Learning Domain-Invariant Discriminative Features for Heterogeneous Face Recognition

SHANMIN YANG, KEREN FU, XIAO YANG, YE LIN, JIANWEI ZHANG, AND CHENG PENG

Abstract

Heterogeneous face recognition (HFR), referring to matching face images across different domains, is a challenging problem due to the vast cross-domain discrepancy and insufficient pairwise cross-domain training data. This article proposes a quadruplet framework for learning domain-invariant discriminative features (DIDF) for HFR, which integrates domain-level and class-level alignment in one unified network. The domain-level alignment reduces the cross-domain distribution discrepancy. The class-level alignment based on a special quadruplet loss is developed to further diminish the intra-class variations and enlarge the inter-class separability among instances, thus handling the misalignment and adversarial equilibrium problems confronted by the domain-level alignment. With a bidirectional cross-domain data selection strategy, the quadruplet loss-based method prominently enriches the training set and further eliminates the cross-modality shift. Benefiting from the joint supervision and mutual reinforcement of these two components, the domain invariance and class discrimination of identity features are guaranteed. Extensive experiments on the challenging CASIA NIR-VIS 2.0 database, the Oulu-CASIA NIR&VIS database, the BUAA-VisNir database, and the IIIT-D viewed sketch database demonstrate the effectiveness and preferable generalization capability of the proposed method.

Index Terms

Adversarial domain adaptation, domain-invariant feature, heterogeneous face recognition, quadruplet loss.

I. INTRODUCTION

Deep convolution neural network (CNN) based face recognition (FR) has made impressive progress in recent years [1]. However, the performance of most face recognition systems degrades severely in specific real-world applications, e.g., identifying faces captured with near-infrared (NIR) sensory devices in surveillance under night-time and low-light conditions [2] and (or) recognizing sketch drawings based on the description of witnesses [3]. This problem mainly results from that most pre-enrolled face databases are collected in visual (VIS) conditions, which have substantial appearance differences from NIR and sketch counterparts. With the effectiveness and convenience, NIR and sketch face recognition plays a more and more important role in security control and law enforcement applications. Consequently, the demand for robust heterogeneous face recognition (HFR) proliferates.

As one kind of fine-grained classification task [1], face recognition is a challenging problem due to the high inter-class similarity and the rich intra-class diversity. Face images of different identities may look like each other owing to blood relationships, environments, and other factors, whereas different face images of the same identity may exhibit notable appearance variations because of changes in illumination, pose, expression, and so forth. Compared with traditional face recognition, HFR further faces more challenges.

One of the challenges is lacking in sufficient pairwise cross-modality training data. The success of deep CNN-based methods largely relies on large scale training datasets. However, there is few publicly available large-scale pairwise HFR dataset. It is prohibitively expensive and time-consuming...
FIGURE 1. An overview of the proposed quadruplet heterogeneous face recognition architecture. GRL represents a gradient reversal layer, which works as an identity transformation during forward-propagation. While in back-propagation, it transmits the gradient from the subsequent layer to the previous layer after multiplying the gradient by a factor of $-\lambda$.

to gather large scale pairwise cross-modality face images for training. Hence, how to train deep models on small-scale paired heterogeneous face dataset remains a critical problem.

The other problem arises from the discrepancy of data distributions between different domains. This difference brings in significant appearance variations even for the same identity, e.g., the overall pixel intensity distribution and local component details of face images collected from different domains vary considerably, causing great difficulty for heterogeneous face recognition.

Numerous researches have been dedicated to reducing the domain gap for HFR. One most straightforward and accessible strategy is learning domain invariant face representations, like [2], [4]–[8]. However, most of these methods mainly emphasized minimizing the domain discrepancy but ignored the underlying data structure. It may lead to poor generalization performance and even the identity misalignment problem [9], that is, images from different domains belonging to different identities might be aligned closer (together) in the feature space, which is disastrous to face recognition.

Enlightened by these observations, this article proposes a deep quadruplet framework (as shown in Fig. 1) to learn domain-invariant discriminative features (DIDF) for HFR. It exploits the underlying data structure information in the semantic label space besides matching distributions in the feature space. Specifically, to mitigate the distribution discrepancy between different sensing modalities, an adversarial training mechanism (ATM) is introduced. It implicitly matches the source and target domain distributions with a bidirectional cross-domain sampling strategy, benefiting to explore the underlying relationships among cross-domain face images. Secondly, according to the characteristic of large cross-domain variations in HFR and metric learning ideas, a special quadruplet loss is proposed, which explicitly constrains the distances between cross-domain positive pairs and negative pairs in the feature space, not only reducing the intra-class variations effectively but also bringing in a global constraint on the inter-class distance. Such that guides the network to further eliminate the cross-domain discrepancy as well as solve the class misalignment and adversarial equilibrium [10], [11] problems confronted by domain-level alignment, enhancing the domain-invariant and class-discrimination of identity features. As shown in Fig. 2(d), with the collaboration of these two components, DIDF aligns the domains well without class mismatch problem and the extracted features are intra-class compact and inter-class separable.

Furthermore, a class-level alignment method based on a special quadruplet loss is developed. It firstly generates a large number of cross-domain image tuples (namely quadruplets) with a bidirectional cross-domain sampling strategy, benefiting to explore the underlying relationships among cross-domain face images. Secondly, according to the characteristic of large cross-domain variations in HFR and metric learning ideas, a special quadruplet loss is proposed, which explicitly constrains the distances between cross-domain positive pairs and negative pairs in the feature space, not only reducing the intra-class variations effectively but also bringing in a global constraint on the inter-class distance. Such that guides the network to further eliminate the cross-domain discrepancy as well as solve the class misalignment and adversarial equilibrium [10], [11] problems confronted by domain-level alignment, enhancing the domain-invariant and class-discrimination of identity features. As shown in Fig. 2(d), with the collaboration of these two components, DIDF aligns the domains well without class mismatch problem and the extracted features are intra-class compact and inter-class separable.

The proposed DIDF method is extensively evaluated on three challenging NIR-VIS databases and one Sketch-Photo database. Comparisons with state-of-the-art HFR methods, along with ablation studies, demonstrates the effectiveness of the DIDF framework.

In summary, the main contributions of this work are:

1) An end-to-end quadruplet architecture is proposed to learn domain-invariant discriminative features for HFR, which optimizes the domain-level alignment and class-level alignment simultaneously in one unified network. To the best of our knowledge, it is the first time that the two kinds of alignment are explicitly considered at the same time in HFR.

2) A class-level alignment method based on a special quadruplet loss is introduced. It not only produces a large
number of cross-domain face images pairs, but also constrains the images with the same identities to be closer to each other in the feature space than those with different identities. With the joint supervision and mutual enhancement of this kind of class alignment method and the domain alignment method, we can reduce the cross-domain discrepancy effectively and enhance the discrimination of learned identity features.

(3) Extensive experiments are conducted on four challenging benchmarks, quantitative comparisons against some state-of-the-art HFR methods demonstrate the effectiveness and superiority of the proposed DIDF method in heterogeneous face recognition.

The rest of this article is organized as follows: Section II briefly reviews some related works on HFR. Then the proposed method is detailed in Section III. In Section IV, we report the experimental results on four commonly adopted HFR datasets. Finally, Section V concludes the paper.

**II. RELATED WORK**

Heterogeneous face recognition has drawn increasing attention in biometrics, existing methods can be roughly divided into three categories [2]–[4]: (i) data synthesis [16]–[20]; (ii) latent subspace learning [21], [22]; (iii) modality-invariant feature learning. Here we only look back on the most related modality-invariant feature learning-based methods; for more details, please refer to [3]. What is more, we briefly review the adversarial domain adaptation approaches and deep metric learning algorithms that are associated with HFR.

**A. MODALITY-INVARIANT FEATURE LEARNING**

Modality-invariant feature learning based HFR methods focus on designing or learning features that are only related to face identities. Early algorithms of this category mainly relied on handcrafted features [2], [3], [6]. With the development of deep learning, numerous CNN-based algorithms have been proposed and shown great potentials in learning domain-invariant features for HFR. [23] explored different strategies using deep CNNs pre-trained on VIS face datasets to solve HFR problems. [7] proposed a multi-view deep network to learn a non-linear discriminant and view-invariant representation shared between multiple views. [4] designed an Invariant Deep Representation (IDR) network integrating invariant feature extraction and subspace learning. [8] learned a non-linear mapping from VIS to the thermal spectrum, bridging the modality gap while retaining the identity information. [6] designed a Wasserstein CNN aiming to minimize the Wasserstein distance between NIR and VIS distributions. [24] utilized a trace norm and a block-diagonal prior to enforce the correlation across distinct modalities for each identity, and employed an across modal ranking to maximize the inter-class margin. In [5], a generative adversarial network was designed to perform cross-spectral face hallucination in the pixel space, and two kinds of loss were employed to reduce the discrepancy between real VIS and generated VIS (generated from NIR) distributions in the feature space. In [2], a Disentangled Variational Representation (DVR) network was employed to reduce the discrepancy between NIR and VIS.

Unlike these methods, we learn domain-invariant embeddings by aligning the feature-level domain distributions in a way of adversarial training. Through the min-max two player game between the feature extractor and an additional domain discriminator, the feature extractor is guided to directly derive feature representations that are invariant between different domains. Moreover, the domain alignment performance is guaranteed and reinforced by our class-level alignment.

**B. ADVERSARIAL DOMAIN ADAPTATION**

Deep adversarial domain adaptation, based on adversarial learning, provides an attractive way to align data distributions. It has been widely applied to mitigate the distribution discrepancy between different domains (tasks). The essence of adversarial domain adaptation [9], [11], [14], [25]–[27] is...
to train a domain discriminator besides the feature extractor, and the two networks are optimized in an adversarial manner. The domain discriminator exerts to distinguish the source domain representations from the target domain ones, while the feature extractor, acting as the generator in typical GAN setting, tries to learn domain-invariant features by fooling the discriminator. Once the domain discriminator is completely confused by the feature extractor, the two domains are considered to be aligned.

Nevertheless, considering only the domain-level alignment without class constraint would result in the semantic misalignment problem [9], e.g., misaligning a target domain image of identity $c_a$ to a source domain image of different identity $c_b$ in the feature space. Another problem is the equilibrium challenge inherent in adversarial learning [10], [11], that is, the two domains are not ensure to be really aligned even if the domain discriminator is fully confused. To circumvent these problems, we present a class-level alignment method, details are in Section III-C.

C. DEEP METRIC LEARNING

To make intra-class samples nearby and inter-class ones far from one another in the feature space, deep metric learning has been widely studied and applied in face recognition [28]–[30], person re-identification [31], [32], and so on. For deep metric learning and face recognition, a lot of loss functions [33]–[41] are proposed. Due to the large cross-modality appearance difference in HFR, extracted features of different modalities usually distributed as separate clusters in the feature space. Thus, the performance of most loss functions developed for single-modality cases (e.g., Softmax [33], Center Loss [36], L-Softmax [37], Angular-Softmax Loss [38], AM-Softmax [39], CosFace [40], and ArcFace [41]) would be seriously influenced. Contrastive loss [34] and triplet loss [35], aiming at constraining the distances of matching pairs and non-matching pairs, are naturally reasonable metrics for the cross-modality problem and have been applied to HFR in [15], [24], [42]–[44]. However, Contrastive loss and triplet loss can only focus on one negative pair when selecting each training sample, lacking the ability to distinguish other classes of samples. Such that they fail to achieve reliable generalization [31], [32], [45]. For example, the model trained with a triplet loss would still produce relatively large intra-class variations and small inter-class differences when applied to unseen target testing identities, leading to bad separation and even misclassification (as shown in Fig. 2(c)).

Afterward, [32] improved the triplet loss and proposed the quadruplet loss by introducing an additional distance loss to triplet loss. The added loss aims to optimize the distance between the positive and negative pairs with no common anchor. Recently, [45] generalized the cases whether the negative pairs have a shared anchor with the positive pair, and proposed a new version of quadruplet loss. [46] applied the quadruplet loss to the hierarchical clothing retrieval task. Despite the success, the problems of [32], [45], [46] are that: i) they can only focus on the local (class-level) relationships among instances rather than the global (domain-level) data distributions. ii) during the process of selecting quadruplet pairs, they did not consider the challenging application scenarios where image samples come from different distributions (domains). These questions would cause their performance to be limited by the notable cross-modality gap.

In contrast, we are essentially concerned about the cross-modality situations in face recognition, and design a special quadruplet loss for HFR. This quadruplet loss explicitly constrains the largest cross-modality intra-class distance to be smaller than the smallest (cross-domain and within-domain) inter-class distance in embedding space. Therefore, it reduces the intra-class variations as well as increases the inter-class separability more efficiently and effectively, improving the discrimination of identity features. Moreover, the mining of quadruplet pairs is bidirectional, which helps to explore the underlying commonness and difference among images and makes fine-tuning deep models on small data-sets possible. We combine this kind of class-level alignment method with the domain-level alignment method to align the source and target domain features at both global and local levels.

III. PROPOSED METHOD

The proposed deep framework for HFR consists of three primary modules: a feature extraction module, a class alignment module, and a domain discriminator. The CNN-based feature extraction module is trained to minimize the class alignment error by collaborating with the class alignment module, and to maximally confuse the domain discriminator; the domain discriminator is trained adversarial with the feature extraction module to classify the domain (source or target).

A. PROBLEM FORMULATION

Let $X_i$ and $X_j$ be the source (VIS / Photo) and target (NIR / Sketch) domain images respectively. They share the same feature space but different marginal data distributions $P(X_i)$ and $P(X_j)$ ($P(X_i) \neq P(X_j)$, as shown in Fig. 4). We describe the CNN feature extraction process as $f_i = G_f(x_i; \theta_f)$, where $G_f$ is the feature extractor defined by convolution neural network with parameters $\theta_f$, $x_i \in \{X_i \cup X_j\}$ is the $i$th training sample, $f_i \in \mathbb{R}^d$ is the face representation. Each $x_i$ has an identity label $c_i$, and a binary domain label $d_i$, indicating whether $x_i$ comes from the source domain ($d_i = 0$) or the target domain ($d_i = 1$). The goal of the proposed method is to extract face representations that are domain-invariant and identity-discriminative.

B. DOMAIN-LEVEL ALIGNMENT

Our domain-level alignment aims at eliminating the domain-related information and generating domain-invariant face representations. That is, making the distributions $P(f_i) = \{G_f(x_i; \theta_f) \mid x_i \sim P(X_i)\}$ and $P(f_j) = \{G_f(x_i; \theta_f) \mid x_i \sim P(X_j)\}$ to be aligned. We follow the idea of adversarial domain adaptation, as [11], [14], [25]–[27]. Specifically, we train a
domain discriminator $G_d$ (parameterized with $\theta_d$) along with the feature extractor $G_f$ in a min-max way. $G_d$ is optimized to minimize the domain classification loss as defined in (1), whereas $G_f$ is trained to maximize the domain objective loss:

$$L_d = \frac{1}{N} \sum_{i=1}^{N} [-d_i \log (G_d (f_i; \theta_d)) - (1 - d_i) \log (1 - G_d (f_i; \theta_d))].$$

(1)

Once $G_d$ is fully confused by $G_f$, determining the source of the representations by random guess, it is consider that $P(f_i)$ and $P(f_j)$ are aligned. The advantage of this method is that it does not struggle for the concrete distribution form of $P(f_i)$ and $P(f_j)$, nor defining a suitable (dis)similarity measurement metric between $P(f_i)$ and $P(f_j)$.

The adversarial domain adaption approach could achieve a good global alignment between the source and the target domain representations, thus eliminating the domain discrepancy between them. However, it suffers from class misalignment and adversarial equilibrium challenges. Therefore, we introduce a novel quadruplet-based class alignment method, which complements and promotes mutually with the adversarial domain-level alignment. The quadruplet function not only enlarges the training set, but also constrains the distance among instances, which is beneficial to further reduce the global discrepancy and ensure different domains well aligned. As confirmed in the experiments, the joint optimization between these two components obtains state-of-the-art performance, better than using each alone.

C. QUADRUPLET BASED CLASS ALIGNMENT

The quadruplet based class alignment method is introduced to explore the relationships among the samples, align the cross-modality representations belonging to the same identity closely and clearly separate the representations belonging to different identities. To this end, we construct quadruplets $\{(x_q, x_p, x_n^1, x_n^2)\}$ as following: We firstly set a target domain image as an anchor $x_a$, a source domain image having the same identity as a positive example $x_p$, and another two source domain images as negative examples $x_n^1$ and $x_n^2$, respectively, where the identities of $x_p$, $x_n^1$, and $x_n^2$ are different from one another (see Fig. 3). Then, to balance the importance of the source and target domain images and further augment the training samples, we inversely generate quadruplets whose $x_a$ is from the source domain, but $x_p$, $x_n^1$, and $x_n^2$ are from the target domain. That is, $(x_a, x_p)$ and $(x_a, x_n^1)$ are the cross-modality positive and negative pairs respectively, $(x_n^1, x_n^2)$ is another same-domain negative pair. The quadruplet loss can be formulated as:

$$L_{quad} = \frac{1}{N} \sum_{a,p,n_1,n_2}^{N} \left( (m_1 + D_{a,p} - D_{a,n_1})^+ \right. \left. + \gamma (m_2 + D_{a,p} - D_{n_1,n_2})^+ \right)$$

$$D_{i,j} = 1 - \frac{f_i \cdot f_j}{\|f_i\| \cdot \|f_j\|},$$

(2)

where $m_1$ and $m_2$ are margins enforced between positive and negative pairs, $D_{i,j}$ is the cosine distance between two features $f_i$ and $f_j$, $[.]^+ = \max(., 0)$, and $\gamma$ is a hyper parameter between 0 and 1 to control the intensity of the second term. As shown in (2), the first term is the commonly used triplet loss, which concentrates on the relative distances between matching and non-matching pairs. The second term brings in a new constraint, which enforces the smallest inter-class distance to be larger by a margin of $m_2$ than the largest intra-class distance. It helps to further improve the inter-class separability and boost the generalization performance.

However, the possible number of quadruplets is overwhelming even for a small dataset, optimizing all quadruplets is unfeasible in computation and training time. On the other hand, random selection of quadruplet tuples are easy to satisfy the constraint in (2) and thus less contribute to the training. Hence, how to select of suitable quadruplets is critical. In this article, we select quadruplet tuples at each training epoch. We set $(x_a, x_p)$ as the most dissimiliar cross-domain positive pair, $(x_a, x_n^1)$ as the most similar cross-domain negative pair. A formula expression is:

$$\begin{align*}
\{x_p \in \{x_i|c_i = c_a, d_i \neq d_a\} \\
x_p = \arg \max_{x_p} D_{a,p}
\end{align*}$$

$$\begin{align*}
\{x_n^1 \in \{x_i|c_i \neq c_a, d_i \neq d_a, D_{a,p} + m_1 > D_{a,n_1}\} \\
x_n^1 = \arg \min_{x_n^1} D_{a,n_1},
\end{align*}$$

(3)

where the parameter $m_1$ is the first margin in (2). As to $(x_n^1, x_n^2)$, we randomly choose one of the margin-based hard samples, which meets the following requirements:

$$\begin{align*}
D_{a,p} + m_2 > D_{n_1,n_2} \\
c_{n_2} \neq c_a, c_{n_2} \neq c_p, c_{n_2} \neq c_{n_1} \\
d_{n_2} \neq d_a, d_{n_2} = d_p = d_{n_1},
\end{align*}$$

(4)

where the parameter $m_2$ is the second margin in (2).

The benefits of this choice are twofold: i) The bidirectional cross-domain quadruplet sampling method can generate positive and negative training samples, thus mine the underlying differences and commonalities among source and target domain images more effectively. ii) With the constraint of cross-domain quadruplet loss, the network pays more attention to the individual distinction, so that the cross-domain
shift can be further weakened or even eliminated, the learned features would be much more discriminative.

**D. OVERALL OBJECTIVE**

In the presented framework for HFR, the feature extraction module is designed to receive batches of training face samples \( \{x_d, x_p, x_n, x_s\} \) as input, and output feature representations to the class alignment module \( G_c \) (parameterized by \( \theta_c \)) and also to the domain discriminator \( G_d \). The extraction module and discriminator \( G_d \) are trained adversarially by playing a two-player min-max game, where the extraction module tries to maximize the domain discriminant loss (and minimize the class alignment loss at the same time) to fool the discriminator such that the discriminator is unable to distinguish from which domain the feature representations come; while the discriminator focuses on minimizing the domain discriminant loss to correctly discriminate the domain source. Formally, we formulate the minimax game with a value function as follow:

\[
\min_{G_f, G_c} \max_{G_d} V(G_f, G_d) = L(\theta_f, \theta_c, \theta_d) = L_{\text{quad}} - \lambda L_d, \quad (5)
\]

where the parameter \( \lambda \) balances the trade-off between identity prediction loss and domain discriminant loss. The goal of adversarial learning is to optimize the parameters \( (\theta_f, \theta_c, \theta_d) \):

\[
(\theta_f^*, \theta_c^*, \theta_d^*) = \arg \min_{\theta_f, \theta_c, \theta_d} L(\theta_f, \theta_c, \theta_d). \quad (6)
\]

At the saddle point, the extracted features \( f_i = G_f(x_i; \theta_f^*) \) \((i = 1, 2, \ldots, n)\) are discriminative in their identities and indistinguishable in their domain distributions. The adversarial learning is optimized with standard stochastic gradient (SGD), by adding a gradient reversal layer (GRL) before the domain discriminator (see Fig. 1). The idea of GRL is proposed by [14]. GRL functions as an identity transformation in forward propagation. While during backward propagation, it transmits the gradient from the subsequent layer to the previous layer after multiplying it by a factor of \(-\lambda\).

**E. NETWORK ARCHITECTURE**

The proposed framework is composed of three key components: i) feature extraction module: it contains four branches of shared feature extractors, where LightCNN-9 and LightCNN-29 [12] are used as backbones. ii) class alignment module: it is built by a “FC + MFM” layer (8192 \( \rightarrow \) 256), which maps the identity representation \( f_i \) to a 256-D feature vector under the supervision of the loss function defined in (2). MFM (Max-Feature-Map) is an activation function proposed by [12]. It adopts a competitive relationship rather than a threshold (or bias) to activate a neuron, demonstrated by [12] to be very powerful in feature selection and suitable for different CNN architectures. It is worth noting that the 256-D identity feature vectors are directly utilized for face comparison in both the training and testing phase, without any fully connected (FC) classifier layers followed. The main reason is that extra FC layer(s) would lead to a large increase in network parameters, which even grows with the number of subjects in training data, increasing the difficulty of network training. iii) domain discriminator: it consists of “FC + MFM + FC” layers (8192 \( \rightarrow \) 256 \( \rightarrow \) 2), with a loss function as shown in (1). We train these three parts jointly in an end-to-end manner, with the weighted loss function defined in (5).

**IV. EXPERIMENTAL RESULTS**

In this section, the proposed HFR framework is systemically evaluated against several state-of-the-art HFR methods on four widely used HFR benchmarks. Besides, ablation studies are implemented to explore the importance of each component in the framework.

**A. DATASETS AND PROTOCOLS**

1) THE CASIA NIR-VIS 2.0 FACE DATABASE [47]

Being the largest publicly available and most challenging (large variations in eyeglasses, distance, lighting, expression, and pose) NIR-VIS face recognition database, it has been an important benchmark dataset for NIR-VIS HFR evaluation. There are a totally of 725 persons in this database, with 1 to 22 VIS and 5 to 50 NIR images per person. Since all images are randomly collected, there is no one-to-one correspondence between NIR and VIS images. Fig. 4(a) lists some samples of aligned NIR and VIS faces.

For fair comparisons against other methods, we follow the standard training and testing protocol, which contains 10-fold experiments and each fold includes training and testing lists. Nearly equal numbers of identities are included in the training and testing lists, and they are kept disjoint from each other. In each fold, there are about 6100 NIR images and 2500 VIS images from about 360 identities for training; as to testing, the gallery comprises of 358 identities and each has only one VIS image, the probe has over 6000 NIR images from the same 358 subjects. All the NIR images in the probe set are matched with the VIS images in the gallery set. Rank-1 accuracy and verification rate (VR, equal means TPR) at false acceptance rate (FAR) are used as the evaluation metric.

2) THE OULU-CASIA NIR&VIS FACIAL EXPRESSION DATABASE [48]

This database consists a total of 80 subjects between 23 to 58 years old coming from Oulu University and CASIA. All of them are captured under three different illumination conditions (normal, weak, and dark) with six diverse expressions (anger, disgust, fear, happiness, sadness, and surprise). Fig. 4(b) displays some cropped VIS and NIR face samples. Following the protocols in [6], we randomly select 48 NIR images and 48 VIS images (eight face images from each expression) for each person, and we randomly choose 20 persons for training and another 20 persons for testing.
All the NIR and VIS images in the testing are used as the probe and gallery respectively. The Rank-1 accuracy, VR@FAR = 1%, and VR@FAR = 0.1% are reported.

3) THE BUAA-VisNir DATABASE [49]
This database contains 150 persons, each with 9 VIS images and 9 NIR face images. Some cropped VIS and NIR samples are illustrated in Fig. 4(c). Following the protocols in [2], [6], it is randomly divided into the training set with 50 persons and the testing set with the remaining 100 persons. For the testing set, only one VIS image of each person is selected into the gallery set, and all NIR images are used as the probe. The performances of rank-1 accuracy, VR@FAR = 1%, and VR@FAR = 0.1% are used as evaluation metrics.

4) THE IIIT-D VIEWED SKETCH DATABASE [50]
It collects 238 sketch-digital image pairs, where the sketches are drawn by a professional sketch artist according to the corresponding digital images. Fig. 4(d) exhibits some samples of cropped sketch and photo faces. Following the experimental setup in [24], we use the CUHK Face Sketch FERET (CUFSF) Database [51] (a viewed sketch-photo face database that includes 1194 persons, each has one sketch-photo image pair) as the training dataset and conduct probe-gallery face identification test on the IIIT-D Viewed Sketch Database. The Rank-1 accuracy and VR@FAR = 1% are reported for comparisons.

B. IMPLEMENTATION DETAILS
For fair comparisons with state-of-the-art HFR methods, two different networks, LightCNN-9 and LightCNN-29, are employed as the backbone, respectively, which are pre-trained on the MS-Celeb-1M dataset [13]. In detail, four branches of LightCNN-9 (LightCNN-29) with the FC layers removed are used as the feature extraction module. The class alignment module is initialized with the first FC layer of LightCNN-9 (LightCNN-29). The domain discriminator is randomly initialized. All HFR images are converted to grayscale, aligned to 144×144 according to five facial landmarks, and center-cropped to 128×128. Specially, HFR training images are horizontal mirror augmented to fine-tune the proposed DIDF framework.

DIDF is implemented on Pytorch with python 3.6. The parameters $m_1$, $m_2$, and $γ$ are empirically set to 0.7, 0.6, and 0.5, respectively; the value of $λ$ is changed from 0.0 to 1.0 with the learning progress, with a schedule like that in [14]. Additionally, Stochastic gradient descent (SGD) is used for training, with the momentum of 0.9 and weight decay of $1e^{-4}$. The dropout ratio for the FC layers are set to 0.5, the learning rate is set to $1e^{-4}$.

C. RESULTS ON THE CASIA NIR-VIS 2.0 DATABASE
On the challenging CASIA NIR-VIS 2.0 database, the performance of the proposed method is compared with some state-of-the-art HFR methods, including traditional methods Gabor + JB [52], Gabor + RBM [54], Recon. + UDP [53], Gabor + HJB [55], and CDFL [56], as well as recently proposed deep CNN-based methods IDNet [42], HFR-CNN [23], Hallucination [16], TRIVET [15], IDR [4], CDL [24], W-CNN [6], ADFL [5], DVR [2], DVG [18], and Bae’s [57]. Since most of these methods follow the standard protocols of this database, the results reported in Table 1 are directly from their published papers.

| Method              | Rank-1 (%) | VR@FAR=0.1% (%) |
|---------------------|------------|-----------------|
| Gabor+JB [52]       | 89.5 ± 0.8 | 83.2 ± 1.0      |
| Recon.+UDP [53]     | 78.5 ± 1.7 | 85.8            |
| Gabor+RBM [54]      | 86.2 ± 1.0 | 81.3 ± 1.8      |
| Gabor+HJB [55]      | 91.6 ± 0.8 | 89.9 ± 0.9      |
| CDL [56]            | 75.1 ± 1.4 | 55.10           |
| IDNet [42]          | 87.1 ± 0.9 | 74.5            |
| HFR-CNN [23]        | 85.9 ± 0.9 | 78.0            |
| Hallucination [16]  | 86.9 ± 0.9 |                |
| TRIVET [15]         | 95.7 ± 0.5 | 91.0 ± 1.3      |
| IDR [4]             | 97.3 ± 0.4 | 95.7 ± 0.7      |
| CDL [24]            | 98.6 ± 0.2 | 98.3 ± 0.1      |
| W-CNN [6]           | 98.7 ± 0.3 | 98.4 ± 0.4      |
| ADFL [5]            | 98.2 ± 0.3 | 97.2 ± 0.3      |
| DVR(LightCNN-9) [2] | 99.1 ± 0.2 | 98.6 ± 0.2      |
| DVG(LightCNN-9) [18] | 99.2 ± 0.2 | 98.8 ± 0.2    |
| DIDF(LightCNN-9)    | 99.5 ± 0.2 | 99.1 ± 0.2      |
| Bae’s [57]          | 99.4       | 98.7            |
| DVR(LightCNN-29) [2] | 99.7 ± 0.1 | 99.6 ± 0.3      |
| DVG(LightCNN-29) [18] | 99.8 ± 0.1 | 99.8 ± 0.1      |
| DIDF(LightCNN-29)   | 99.8 ± 0.1 | 99.8 ± 0.1      |

As can be seen in Table 1, our DIDF method outperforms all handcrafted feature-based methods, e.g. compared with Gabor + HJB, DIDF has a superiority of 8.2% on Rank-1 accuracy and 9.9% on VR@FAR = 0.1%. For fair comparisons with the deep CNN-based competitors,
including IDNet, HFR-CNN, Hallucination, TRIVET, IDR, CDL, W-CNN, ADFL, DVR(LightCNN-9), and DVG(LightCNN-9), we evaluate the proposed DIDF method on the LightCNN-9 backbone. It acquires 99.5% on Rank-1 accuracy and 99.1% on VR@FAR = 0.1%, outperforming the previous best method DVG on the same backbone, e.g., 99.2% vs. 99.2% on Rank-1 accuracy and 99.1% vs. 98.8% on VR@FAR = 0.1%. Especially, based on DVG, DIDF relatively reduces the error rate (1-VR) at FAR = 0.1% by 25%. When verified on the LightCNN-29 backbone, same as DVR(LightCNN-29) and DVG(LightCNN-29), DIDF achieves 99.8% on both Rank-1 accuracy and VR@FAR = 0.1%, the best performance on this database. All these experimental results suggest that the proposed method can learn domain-invariant discriminative face representations and is effective for NIR-VIS heterogeneous face recognition.

### D. RESULTS ON THE OULU-CASIA NIR&VIS DATABASE

In this subsection, the performance of our proposed method is verified on the Oulu-CASIA NIR&VIS database and compared with KDSR [58], H2(LBP3) [59], TRIVET [15], IDR [4], CDL [24], W-CNN [6], ADFL [5], DVR [2], and DVG [18]. Except that the results of TRIVET, IDR, and CDL are from [2], all others are directly from their published papers.

From Table 2 we can observe that the proposed DIDF method on the LightCNN-9 backbone obtains 100.0% on Rank-1 accuracy, 97.8% on VR@FAR = 1%, and 89.9% on VR@FAR = 0.1%, which outperforms traditional methods KDSR and H2(LBP3), and deep learning based competitors, such as TRIVET, IDR, CDL, W-CNN, ADFL, DVR(LightCNN-9), and DVG(LightCNN-9). On the LightCNN-29 backbone, DIDF further improves VR@FAR = 1% to 99.0% and VR@FAR = 0.1% to 93.3%, outperforming the state-of-the-art methods DVR and DVG on the same backbone. Based on DVG, the error rates at FAR = 1% and FAR = 0.1% are relatively reduced by 37.5% and 5.6%, respectively.

#### TABLE 2. Rank-1 Accuracy (%) and Verification Rate (%) on the Oulu-CASIA NIR&VIS Database.

| Method            | Rank-1 | FAR@1% | FAR@0.1% |
|-------------------|--------|--------|----------|
| KDSR [58]         | 66.9   | 56.1   | 31.9     |
| H2(LBP3) [59]     | 70.8   | 62.0   | 33.6     |
| TRIVET [15]       | 92.2   | 67.9   | 33.6     |
| IDR [4]           | 94.3   | 73.4   | 46.2     |
| CDL [24]          | 94.3   | 81.6   | 53.9     |
| W-CNN [6]         | 98.0   | 81.5   | 54.6     |
| ADFL [5]          | 95.5   | 83.0   | 60.7     |
| DVR(LightCNN-9) [2] | 99.3 | 89.7   | 65.8     |
| DVG(LightCNN-9) [18] | 100.0 | 97.5   | 89.9     |
| DIDF(LightCNN-9)   | 100.0 | 97.8   | 89.9     |
| DVR(LightCNN-29) [2] | 100.0 | 97.2   | 84.9     |
| DVG(LightCNN-29) [18] | 100.0 | 98.4   | 92.9     |
| DIDF(LightCNN-29) | 100.0 | 99.0   | 93.3     |

The prominent performance of our method may attribute to the quadruplet loss to some extent. Due to the small-scale training set of this database, the other referenced deep learning methods, such as TRIVET, IDR, CDL, W-CNN, ADFL, and DVR, may be potentially affected. On contrast, our method can produce a relatively large number of training samples with the help of the bidirectional cross-domain quadruplets, such that captures the intrinsic domain-invariant discriminative identity features more effectively. These results demonstrate the effectiveness and superiority of our method in reducing the NIR-VIS modality discrepancy.

### E. RESULTS ON THE BUAA-VisNir DATABASE

On this database, the proposed method DIDF is compared with the perviously proposed KDSR [58], H2(LBP3) [59], TRIVET [15], IDR [4], CDL [24], W-CNN [6], ADFL [5], DVR [2], and DVG [18] methods. Comparison results are tabled in Table 3, in which we can observe that: on both the LightCNN-9 and the LightCNN-29 backbones, DIDF exceeds all its competitors, including the traditional ones (located in the second part of the table, from top to down), and the deep CNN-based ones (located in the third and fourth parts of the table, from top to down). Especially, DIDF outperforms the previous best method DVR on the LightCNN-9 backbone by a margin of 1.7% on Rank-1 accuracy, 2.4% on VR@FAR = 1%, and 3.6% on VR@FAR = 0.1%. Besides, its advantages over DVG on the LightCNN-29 backbone are 0.6% on Rank-1 accuracy, 1.2% on VR@FAR = 1%, and 1.7% on VR@FAR = 0.1%. These results again indicate the effectiveness and advancement of the proposed method for the NIR-VIS HFR problem.

#### TABLE 3. Rank-1 Accuracy (%) and Verification Rate (%) on the BUAA-VisNir Database.

| Method            | Rank-1 | FAR@1% | FAR@0.1% |
|-------------------|--------|--------|----------|
| KDSR [58]         | 83.9   | 86.5   | 69.5     |
| H2(LBP3) [59]     | 88.8   | 88.8   | 73.4     |
| TRIVET [15]       | 93.9   | 93.0   | 80.9     |
| IDR [4]           | 94.3   | 93.4   | 84.7     |
| CDL [24]          | 96.9   | 95.9   | 90.1     |
| W-CNN [6]         | 97.4   | 96.0   | 91.9     |
| ADFL [5]          | 95.2   | 95.3   | 88.0     |
| DVR(LightCNN-9) [2] | 97.9 | 97.0   | 92.8     |
| DVG(LightCNN-9) [18] | 98.0 | 97.2   | 93.0     |
| DIDF(LightCNN-9)   | 99.7   | 99.6   | 96.6     |
| DVR(LightCNN-29) [2] | 99.2 | 98.5   | 96.9     |
| DVG(LightCNN-29) [18] | 99.2 | 98.5   | 97.3     |
| DIDF(LightCNN-29) | 99.8   | 99.7   | 99.0     |

### F. RESULTS ON THE IIIT-D VIEWED SKETCH DATABASE

In this subsection, we evaluate the proposed method on the IIIT-D viewed sketch-photo face recognition database, against some state-of-the-art HFR methods, including conventional handcrafted feature-based methods such as Original WLD [60], SIFT [61], EUCLBP [62], LFDA [63], and MCWLD [64], as well as deep learning based methods like VGG [65], CenterLoss [36], LightCNN [12], CDL [24], and DVG [18]. Table 4 shows the comparison results of Rank-1 accuracy and VR@FAR = 1%, in which it is obvious that the proposed DIDF method exceeds all the conventional methods, such as TRIVET, IDR, CDL, W-CNN, ADFL, and DVR, by a margin of 1.8% on Rank-1 accuracy, 2.5% on VR@FAR = 1%, and 3.7% on VR@FAR = 0.1%. Besides, its advantages over DVG on the LightCNN-29 backbone are 0.7% on Rank-1 accuracy, 1.5% on VR@FAR = 1%, and 1.8% on VR@FAR = 0.1%. These results again indicate the effectiveness and advancement of the proposed method for the IIIT-D HFR problem.

#### TABLE 4. Rank-1 Accuracy (%) and Verification Rate (%) on the IIIT-D Viewed Sketch Database.

| Method            | Rank-1 | FAR@1% | FAR@0.1% |
|-------------------|--------|--------|----------|
| Original WLD [60] | 64.9   | 48.2   | 41.5     |
| SIFT [61]         | 70.8   | 58.6   | 51.3     |
| EUCLBP [62]       | 98.0   | 81.5   | 54.6     |
| LFDA [63]         | 95.5   | 83.0   | 60.7     |
| MCWLD [64]        | 99.3   | 89.7   | 65.8     |
| VGG [65]          | 100.0  | 97.5   | 89.3     |
| CenterLoss [36]   | 100.0  | 97.8   | 89.9     |
| LightCNN [12]     | 100.0  | 97.2   | 84.9     |
| CDL [24]          | 100.0  | 98.4   | 92.9     |
| DVG(LightCNN-9) [18] | 100.0 | 99.0   | 93.3     |
methods, e.g., the Rank-1 accuracy on the LightCNN-29 backbone is 11.56% higher than that of MCWLD. Besides, DIDF outperforms the previous best deep method DVG on the same LightCNN-9 backbone by a margin of 1.61% on Rank-1 accuracy and 2.3% on VR@FAR = 1%, and achieves comparable performance with DVG on the LightCNN-29 backbone. These experimental results demonstrate the effectiveness and potential of DIDF in learning modality-invariant discriminative representations for sketch-photo face recognition.

G. ABLATION STUDY
To investigate the contribution of supervision terms domain-level and class-level alignment, we conduct experiments on the CASIA NIR-VIS 2.0 Database. We mark the baseline network (LightCNN-9, trained on the MS-Celeb-1M dataset [13]) as “B(Basel.),” and we tag the method fine-tuning with softmax loss as “Softmax.” Similarly, “B + D” and “B + C” signify fine-tuning with domain-level and class-level alignment constraints respectively, and DIDF is our full system (fine-tuning with both domain-level and class-level alignment supervision).

1) EFFECTS ON THE CROSS-DOMAIN DISCREPANCY
In domain adaptation theory, $\mathcal{A}$-distance is introduced to measure the distribution discrepancy between two domains [66]–[68]. The $\mathcal{A}$-distance is defined as:

$$d_{\mathcal{A}}(D_s, D_t) = 2(1 - err(h)), \quad (7)$$

where $D_s$ and $D_t$ represent the source and target domains respectively, $err(h)$ is the classification loss of a classifier $h$ trained to distinguish whether a feature belongs to the source or target domain. We report the $\mathcal{A}$-distance on features of different models (“B(Basel.),” “Softmax,” “B + D,” “B + C,” and DIDF) in Fig. 5. From which, we can observe that the $\mathcal{A}$-distance on “B + D,” “B + C,” and DIDF features become smaller and smaller, and all are smaller than that on “B(Basel.)” and “Softmax.” This implies that all these three models can reduce the distribution discrepancy. Especially, DIDF is most effective, declining the discrepancy to the greatest extent. All these results demonstrate the effect and significance of the collaboration between domain-level and class-level alignment (having a more exceptional ability to reduce the cross-modality discrepancy). For the larger value of “B + D” than “B + C,” it may be ascribed to two reasons: 1) The performance of “B + D” is depressed by the limited training data; 2) As explained above, cross-domain quadruplet loss focuses on reducing the intra-class cross-modal distance and enlarging inter-class distance, thus facilitates decreasing the distribution discrepancy between domains. Furthermore, it manages to enrich the training samples effectively.

2) EFFECTS ON FEATURE DISTRIBUTIONS
Fig. 6 shows the distributions of intra-class and inter-class distances on training and testing set from models trained with different supervised terms. From which we can observe that compared with “B(Basel.),” “Softmax,” the models “B + D,” “B + C,” and DIDF gradually obtain decreasing average intra-class distance and increasing average inter-class distance, indicating that both domain-level and class-level alignment help learning intra-class compact and inter-class separable features. Particularly, their combination DIDF is most resultful. The features learned by DIDF have the best intra-class compactness as well as the best inter-class separation. These results further validate the effectiveness of the combination of domain-level and class-level alignment.

3) EFFECTS ON RECOGNITION PERFORMANCE
The recognition results of different modes are presented in Table 5 and Fig. 7. We can find in Table 5 that the methods can be nearly ordered in ascending Rank-1 accuracy (VR@FAR = 1% and VR@FAR = 0.1%) as “B(Basel.),” “Softmax,” “B + D,” “B + C,” and DIDF. DIDF raises the Rank-1 accuracy from 92.5% to 99.5%, VR@FAR = 0.1% from 88.5% to 99.1%, and VR@FAR = 0.01% from 70.6% to 97.1% by comparison with the baseline.

| Method       | Rank-1 | VR@FAR=1% |
|--------------|--------|-----------|
| Original WLD | 74.34  | -         |
| SIFT         | 76.28  | -         |
| EUCLBP       | 79.36  | -         |
| LFDA         | 81.43  | -         |
| MCWLD        | 84.24  | -         |
| VGG          | 80.89  | 72.08     |
| CenterLoss   | 84.07  | 76.20     |
| LightCNN-9   | 84.07  | 75.30     |
| CDL          | 85.35  | 82.52     |
| DVG          | 86.63  | 92.24     |
| DIDF(LightCNN-9) | 88.24 | 94.54     |
| LightCNN-29  | 83.24  | 81.04     |
| DVG(LightCNN-29) | 96.98 | 97.84     |
| DIDF(LightCNN-29) | 95.80 | 98.32     |
TABLE 5. Rank-1 Accuracy (%) and VR@FAR (%) of Models With Different Supervised Terms Evaluated on the CASIA NIR-VIS 2.0 Database.

| Method     | Rank-1 | FAR@0.1% | FAR@0.01% |
|------------|--------|----------|-----------|
| B(Basel.)  | 92.5   | 88.5     | 70.6      |
| Softmax    | 97.9   | 96.5     | 90.6      |
| B + D      | 98.6   | 97.3     | 92.3      |
| B + C      | 99.4   | 98.9     | 96.2      |
| DIDF       | 99.5   | 99.1     | 97.1      |

V. CONCLUSION

This article proposes a novel framework for heterogeneous face recognition, integrating domain-level and class-level alignment in one unified network. As far as we know, it is the first time in HFR that the two levels of optimization are explicitly considered simultaneously. The domain-level alignment reduces the cross-domain distribution discrepancy in an adversarial learning manner. The class-level alignment based on a specially designed quadruplet loss boosts the feasibility of fine-tuning deep models on small datasets and improves the discrimination of identity features. These two components have been proved to be essential, working together and mutually reinforcing. Extensive experiments on
four challenging HFR benchmark databases and two different backbones demonstrate the effectiveness and superiority of the proposed DIDF framework in learning domain-invariant discriminative features for HFR. Besides, DIDF is a general framework in which inner modules can be replaced/improved for other problems such as pose/lighting/expression-invariant face recognition.

REFERENCES

[1] M. Wang and W. Deng, “Deep face recognition: A survey,” 2018, arXiv:1804.06655 [Online]. Available: http://arxiv.org/abs/1804.06655
[2] 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., vol. 33, 2019, pp. 9005–9012.
[3] 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.
[4] R. He, X. Wu, Z. Sun, and T. Tan, “Learning invariant deep representation for NIR-VIS face recognition,” in Proc. 31st AAAI Conf. Artif. Intell., 2017, pp. 2000–2006.
[5] L. Song, M. Zhang, X. Wu, and R. He, “Adversarial discriminative heterogeneous face recognition,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 7355–7362.
[6] 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.
[7] M. Kan, S. Shan, and X. Chen, “Multi-view deep network for cross-view classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4847–4855.
[8] M. S. Sarfraz and R. Stiefelhagen, “Deep perceptual mapping for cross-modal face recognition,” Int. J. Comput. Vis., vol. 122, no. 3, pp. 426–438, Mar. 2017.
[9] W. Deng, L. Zheng, and J. Jiao, “Domain alignment with triplets,” 2018, arXiv:1812.00893 [Online]. Available: https://arxiv.org/abs/1812.00893
[10] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, “Generalization and equilibrium in generative adversarial nets (gans),” in Proc. 34th Int. Conf. Mach. Learn., 2017, pp. 224–232.
[11] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, “Maximum density divergence for domain adaptation,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Apr. 28, 2020, doi: 10.1109/TIP.2020.2991050.
[12] 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.
[13] Y. Guo, Z. 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.
[14] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by back-propagation,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1180–1189.
[15] 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.
[16] 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. 28–66.
[17] B. Cao, N. Wang, J. Li, and X. Gao, “Data augmentation-based joint learning for heterogeneous face recognition,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 6, pp. 1731–1743, Jun. 2019.
[18] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, “Dual variational generation for low shot heterogeneous face recognition,” in Adv. Neural Inf. Process. Syst., 2019, pp. 2674–2683.
[19] R. He, J. Cao, J. 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.
[20] M. Zhu, J. Li, N. Wang, and X. Gao, “A deep collaborative framework for face Photo–Sketch synthesis,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 10, pp. 3096–3108, Oct. 2019.
[21] Y. Jin, J. Li, C. Lang, and Q. Ruan, “Multi-task clustering ELM for VIS-NIR cross-modal feature learning,” Multidimensional Syst. Signal Process., vol. 28, no. 3, pp. 905–920, Jul. 2017.
