identity from noise. Finally, these generated images are used to boost HFR network. Extensive qualitative and quantitative experimental results on four databases have shown the superiority of our DVG.

References

[Bhatt *et al.*, 2012] Himanshu S Bhatt, Samarth Bharadwaj, Richa Singh, and Mayank Vatsa. Memetic approach for matching sketches with digital face images. Technical report, 2012.

[Chen *et al.*, 2009] 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 *CVPR*, 2009.

[Deng *et al.*, 2019] Zhongying Deng, Xiaojiang Peng, and Yu Qiao. Residual compensation networks for heterogeneous face recognition. In *AAAI*, 2019.

[Dixit *et al.*, 2017] Mandar Dixit, Roland Kwitt, Marc Niethammer, and Nuno Vasconcelos. Aga: Attribute-guided augmentation. In *CVPR*, 2017.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, 2014.

[Goswami *et al.*, 2011] Debaditya Goswami, Chi-Ho Chan, David Windridge, and Josef Kittler. Evaluation of face recognition system in heterogeneous environments (visible vs nir). In *ICCV Workshops*, 2011.

[Guo *et al.*, 2016] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[He *et al.*, 2017] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for NIR-VIS face recognition. In *AAAI*, 2017.

[He *et al.*, 2018] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein CNN: learning invariant features for NIR-VIS face recognition. *TPAMI*, 2018.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. In *NIPS*, 2017.

[Hu *et al.*, 2018] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *CVPR*, 2018.

[Huang *et al.*, 2012] D. Huang, J. Sun, and Y. Wang. The BUAA-VisNir face database instructions. Technical report, 2012.

[Huang *et al.*, 2017] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017.

[Huang *et al.*, 2018] Huaibo Huang, Zhihang Li, Ran He, Zhenan Sun, and Tieniu Tan. Introvae: Introspective variational autoencoders for photographic image synthesis. In *NIPS*, 2018.

[Juefei-Xu *et al.*, 2015] Felix Juefei-Xu, Dipan K. Pal, and Marios Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *CVPR Workshops*, 2015.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Klare *et al.*, 2011] Brendan Klare, Zhifeng Li, and Anil K. Jain. Matching forensic sketches to mug shot photos. *TPAMI*, 2011.

[Lezama *et al.*, 2017] José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *CVPR*, 2017.

[Li *et al.*, 2013] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *CVPR Workshops*, 2013.

[Liu and Tuzel, 2016] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.

[Liu *et al.*, 2016] X. Liu, L. Song, X. Wu, and T. Tan. Transferring deep representation for nir-vis heterogeneous face recognition. In *ICB*, 2016.

[Peng *et al.*, 2019] Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. Dlface: Deep local descriptor for cross-modality face recognition. *PR*, 2019.

[Reale *et al.*, 2016] Christopher Reale, Nasser M. Nasrabadi, Heesung Kwon, and Rama Chellappa. Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition. In *CVPR Workshops*, 2016.

[Saxena and Verbeek, 2016] Shreyas Saxena and Jakob Verbeek. Heterogeneous face recognition with cnns. In *ECCV Workshops*, 2016.

[Schwartz *et al.*, 2018] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Rogerio Feris, Abhishek Kumar, Raja Giryes, and Alex M Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *NIPS*, 2018.

[Shu *et al.*, 2018] Zhixin Shu, Mihir Sahasrabudhe, Rıza Alp Güler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018.

[Song *et al.*, 2018] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. In *AAAI*, 2018.

[Wang *et al.*, 2018] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018.

[Wu *et al.*, 2018a] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *TIFS*, 2018.

[Wu *et al.*, 2018b] Xiang Wu, Lingxiao Song, Ran He, and Tieniu Tan. Coupled deep learning for heterogeneous face recognition. In *AAAI*, 2018.

[Wu *et al.*, 2019] Xiang Wu, Huaibo Huang, Vishal M. Patel, Ran He, and Zhenan Sun. Disentangled variational representation for heterogeneous face recognition. In *AAAI*, 2019.

[Zhang *et al.*, 2011] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, 2011.

[Zhang *et al.*, 2019] He Zhang, Benjamin S. Riggan, Shuowen Hu, Nathaniel J. Short, and Vishal M. Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *IJCV*, 2019.