A Unified Multi-Domain Face Normalization Framework for Cross-Domain Prototype Learning and Heterogeneous Face Recognition

Meng Pang, *Member, IEEE*, Wenjun Zhang, Yang Lu, *Senior Member, IEEE*, Yiu-ming Cheung, *Fellow, IEEE*, and Nanrun Zhou

***Abstract*—Face normalization is a critical technique for improving the robustness and generalizability of face recognition systems by reducing intra-personal variations arising from expressions, poses, occlusions, illuminations, and domain shifts. Existing normalization methods, however, often lack the flexibility to handle multi-factorial variations and exhibit limited cross-domain adaptability. To address these challenges, we propose a Unified Multi-Domain Face Normalization Network (UMFN), which is designed to process facial images with diverse variations from various domains and reconstruct frontal, neutralized facial prototypes in the target domain. As an unsupervised domain adaptation model, the UMFN facilitates concurrent training across multiple cross-domain datasets and demonstrates robust prototype reconstruction capabilities. Notably, the UMFN functions as a joint prototype and feature learning framework, extracting domain-agnostic identity features through a decoupling mapping network and adversarial training with a feature domain classifier. Furthermore, we design an efficient Heterogeneous Face Recognition (HFR) network that integrates these domain-agnostic features and the identity-discriminative features extracted from normalized prototypes, enhanced by contrastive learning to improve identity recognition accuracy. Empirical evaluation on multiple cross-domain benchmark datasets validate the effectiveness of the UMFN for face normalization and the superiority of the HFR network for heterogeneous face recognition.**

***Index Terms*—Face normalization, cross-domain prototype learning, heterogeneous face recognition, adversarial learning, contrastive learning.**

Received 12 December 2024; revised 29 March 2025 and 5 May 2025; accepted 5 May 2025. Date of publication 14 May 2025; date of current version 4 June 2025. This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant 62466036, Grant 62162041, and Grant 62376233; in part by NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21; in part by the General Research Fund of RGC under Grant 12202622 and Grant 12202924; in part by RGC Senior Research Fellow Scheme under Grant SRFS2324-2S02; in part by the Science and Technology Planning Project of Shanghai under Grant 23010501800; in part by the High-Level and Urgently Needed Overseas Talent Programs of Jiangxi Province under Grant 20232BCJ25024; in part by the Natural Science Foundation of Jiangxi Province under Grant 20232BAB212025; in part by the Natural Science Foundation of Fujian Province under Grant 2024J09001; and in part by the Xiaomi Young Talents Program. The associate editor coordinating the review of this article and approving it for publication was Prof. Zhen Lei. *(Corresponding author: Nanrun Zhou.)*

Meng Pang and Wenjun Zhang are with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, China (e-mail: pangmeng1992@gmail.com).

Yang Lu is with the Department of Computer Science and Technology, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: luyang@xmu.edu.cn).

Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China (e-mail: ymc@comp.hkbu.edu.hk).

Nanrun Zhou is with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, China, and also with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China (e-mail: nrzhou@sues.edu.cn).

Digital Object Identifier 10.1109/TIFS.2025.3570121

I. Introduction

FACIAL normalization, a pivotal technique in computer vision, focuses on mitigating disruptive variations in facial images, including pose, expression, occlusion, and lighting, to generate a neutral-expression, frontal-view facial prototype that is unoccluded and uniformly illuminated [1]. Its utility spans diverse domains such as criminal investigations, identity verification, and human-computer interaction [2], [3], [4]. Despite its potential, current normalization approaches struggle to address the escalating complexity and diversity of real-world scenarios, emphasizing the pressing demand for more robust and adaptable solutions capable of accommodating dynamic environmental and conditional changes.

The limitations of existing face normalization methods are mainly manifested in two aspects: single contaminant element processing constraints, and single-domain applicability constraints. The former indicates that most current face normalization methods [5], [6] primarily focus on modeling and processing a single contaminant element. For example, they may be designed for only one specific category of variation factors such as pose, expression, illumination, or occlusion, and are unable to effectively handle other categories of variation factors, let alone cope with situations where multiple variation factors coexist. Additionally, some unsupervised normalization methods [7], [8] may inadvertently lose some crucial identity information of the input face during processing, resulting in an inconsistent identity between the generated facial prototype and the input face. The latter constraint points out that existing normalization models are usually only applicable to facial images within a single domain and lack universality across multi-domain scenarios [9], [10]. However, from a practical perspective, the images received by the system may come from various different domains, such as visible light images, near-infrared images, and even sketch images commonly used in criminal investigation. Therefore, there is an urgent need for a face normalization

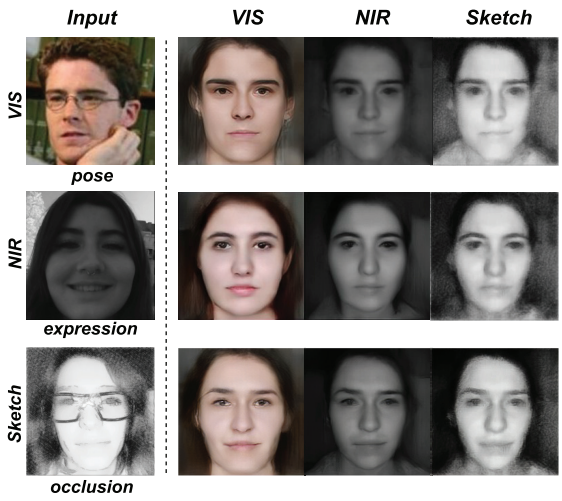


Fig. 1. Illustration of a desired face normalization method capable of handling multiple cross-domain face normalization tasks and addressing various types of facial variations.

method that can span multiple domains and simultaneously handle various types of facial variations, as illustrated in Fig. 1.

To tackle the aforementioned challenges, we propose an innovative Unified Multi-domain Face Normalization Network (UMFN) model. This model demonstrates proficiency in managing facial images originating from diverse domains and contaminated with various elements such as pose variations, expressions, occlusions, among others. It is capable of reconstructing a frontal, neutral facial prototype within the target domain, while concurrently extracting identity features that remain independent of domain-specific information. It is worth noting that, despite the remarkable advancements of diffusion models in image generation, their constraints in interpretable representation learning render them unsuitable for our requirements of joint prototype and feature learning [11]. Consequently, this paper embraces a fusion framework that integrates an Autoencoder with a Generative Adversarial Network (GAN), as illustrated in Fig. 2.

Specifically, the UMFN model incorporates a feature decoupling network alongside a domain classifier for adversarial training, effectively separating domain-agnostic features. By utilizing concatenated masks, it meticulously controls the generation of clean prototypes within designated domains or reconstructs the original images, thereby ensuring identity consistency throughout the process. Moreover, a multi-scale InstanceNorm Patch-based discriminator [12] is employed to refine image details and textures, significantly enhancing the visual quality of the outputs. As an unsupervised domain adaptation model, the UMFN facilitates the simultaneous training of diverse, unlabeled datasets within a unified network framework, showcasing its flexibility and practicality in cross-domain prototype reconstruction. Furthermore, we have devised an efficient heterogeneous face recognition (HFR) network that operates in conjunction with the UMFN's encoder. This network is tasked with extracting identity-discriminative features from the generated prototypes. These features are then fused with the domain-agnostic features of the input images for the purpose of HFR. During the training of the HFR network, a contrastive learning mechanism is leveraged to minimize the feature distance between the generated and authentic prototypes, thereby bolstering the network's proficiency in accurately recognizing identities.

The contributions of the paper are summarized below:

- We present a unified framework for face normalization, referred to as UMFN, which tackles the shortcomings of current approaches in managing diverse facial variations and cross-domain images. This framework efficiently processes multiple types of facial variations and is adaptable to face images from various domains.
- We develop a high-performance HFR network that combines domain-independent features from input samples with identity-specific features extracted from generated prototypes. Additionally, a contrastive learning strategy is incorporated to enhance the network's accuracy in identity recognition.
- As an unsupervised domain adaptation framework, the UMFN supports the concurrent training of multiple datasets from different domains without involving identity labels within a single network. It excels in reconstructing prototypes across multiple domains, improving the model's adaptability and practicality, while reducing the dependence on labeled data.
- Extensive experiments on diverse cross-domain face datasets with varying facial variations demonstrate the UMFN's superior performance in face normalization tasks, in both single-modal and cross-modal settings, as well as in HFR tasks.

The rest of this paper is organized as follows. Section II makes an overview of the related works on face normalization and heterogeneous face recognition. Section III details the proposed UMFN and HFR network. In Section IV, we conduct experiments on six real-world datasets to evaluate the performance of the proposed algorithms. Finally, we draw a conclusion in Section V.

II. Related Works

A. Face Normalization

Face normalization, an emerging and highly regarded research domain, has demonstrated significant practical importance in various applications, including criminal identification and facial evidence collection. The mainstream approaches in face normalization primarily focus on mitigating facial variations within single-modal contexts, striving to eliminate factors such as varying lighting conditions, exaggerated facial expressions, extreme head poses, or partial occlusions, in order to restore a standardized facial prototype. Existing single-modal face normalization algorithms can be broadly categorized into two categories [13]. The first category constructs clustering models utilizing auxiliary sample information to estimate the facial prototypes of contaminated samples. The second category harnesses generative AI models, such as GANs [14], [15] and Diffusion Models (DMs) [16], to learn the mapping

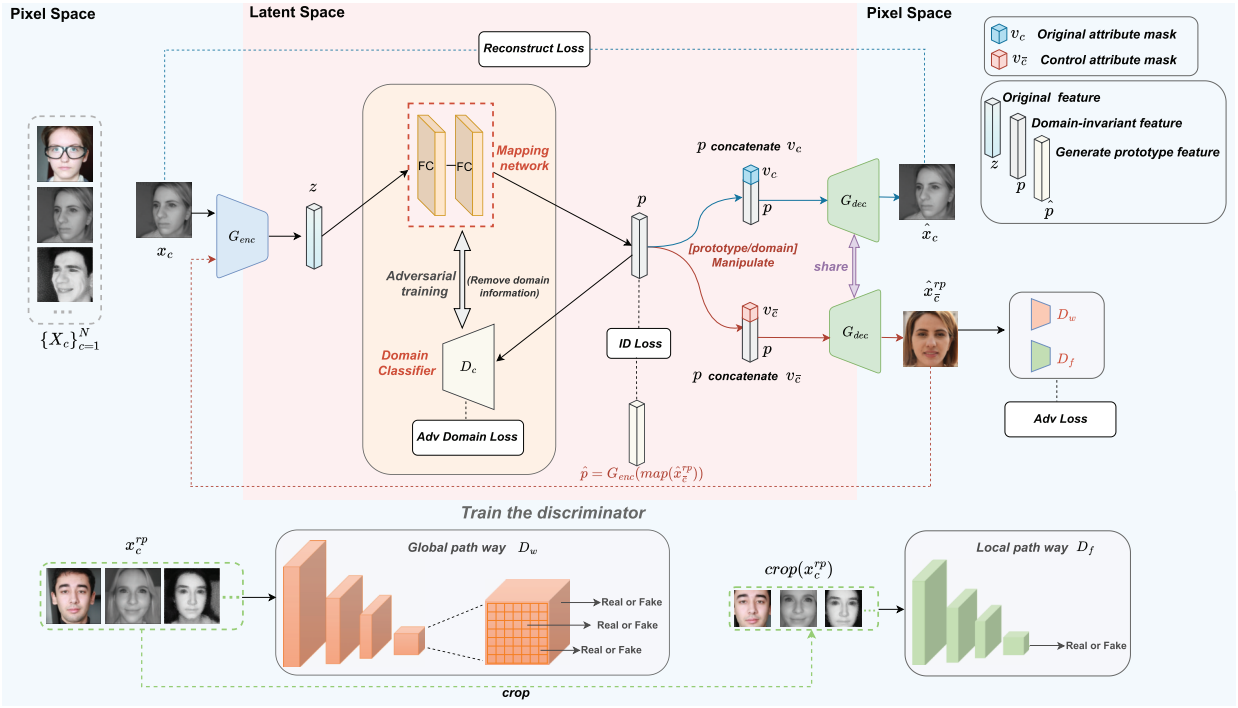


Fig. 2. Overview of the UMFN architecture. $x_c \in X_c$ represents an input image from domain c. The encoder G_{enc} and mapping network extract domain-agnostic features p from x_c. v_c and $v_{\bar{c}}$ are control masks used for reconstructing the original input ($\hat{x}_c$) and generating the target prototype ($\hat{x}_{\bar{c}}^{rp}$) in domain $\bar{c}$, respectively. ${x_c}^{rp}$ denotes the real prototype corresponding to x_c in its native domain c.

relationship between contaminated and standardized facial samples.

In the first category, Gao et al. [17] integrated unlabeled facial samples with a labeled training set, applying a Gaussian Mixture Model for clustering to estimate prototypes. Pang et al. [18] introduced a more robust semi-supervised low-rank representation model for clustering and prototype estimation, which is updated using labeled retrieval samples. While these methods are effective in prototype reconstruction, they require prior access to facial samples, making them unsuitable for real-time applications. In the second category, researchers have leveraged the image generation and mapping capabilities of generative AI models to propose a series of generative learning methods. These methods target specific facial variations, such as expression [5], [19], lighting [6], [7], pose [8], [20], [21], and partial occlusions [10], [22], [23], to generate facial prototypes with consistent identities. However, these methods are typically limited to specific types of facial variations and struggle to handle combinations of multiple variations. Pang et al. [9] attempted to address multiple types of contaminations by designing a Variation Disentangling GAN (VD-GAN). Nevertheless, this approach has limitations in preserving identity characteristics and is only effective for single-domain contamination.

Recently, several cross-domain face synthesis methods [24], [25], [26], [27], [28], [29], [30] have been introduced, including Cycle-consistency GAN (CycleGAN) [24], Parallel Multistage GAN (PMSGAN) [25], Brownian Bridge Diffusion Model (BBDM) [27], and NIR-FER Stochastic Differential Equations (NFER-SDE) [28]. Though these methods focus on transforming domain styles while preserving facial details, they are not designed to eliminate facial variations from input images. Given these limitations, this paper aims to develop a multi-domain face normalization framework capable of addressing universal facial variation removal and to explore the effective integration and optimization of domain transformation and prototype learning within a unified structure.

B. Heterogeneous Face Recognition

HFR pertains to the challenge of matching identities across disparate image domains, such as near-infrared (NIR) and visible light (VIS) imagery. The substantial distributional differences between these image domains significantly elevate the complexity of the matching and recognition tasks for the system. Existing HFR approaches can be broadly classified into three primary strategies [31]: modality-invariant feature representation, subspace learning-based techniques, and cross-modal synthesis methods.

Modality-invariant feature representation methods [32], [33], [34], [35] strive to extract common features from heterogeneous face data, aiming to minimize the discrepancies between different modalities by identifying cross-modal consistency during the feature extraction process. Liu et al. [33] utilized a multi-scale Difference of Gaussian (DoG) filter to capture illumination-invariant features from NIR and VIS face images. Gong et al. [32] proposed a universal encoding model that adopts a pixel-wise encoding approach to distill shared features from heterogeneous faces. Yi et al. [34] combined Gabor filters with Restricted Boltzmann Machines to extract common features from such faces, while also employing Principal Component Analysis (PCA) to reduce heterogeneity

and redundancy. Furthermore, [35] introduced a synchronized learning strategy for Local Binary Patterns (LBP) and encoding, effectively mitigating the impact of expression and illumination variations through binary encoding. However, these methods often necessitate substantial prior knowledge, and the extracted features can be redundant or lack compact, effective representations, thereby limiting their widespread application.

Subspace learning-based methods [36], [37], [38], [39], [40] endeavor to project face images from disparate modalities into a common subspace, where identity classification performance is optimized by maximizing inter-class distances and minimizing intra-class distances for more effective classification. He et al. [36] mapped NIR and VIS faces onto orthogonal subspaces, incorporating the Wasserstein distance to diminish modal discrepancies. Hu et al. [38] represented cross-modal faces as a blend of domain-specific and identity-related factors, then proposed a cross-domain factor separation module to erase modality differences in NIR-VIS pairs. Mudunuri et al. [39] introduced an orthogonal dictionary alignment method to align low-resolution NIR and VIS face images. Meanwhile, Cho et al. [40] presented a relational graph module that mitigates dependence on local texture details by integrating global information from cross-domain faces. Hu et al. [41] proposed the Dual Facial Alignment Learning (DFAL) method, addressing domain discrepancies in face recognition through feature- and image-level alignment, along with cross-domain compact representation. Yang et al. [42] introduced the Robust cross-domain Pseudo-labeling and Contrastive learning (RPC) network, enhancing NIR-VIS face recognition via pseudo-label sharing and intra- and cross-domain contrastive learning. Hu et al. [43] developed the Adversarial Disentanglement and Modality-Invariant Representation Learning (DMiR) method, eliminating spectral differences in cross-modal features using domain-relevant disentanglement, modality-invariant representation, and orthogonal decorrelation.

Cross-modal synthesis methods [1], [44], [45], [46], [47] aim to eliminate modality discrepancies by synthesizing heterogeneous faces into a unified modality at the pixel level. He et al. [44] proposed a cross-spectral face synthesis network, which includes two modules: texture restoration, which generates realistic textures, and pose correction, which produces front-facing pose images. To address the pose mismatch issue in NIR-VIS datasets, Yu et al. [45] introduced a pose-preserving cross-spectral face refinement method to generate NIR-VIS paired images that maintain pose consistency at the pixel level. Pang et al. [1] proposed a unified framework that generates cross-domain prototypes through bidirectional prototype learning, eliminating domain information and noise from input faces at the pixel level. Fu et al. [47] introduced the Dual Variational Generation (DVG-Face) framework, which generates large-scale heterogeneous data by sampling from noise. Our proposed UMFN method can be regarded as a combination of cross-modal synthesis methods and subspace learning-based methods, distinguished by its ingenious utilization of synthesized standard prototypes to retroactively train the HFR network. It employs a contrastive learning mechanism to learn highly identity-discriminative features and integrates them with domain-agnostic features extracted from the original input by the UMFN, facilitating efficient and accurate HFR.

TABLE I
MEANING OF THE NOTATIONS IN UMFN

Notation	Meaning
G_{enc}	Encoder
G_{dec}	Decoder
map	Mapping network
D_c	Domain feature classifier
D_w	Global path way discriminator
D_f	Local path way discriminator
x_c	Image in domain c
$\hat{x}_c$	Reconstructed image of x_c
x_c^{rp}	Real prototype of x_c in domain c
$\hat{x}_{\bar{c}}^{rp}$	Generated prototype of x_c in domain $\bar{c}$
z	Feature encoded by G_{enc}
p	Domain-agnostic feature extracted by map
$\hat{p}$	Generated prototype's latent feature
v_c	Control mask for reconstructing original input
$v_{\bar{c}}$	Control mask for generating target prototype

III. The Proposed Method

This paper introduces a generalized framework for facial normalization, named UMFN, with its structural overview depicted in Fig. 2. The UMFN framework is built to process facial images containing diverse variations (e.g., lighting, pose, occlusion) across multiple image domains, generating standardized facial prototypes aligned with a specified target domain. The system integrates two primary modules: (1) resolving cross-domain discrepancies by converting input images from heterogeneous sources into the unified target domain, and (2) simultaneously addressing multiple facial variations while ensuring identity-preserving prototype generation. These modules are thoroughly discussed in Sections III-A and III-B, respectively. Section III-C further elaborates on the HFR network, which is optimized through integration with UMFN to enable seamless cross-domain recognition. Section III-D outlines the overall algorithm logic and related relationships. Key terminologies and symbols used in this work are consolidated in Table I for clarity.

A. Feature Decoupling and Cross-Domain Generation

To address the issue of input face images potentially originating from diverse domains, we have developed a feature decoupling mapping network. This network is capable of isolating domain-specific information from the latent space, thereby producing domain-invariant representations. Such decoupled, domain-agnostic identity features can subsequently be utilized for HFR tasks.

Feature Decoupling: We design a feature decoupling mapping network that first extracts the primary feature $z = G_{enc}(x_c)$ from the input image x_c via a pre-trained encoder G_{enc}, and subsequently generates its latent feature representation $p = \text{map}(z)$ using the mapping network map. Specifically, the domain feature classifier D_c processes the latent representations p as input and produces the domain prediction code (i.e., the image domain to which p belongs). The objective of D_c

is to precisely identify the domain of p, while the mapping network map is designed to disrupt D_c's accurate domain prediction. Through adversarial training, the dynamic interaction between the two facilitates the decoupling of domain information. Consequently, the objective function for training D_c is defined as follows:

$$L_{D_c} = E\left[\log P(l_p^{pre} = l_{x_c}^{real} \mid p)\right], \tag{1}$$

where l_p^{pre} is the domain label predicted by D_c, $l_{x_c}^{real}$ is the real domain label of x_c. On the other side, the objective function for training the mapping network *map* is as

$$L_{map} = -E\left[\log P(l_p^{pre} = l_{x_c}^{real} \mid p)\right]. \tag{2}$$

Cross-Domain Generation: We propose integrating a control mask with reconstruction learning to enable cross-domain image generation. The primary goal of reconstruction learning is to train the decoder G_{dec} to decode domain-agnostic latent representations $p = \text{map}(G_{enc}(x_c))$ conditioned on the original attribute mask v_c, thereby reconstructing the input image x_c. This framework ensures the decoder not only learns to interpret and apply mask semantics but also leverages identity-related features embedded in the latent representation p. Consequently, by modifying the mask, images from target domains can be generated in subsequent stages. The reconstruction learning objective function is formulated as

$$L_{rec} = E_{x_c \sim p_{data}}\left[\|x_c - \hat{x}_c\|_1\right], \tag{3}$$

where $\hat{x}_c = G_{dec}(p, v_c)$ is the reconstructed image of x_c, v_c is the original attribute mask of the input image x_c, indicating whether x_c is contaminated and its domain information. Specifically, v_c is composed of a four-digit one-hot vector, where the first digit indicates whether the image is contaminated, and the last three digits represent the domain it belongs to. In Eq. (3), l_1 norm is employed instead of the l_2 norm to suppress the issue of blurred generated images.

B. Prototype Generation

Unlike traditional single-contamination normalization approaches, which heavily emphasize categorizing facial variations and meticulously modeling them, our method revisits the core objective of prototype reconstruction. It focuses on directly generating standardized facial prototypes that retain identity information by leveraging identity cues from contaminated facial samples. This strategy effectively circumvents the complexities associated with explicitly modeling facial variations in an innovative way.

Notably, we introduce a prototype adversarial loss, distinct from traditional GAN-based image translation approaches that seek to model the distribution of original face data with diverse variations. Our loss focuses on approximating a condensed distribution of standardized facial prototypes while simultaneously mitigating overfitting to these variations. Additionally, we develop a multi-scale Markov discriminator $D = [D_w, D_f]$ to enhance detail refinement in synthesized faces and stabilize training dynamics. Specifically, the domain-agnostic identity feature $p = \text{map}(G_{enc}(x_c))$ is concatenated with a control mask $v_{\bar{c}}$, where the first bit in $v_{\bar{c}}$ determines whether face normalization is applied (1 for activation), while the remaining three bits specify the target domain for prototype generation. This combined feature is fed into the decoder to synthesize a pixel-level standardized facial prototype within the designated target domain. The adversarial objectives for training the generator and the discriminator are defined below:

$$\begin{aligned} L_D^{adv} = {} & D_w\left(\hat{x}_{\bar{c}}^{rp}\right) + D_f\left(\text{crop}\left(\hat{x}_{\bar{c}}^{rp}\right)\right) \\ & - D_w\left(x_c^{rp}\right) - D_f\left(\text{crop}\left(x_c^{rp}\right)\right), \end{aligned} \tag{4}$$

$$L_{G_{dec}}^{adv} = -\left(D_w\left(\hat{x}_{\bar{c}}^{rp}\right) + D_f\left(\text{crop}\left(\hat{x}_{\bar{c}}^{rp}\right)\right)\right), \tag{5}$$

where D_w and D_f represent the global path discriminator and local path discriminator, respectively. $\hat{x}_{\bar{c}}^{rp} = G_{dec}(\text{map}(G_{enc}(x_c), v_{\bar{c}})$ denotes the specified domain prototype generated by inputting x_c and the control mask $v_{\bar{c}}$, while x_c^{rp} represents the uncontaminated faces in the training set. The function "crop" is used to cut out the facial features from the input face images. Considering that the Wasserstein distance is more stable and can avoid training instability and mode collapse in GANs, we adopt the WGAN-GP [48] loss in our model to replace the traditional cross-entropy loss.

Furthermore, to ensure that the generated prototype $\hat{x}_{\bar{c}}^{rp}$ can effectively preserve the identity information from the input x_c, we have designed an identity feature similarity loss function. This is achieved by feeding the generated prototype back into the encoder and mapping network to obtain the latent feature $\text{map}(G_{enc}(\hat{x}_{\bar{c}}^{rp}))$, and then calculating its similarity to the latent feature p of the input x_c, thereby retaining identity. The loss function is as follows:

$$L_{id} = \left\| p - \text{map}\left(G_{enc}\left(\hat{x}_{\bar{c}}^{rp}\right)\right)\right\|_2^2, \tag{6}$$

where $\|\cdot\|_2^2$ denotes l_2-norm. On the other hand, UMFN transfers the identity feature retention ability from the reconstruction learning task to the face prototype generation task. Since these two tasks share the same input and output domains, the transferability gap [49] between them is minimal. Therefore, the identity feature retention ability learned from the reconstruction learning task can be easily transferred to the prototype reconstruction task. Therefore, the overall loss function of the decoder G_{dec} is

$$L_{G_{dec}}^{adv} = \lambda_1 L_{rec} + \lambda_2 L_{id}, \tag{7}$$

where λ_1 and λ_2 are weighting hyperparameters of L_{rec} and L_{id}, respectively.

C. Heterogeneous Face Recognition

At this stage, we meticulously construct an HFR network, whose core components encompass an encoder F_θ (its network architecture aligns with G_{enc} in UMFN) and a fully connected layer l_θ. The detailed structure is shown in Fig. 3. Specifically, the HFR network consists of two core operations:

Feature Fusion: During the training phase, we fuse domain-agnostic features (derived from the source domain input image via UMFN's encoder G_{enc} and mapping network *map*) with identity-discriminative features (extracted from real target domain prototypes using encoder F_θ). The fused features are processed by a softmax layer for classification. To enhance robustness, the fully connected layer l_θ expands the feature

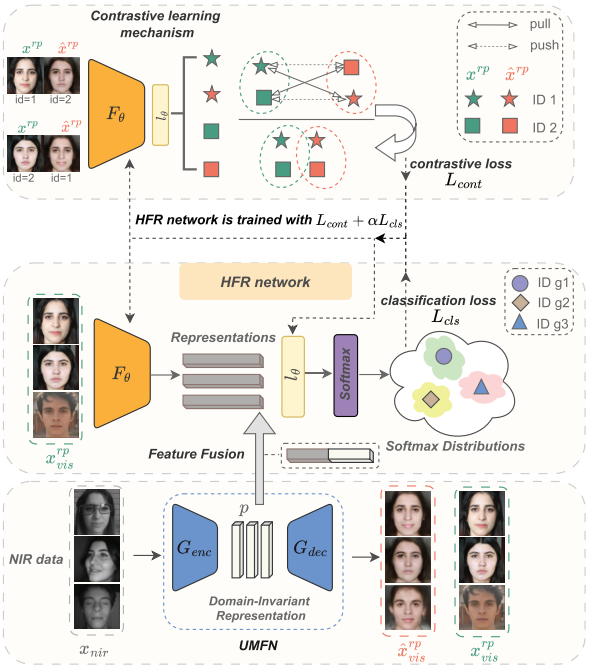


Fig. 3. Overview of the HFR network and training process. The HFR network is trained using the loss function in Eq. (10), which is composed of L_{cls} (classification loss) and αL_{cont} (contrastive loss), where α is a trade-off hyperparameter balancing these two losses.

dimensions, enabling richer integration of multi-dimensional identity information. The classification loss is formulated as

$$\mathcal{L}_{\text{cls}} = \text{softmax}\left(l_\theta\left[F_\theta(x_{v,i}^{rp}) \oplus \text{map}(G_{enc}(x_{n,i}))\right], i\right), \quad (8)$$

where $x_{n,i}$ denotes the source domain image, $x_{v,i}^{rp}$ represents the real target domain prototype, $F_\theta(x_{v,i}^{rp})$ is the identity-discriminative feature of $x_{v,i}^{rp}$, and $\text{map}(G_{enc}(x_{n,i}))$ corresponds to the domain-agnostic feature of $x_{n,i}$. The operator $\oplus$ signifies feature concatenation. During inference, identity-discriminative features from generated target domain prototypes are extracted via the trained F_θ and combined with domain-agnostic features from test-phase source domain images to enable heterogeneous face recognition.

Contrastive Learning Mechanism: During the cross-domain generation process, the generated target domain prototypes may experience identity feature loss or the introduction of noise. Relying directly on such data for identity classification training can result in overfitting to absent or noisy features, thereby compromising classification performance. To address this issue, we innovatively incorporate a contrastive learning mechanism that aligns the generated prototypes with real ones within the feature space, while simultaneously mitigating the impact of noise and missing features. This approach ensures that the encoder F_θ extracts features that closely resemble those of real prototypes.

Specifically, we randomly sample two pairs of real heterogeneous images, $(x_{v,1}^{rp}, x_{n,1})$ and $(x_{v,2}^{rp}, x_{n,2})$, where $x_{n,1}$ and $x_{n,2}$ are NIR images with contamination. We then use UMFN to perform cross-domain restoration on $x_{n,1}$ and $x_{n,2}$, generating the corresponding target VIS domain prototypes $\hat{x}_{v,1}^{rp}$ and $\hat{x}_{v,2}^{rp}$. Next, we form positive pairs $(x_{v,1}^{rp}, \hat{x}_{v,1}^{rp})$ and $(x_{v,2}^{rp}, \hat{x}_{v,2}^{rp})$, and negative pairs $(x_{v,1}^{rp}, \hat{x}_{v,2}^{rp})$ and $(x_{v,2}^{rp}, \hat{x}_{v,1}^{rp})$. The corresponding contrastive loss is

$$L_{cont} = \sum_{j \neq k}(1 - \langle f_j, \tilde{f}_j\rangle) + (1 - \langle f_k, \tilde{f}_k\rangle) + \max(0, \langle f_j, \tilde{f}_k\rangle - m) + \max(0, \langle f_k, \tilde{f}_j\rangle - m), \quad (9)$$

where $f_j = l_\theta[F_\theta(x_{v,j}^{rp}) \oplus \text{map}(G_{enc}(x_{n,j}))]$, $\langle , \rangle$ represents the calculation of the cosine similarity between two features, and m is a margin value set at 0.5. In summary, the ultimate loss function of the HFR network is expressed as

$$L_{hfr} = L_{cls} + \alpha L_{cont}, \quad (10)$$

where α is a trade-off parameter.

D. Summary of the Proposed Algorithm

Finally, we present a summary of the proposed UMFN and the HFR network. The UMFN framework encompasses two primary components: (1) feature decoupling and cross-domain transformation, and (2) prototype generation. It commences by transforming input images into latent representations and decouples domain-specific information to yield domain-agnostic features. Subsequently, these features are harnessed to generate standardized facial prototypes employing adversarial loss and identity feature similarity loss, while a multi-scale Markov discriminator further refines facial details. Regarding the HFR network, it integrates the domain-agnostic features with identity-discriminative features extracted from the prototypes, incorporating contrastive learning to align the generated prototypes with real ones, enabling robust cross-domain face recognition. Together, these components constitute a cohesive framework that provides a comprehensive solution for face normalization and cross-domain face recognition.

IV. Experiments

A. Datasets Descriptions

LAMP-HQ [50] is a newly created, challenging large-scale NIR-VIS dataset comprising over 56,000 NIR and 16,000 VIS face images from 573 individuals, featuring diverse expressions, poses, lighting conditions, occlusions, and complex backgrounds.In the heterogeneous face recognition experiments, we adopted a 10-fold cross-validation setup, ensuring that the training and testing sets in each trial were completely identity-disjoint.

CASIA NIR-VIS 2.0 [51] consists of 725 identities, with each identity containing 1 to 22 VIS images and 5 to 50 NIR images. The experimental setup involves 10 independent trials designed for heterogeneous face recognition tasks. The training set comprises approximately 6,100 NIR images and 2,500 VIS images, covering a total of 360 identities. In each experimental test phase, the gallery set includes 358 VIS images, each representing one of 358 identities, while the probe set contains over 6,000 NIR images corresponding to the same 358 identities. The identities in the training and testing sets are completely disjoint, ensuring the independence and fairness of the experimental results.

BUAA NIR-VIS [52] includes 150 identities, each with 9 NIR and 9 VIS images, covering various head poses and expressions.

CUFSF [53] is a widely used sketch dataset commonly applied in sketch synthesis and recognition tasks. It comprises 1,194 identities sourced from the FERET dataset [54], with each identity represented by a standard facial photograph and an artist-drawn sketch.

IJB-A [55] is one of the most challenging unconstrained face recognition benchmark datasets, featuring uncontrolled pose variations. It includes images and video frames from 500 subjects, comprising 5,397 images and 2,042 videos, which are split into 20,412 frames. On average, each subject has 11.4 images and 4.2 videos, captured in real-world settings to avoid frontal bias.

LFW [56] comprises more than 13,000 images of 5,749 identities, captured in unconstrained environments with significant variations in expressions, poses, lighting conditions, and other factors.

B. Experimental Setup

In the experimental section, we primilary conduct five experiments: (1) Homogeneous Multi-domain Face Normalization Experiment, (2) Heterogeneous Multi-domain Face Normalization Experiment, (3) Domain Disentanglement and Identity Preservation Visualization Experiment, (4) Heterogeneous Face Recognition Experiment, and (5) Ablation Study. The configurations of these experiments are as follows:

1) Homogeneous Multi-domain Face Normalization Experiment: In this experiment, we employ three types of domain data from the LAMP-HQ dataset to train our model, ensuring all data originates from the same set of identities (IDs). The primary objective is to assess the model's facial normalization performance across multi-domain data within a single dataset. Specifically, we utilize frontal pose (FP) data from the LAMP-HQ dataset, where all NIR and VIS images from 100 IDs are designated as the training set. Given that the LAMP-HQ dataset only includes NIR and VIS domain data, we use software[1] to convert NIR photographs into sketch images, thereby simulating a scenario encompassing three types of domain data. Firstly, we present and analyze the in-domain face normalization results of the UMFN model on the LAMP-HQ dataset, as well as its performance when transferred to the LFW and IJB-A datasets. Secondly, we showcase and analyze the cross-domain face normalization results of the UMFN model on the LAMP-HQ dataset. Lastly, we quantitatively assess the face normalization performance of the UMFN model.

2) Heterogeneous Multi-domain Face Normalization Experiment: To showcase the model's ability to operate effectively without requiring paired training data and to validate its generalization across diverse datasets, we leverage domain-specific data from three independent sources. For training, VIS domain data is sourced from 250 identities in the LAMP-HQ dataset, NIR domain data comprises 130 identities from the BUAA dataset, and Sketch domain data is extracted from 400 identities in the CUFSF dataset.

3) Domain Disentanglement and Identity Preservation Visualization Experiment: We conduct a visualization experiment, to validate the effectiveness of the proposed UMFN in decoupling domain information and preserving identity in multimodal scenarios.

4) Heterogeneous Face Recognition Experiment: We conduct heterogeneous face recognition experiments on the LAMP-HQ and CASIA NIR-VIS 2.0 datasets, following the standard evaluation protocol in [57].

5) Ablation Study Experiment: We investigate the contribution of each component within the UMFN framework and explore the roles of the contrastive learning mechanism and feature fusion operation in the HFR network.

C. Implementation Details and Parameter Setting

In the UMFN, we employ a Resnet-50 pre-trained on the MS1M [58] dataset and fine-tuned on VGGFace2 [59] as the encoder G_{enc} to extract 2048-dimensional identity features, keeping G_{enc} fixed during training and testing. The mapping network *map* consists of three linear layers, inputting a 2048-dimensional vector and outputting a 2048-dimensional vector. The feature domain classifier D_c is composed of linear layers and outputs the predicted domain category, while the decoder G_{dec} outputs images of size 224×224. The discriminators D_w and D_f are both patchGAN discriminators with a receptive field of 70, using instance normalization as the normalization layer, and the crop size is 150×150. The network architectures of the decoder G_{dec}, the discriminators D_w and D_f are presented in Table II. The total number of trainable parameters (Params) in the UMFN model is approximately 36.8M. The model training was conducted on a machine equipped with an Intel Xeon Gold 6430 processor, 32GB of memory, an NVIDIA GeForce RTX 4090 GPU, running Ubuntu 22.04.5 LTS, and using PyTorch version 2.1.2.

In the UMFN prototype generation phase, two hyperparameters, λ_1 and λ_2, are introduced to balance the reconstruction loss L_{rec} and the identity feature similarity loss L_{id} in Eq. (7). Specifically, λ_2 is set to 10 as recommended in [60], while λ_1 is dynamically adjusted by $10 - \frac{\text{cur}_{epoch}}{\text{sum}_{epoch}} \times 9$, where cur_{epoch} and sum_{epoch} denote the current and the total training epochs, respectively. This dynamic adjustment incorporates a penalty term tied to the training progress. The design rationale stems from experimental insights: reconstruction learning effectively leverages masks to guide domain-specific image generation in early training stages, but its impact on identity preservation diminishes as training advances. Thus, λ_1 is progressively reduced to address this trade-off. For HFR network training, the hyperparameter α in Eq. (10) is empirically set to 0.001 through grid search. Optimization employs the Adam optimizer with a learning rate of 0.0002 and momentum of 0.5.

It is worth noting that the implementation strategy used in our experiments represents one of many viable configurations tailored to our specific task. We encourage researchers to adapt the network architecture to suit their specific requirements.

[1]The link for the transformation software is at https://fotosketcher.com

TABLE II
THE NETWORK ARCHITECTURES OF DECODER G_{dec}, DISCRIMINATORS D_w AND D_f

Decoder (G_{dec})	**Discriminator** (D_w)	**Discriminator** (D_f)
DeConv$(1024, 6, 2, 0)$, BN, Leaky ReLu	Conv$(64, 4, 2, 1)$, Leaky ReLu	Conv$(64, 4, 2, 1)$, Leaky ReLu
DeConv$(512, 6, 2, 1)$, BN, Leaky ReLu	Conv$(128, 4, 2, 1)$, InstanceNorm, Leaky ReLu	Conv$(128, 4, 2, 1)$, InstanceNorm, Leaky ReLu
DeConv$(256, 6, 2, 2)$, BN, Leaky ReLu	Conv$(256, 4, 2, 1)$, InstanceNorm, Leaky ReLu	Conv$(256, 4, 2, 1)$, InstanceNorm, Leaky ReLu
DeConv$(128, 6, 2, 2)$, BN, Leaky ReLu	Conv$(512, 4, 1, 1)$, InstanceNorm, Leaky ReLu	Conv$(512, 4, 2, 1)$, InstanceNorm, Leaky ReLu
DeConv$(64, 6, 2, 2)$, BN, Leaky ReLu	Conv$(1, 4, 1, 1)$	Conv$(512, 4, 2, 1)$, InstanceNorm, Leaky ReLu
DeConv$(3, 6, 2, 2)$, Tanh		FC(1)

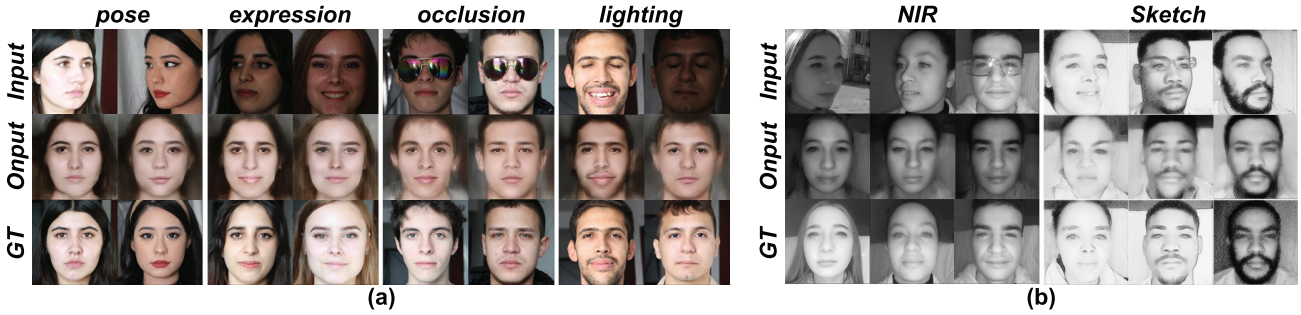


Fig. 4. In-domain face normalization results on the LAMP-HQ dataset. Fig. 4(a) illustrates the in-domain normalization results for input images with various types of contaminations from the visible light domain of the LAMP-HQ dataset. Fig. 4(b) presents the in-domain normalization results for input images with various types of contaminations from the near-infrared and sketch domains of the LAMP-HQ dataset.

D. Homogeneous Multi-Domain Face Normalization

In this subsection, we evaluate the performance of UMFN in homogeneous multi-domain face normalization using the LAMP-HQ dataset.

In-Domain Face Normalization: The in-domain normalization results of the UMFN are presented in Fig. 4. Fig. 4(a) showcases the results of inputting VIS domain faces corrupted by different facial variations such as pose, lighting, occlusion, and expression, and outputting normalized VIS domain faces. Fig. 4(b) displays the outcomes of inputting corrupted faces from the NIR and Sketch domains and outputting normalized faces in their respective domains. The bottom row serves as a reference for the true prototypes. As evident from Fig. 4, the UMFN demonstrates the ability to concurrently normalize diverse types of corrupted faces from multiple distinct domains to their respective normalized states, effectively retaining the identity of the input faces while acquiring facial prototypes.

Subsequently, Fig. 5 presents the in-domain face normalization performance of UMFN after training on the LAMP-HQ dataset and evaluating on the LFW and IJB-A datasets. The results indicate that, even when the input faces are subjected to various types of contaminations, such as occlusions, pose variations, expression distortions, or combinations of these corruptions, the UMFN can efficiently mitigate these facial variations and reconstruct standardized frontal faces, while accurately preserving the crucial identity features of the inputs. *The experiments suggest that our model possesses excellent cross-dataset transferability and versatility in handling diverse facial variations.* Furthermore, we undertake a comparative analysis between our UMFN and the other three recent single-domain face normalization techniques, namely TP-GAN [61], FNM [60] and DisPV [13], using contaminated face images

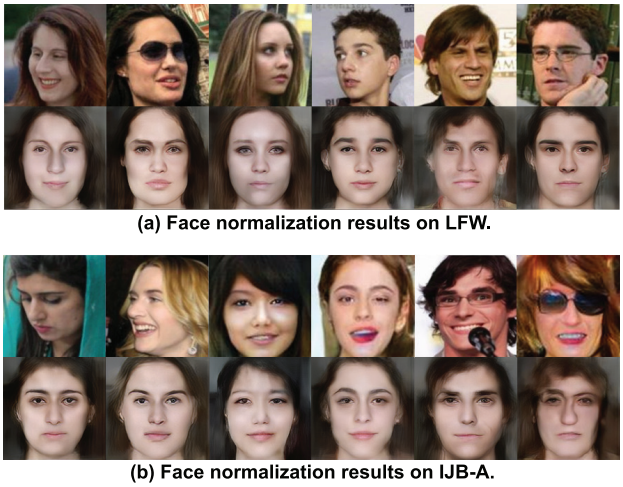


Fig. 5. Face normalization results on LFW and IJB-A datasets. The first row presents the input samples with various facial variations, while the second row shows the corresponding prototypes reconstructed by UMFN.

from the LFW and IJB-A datasets as inputs. The face normalization results for each method are illustrated in Fig. 6. It is evident that the target prototype generated through UMFN normalization demonstrates superior performance in both image quality and identity preservation.

Cross-Domain Face Normalization: Fig. 7 depicts the cross-domain normalization outcomes achieved by the UMFN model on LAMP-HQ dataset. Specifically, the inputs encompass facial images corrupted by various factors from domain A, whereas the outputs represent the facial prototypes generated in domains B and C. The bottom row showcases the reference prototypes of the inputs within the VIS domain. It can

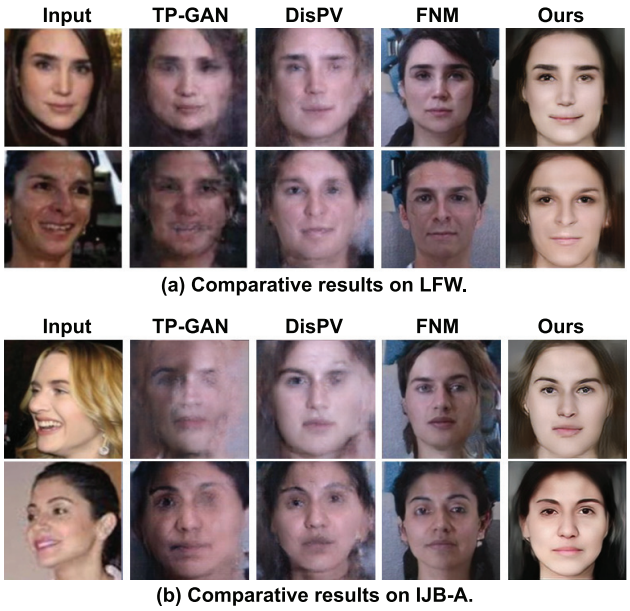


Fig. 6. Comparison with other single-domain face normalization methods on (a) LFW and (b) IJB-A datasets.

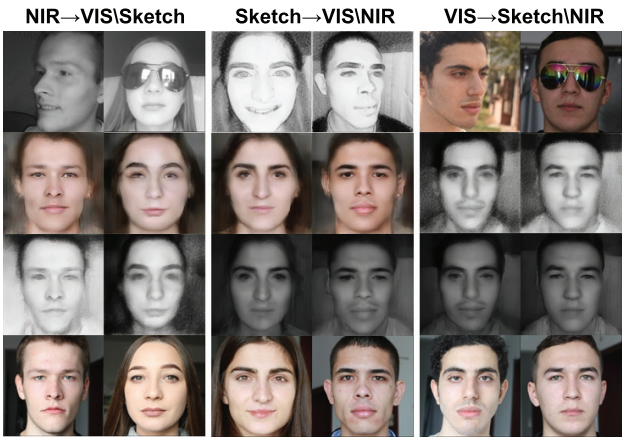


Fig. 7. Cross-domain face normalization results on LAMP-HQ dataset. The second and third rows depict face normalization results in different domains from the inputs shown in the first row. The bottom row shows the real VIS prototypes for reference.

be observed that the UMFN model successfully addresses the challenges of domain transfer and prototype learning, effectively preserving the identity characteristics of the inputs when generating uncontaminated facial prototypes for the target domains. The inspiring results demonstrate that the UMFN is capable of performing effective domain-specific face normalization on facial images sourced from different domains and contaminated by various types of facial variations, all while efficiently retaining the key identity information of the inputs, utilizing a single network architecture.

Quantitative Evaluation: Adhering to the protocols outlined in [23] and [63], we conduct a further quantitative evaluation of the UMFN on LFW and IJB-A datasets. By feeding restored prototypes into the ResNet-18 [62] network for face verification, our objective is to ascertain whether prototypes restored by UMFN enhance verification rate (VR) performance. Additionally, we report the Symmetry Index (SI) [65] to facilitate a comprehensive comparison of the quality of prototypes generated by UMFN, R&R [63], CFR [23], DisPV [13], and the state-of-the-art Mask Rotate [64].

TABLE III
VERIFICATION PERFORMANCE AND SYMMETRY INDEX (SI) ON LFW AND IJB-A DATASETS

Method	Datasets			
	LFW		IJB-A	
	VR↑	SI↓	VR↑	SI↓
ResNet [62]	98.85	-	90.57	-
ResNet+R&R [63]	98.95	20.93	91.98	20.85
ResNet+CFR [23]	99.23	28.31	93.36	24.76
ResNet+DisPV [13]	99.05	31.27	92.48	34.63
ResNet+Mask Rotate [64]	**99.80**	19.27	95.50	20.52
ResNet+UMFN (ours)	99.57	**14.61**	**95.92**	**15.81**

Table III presents the verification performance and SI metrics for UMFN across both datasets. The results reveal that UMFN normalization significantly elevates the verification performance of the ResNet model: on IJB-A, when paired with ResNet-18, it achieves a 5.35% improvement in VR, with accuracy gains also evident on LFW. These experiments corroborate two key findings: (1) the removal of pose, expression, occlusion, and illumination variations from input faces enables the model to concentrate on essential identity features more effectively, and (2) UMFN adeptly retains crucial identity information while mitigating these facial variations. Furthermore, UMFN outperforms the other comparing face normalization methods in terms of SI, generating highly symmetrical faces that rectify facial distortions.

E. Heterogeneous Multi-Domain Face Normalization

In this subsection, we evaluate the performance of UMFN in the heterogeneous multi-domain face normalization experiment. It is particularly noteworthy that there is currently no other face normalization method capable of adequately addressing this complex issue. Fig. 8 showcases the normalization results of images from the LAMP-HQ, BUAA, and CUFSF datasets across different domains (VIS, NIR, Sketch). We observe that regardless of the dataset (or domain) the input comes from, UMFN successfully generates prototypes that are highly matched to the target domain styles provided by another dataset. It effectively eliminates interfering factors such as pose, expression, and occlusion in the input face images while accurately preserving the input identity information. It is worth noting that the changes after normalization are primarily manifested in the mouth area (as indicated by the highlighted boxes in Fig. 8(c)), due to the minimal facial variations in most images of the CUFSF dataset, which generally exhibit simple facial expressions such as smiles. Furthermore, the challenging nature of cross-domain normalization tasks involving sketch images, which contain the least amount of facial detail information, leads to slight local distortions in the generated VIS and NIR prototypes. Despite these challenges, the results demonstrate the UMFN's robust multi-domain and multi-type face normalization capabilities, particularly when

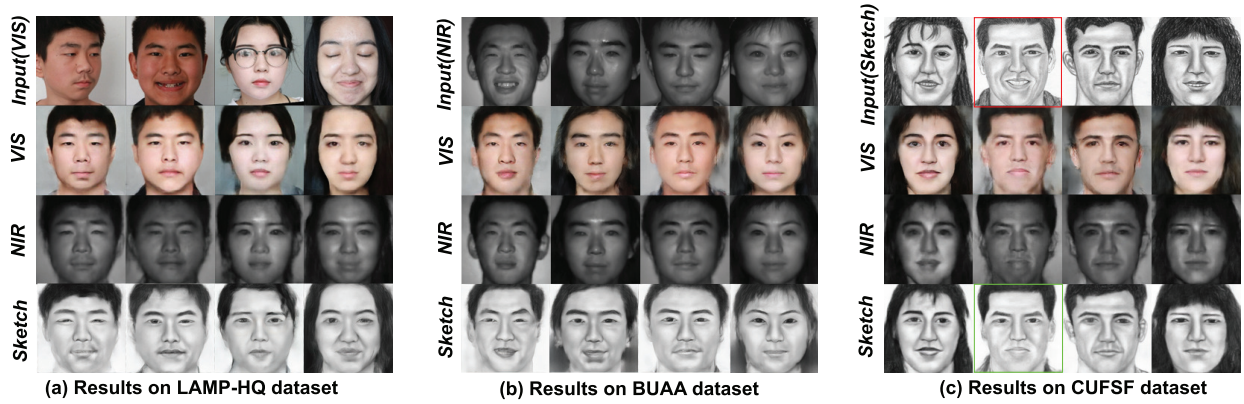


Fig. 8. The normalization results of face images from the LAMP-HQ, BUAA, and CUFSF datasets across different domains.

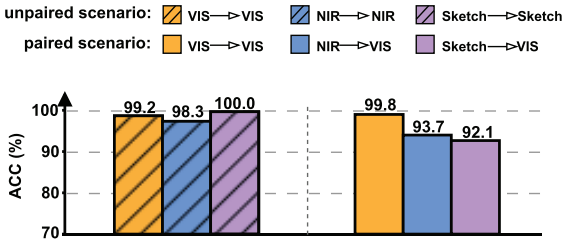


Fig. 9. The face recognition results of UMFN for multi-domain prototype restoration on paired and unpaired scenarios.

trained on unpaired datasets from multiple domains, where its performance remains impressive.

Quantitative Evaluation: Since input faces only have real prototypes within their respective domains and lack corresponding cross-domain real prototypes, we assess the recognition accuracy of the reconstructed prototypes within their own domains. As shown in Fig. 9, in this unpaired scenario, intra-domain prototype restoration across multiple domains achieves a high recognition accuracy exceeding 98%, confirming that our UMFN can perform face normalization with high identity preservation even without paired data for training.

Furthermore, we report the recognition results of our model after cross-domain prototype reconstruction in paired scenarios, where the probe samples with contamination originate from the VIS, NIR, and Sketch domains of LAMP-HQ, while the gallery real prototypes come from the LAMP-HQ VIS domain. We observe that, compared to generating VIS prototypes within the same domain, the cross-domain generation process inevitably leads to slight feature loss or noise introduction, resulting in a decrease in recognition accuracy. This is the key reason why we introduce the contrastive learning mechanism to train the HFR network, aiming to align the generated prototypes with the real ones in the feature space.

F. Domain Disentanglement and Identity Preservation Visualization

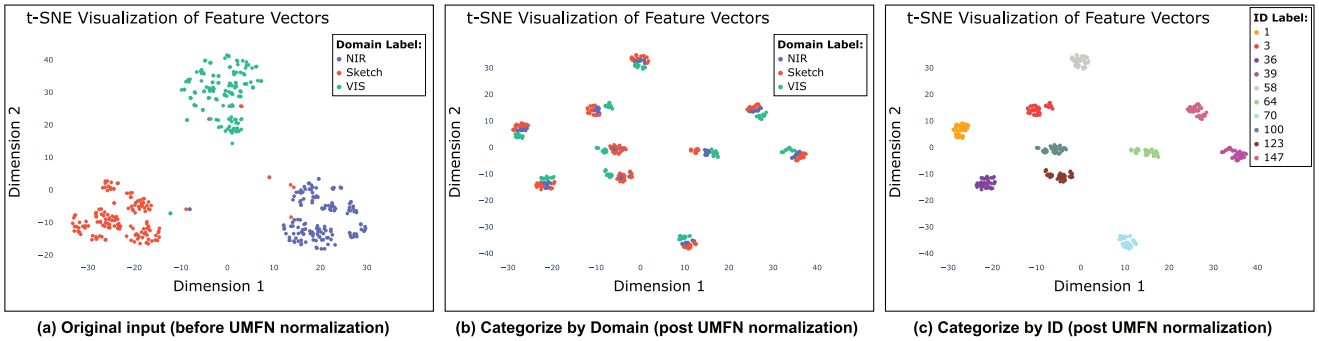
To further validate the effectiveness of our UMFN in decoupling domain-specific information and preserving identity in multimodal scenarios, we conduct a visualization experiment. Specifically, we employ t-SNE to visualize the identity features extracted by the UMFN for VIS, NIR, and sketch samples of 10 randomly selected individuals on LAMP-HQ dataset, as illustrated in Fig. 10. To differentiate between identities and sample domains, we assigned distinct colors to each.

Our analysis reveals a fundamental distinction between the input distribution (before UMFN normalization) and the processed features. In the original feature space, samples from the same domain cluster together irrespective of identity differences, indicating that domain-specific variations dominate over identity-related differences. In contrast, the UMFN's innovative architecture successfully disentangles domain information from identity features during the feature extraction process. This is evidenced by two key observations: (1) distinct separation of identity features across different IDs (see Fig. 10(c)), confirming precise identity information preservation, and (2) effective clustering of same-identity samples from different domains (see Fig. 10(b)), demonstrating robust domain-invariant feature representation. These results collectively validate the UMFN's ability to generate robust, domain-agnostic representations while preserving discriminative identity information.

G. Heterogeneous Face Recognition

In this experiment, the UMFN is utilized to reconstruct cross-domain prototypes from NIR data in the probe set, producing VIS prototypes. Subsequently, we utilize the trained HFR network to obtain fused features (i.e., concatenation of domain-agnostic and identity-discriminative features) for both these generated prototypes and the authentic prototypes in the gallery set. Ultimately, by measuring the cosine similarity between the combined features of the reconstructed prototypes and those of the gallery set, a similarity matrix is derived, which is then used to determine the rank-1 recognition accuracy and VR at FAR =0.001.

Table IV and Table V present a comparison of our UMFN method with other advanced HFR approaches, including ADFL [66], WCNN [36], PCFH [45], PACH [67], ADCANs [68], MMTN [69], NLPF [57], LPL [31], and MDFD-HFA [2], and HERE [70] on CASIA NIR-VIS 2.0 and



(a) Original input (before UMFN normalization) (b) Categorize by Domain (post UMFN normalization) (c) Categorize by ID (post UMFN normalization)

Fig. 10. t-SNE visualization of feature distributions for VIS, NIR, and sketch samples from 10 randomly selected individuals.

TABLE IV

RANK-1 RECOGNITION ACCURACY AND VERIFICATION RATE ON CASIA NIR-VIS 2.0 DATASET

Type	Method	Rank-1 (%)	VR@FAR = 0.1%(%)
Supervised	ADFL (2018)	98.1 ± 0.4	97.1 ± 0.5
	WCNN (2018)	98.7 ± 0.3	98.4 ± 0.4
	PCFH (2019)	98.8 ± 0.3	97.7 ± 0.3
	PACH (2020)	98.9 ± 0.2	98.3 ± 0.2
	ADCANs (2021)	99.1 ± 0.2	98.5 ± 0.2
	MMTN (2022)	99.4 ± 0.0	98.7 ± 0.0
	NLPF (2024)	99.7 ± 0.2	99.5 ± 0.3
Semi-supervised	LPL (2024)	99.6 ± 0.2	99.5 ± 0.2
Unsupervised	MDFD-HFA (2024)	99.4 ± 0.1	-
	HERE (2024)	99.8 ± 0.1	**99.7** ± 0.2
	UMFN (ours)	**99.9 ± 0.1**	**99.7 ± 0.1**

TABLE V

RANK-1 RECOGNITION ACCURACY AND VERIFICATION RATE ON LAMP-HQ NIR-VIS DATASET

Type	Method	Rank-1 (%)	VR@FAR = 0.1% (%)
Supervised	ADFL (2018)	95.8	71.0
	PCFH (2019)	96.4	76.8
	PACH (2020)	96.9	78.7
	ADCANs (2021)	97.8	96.1
	MMTN (2022)	97.7	79.2
	NLPF (2024)	98.8	98.3
Semi-supervised	LPL (2024)	98.7	98.5
Unsupervised	HERE (2024)	98.9	**98.8**
	UMFN (ours)	**99.1**	98.7

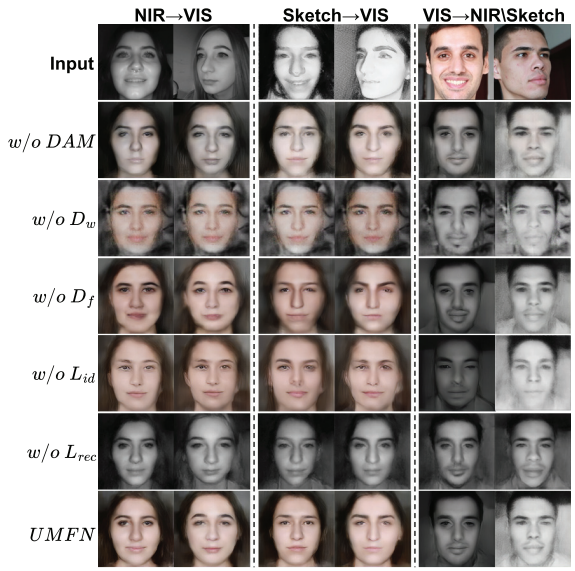


Fig. 11. Face normalization results of the UMFN's variants. The first row depicts the input samples, while the second to sixth rows showcase the face normalization results of UMFN after removing the DAM, the global path discriminator D_w, the local path discriminator D_f, the identity feature similarity loss L_{id}, and the reconstruction loss L_{rec}, respectively. The final row presents the normalization results of the complete UMFN.

LAMP-HQ datasets. It is observed that our method achieves the highest Rank-1 recognition accuracy and verification rate on the CASIA NIR-VIS 2.0 dataset, and performs comparably to the state-of-the-art on the LAMP-HQ dataset. This exceptional performance stems from: (1) The incorporation of a contrastive learning mechanism that aligns the generated prototypes with their real counterparts in the feature space, effectively mitigating the impact of noise and missing features. This enables the HFR encoder F_θ to extract identity-specific features from the generated prototypes that closely resembling those of real prototypes. (2) The fusion of domain-agnostic and identity-discriminative features notably enhances the identity distinguishability of the combined features.

H. Ablation Study

In the UMFN framework, there are six key modules: Encoder (G_{enc}), Decoder (G_{dec}), Mapping Network (*map*), Domain Classifier (D_c), Global Pathway Discriminator (D_w), and Local Pathway Discriminator (D_f). Notably, G_{enc} and G_{dec} act as the foundational encoding and decoding units, respectively, and their removal renders the training process unviable. *map* and D_c are tightly coupled in the domain adversarial training process, forming an integrated module that we refer to as the Domain Adversarial Module (DAM).

To evaluate the contributions of these components, we perform ablation studies by systematically removing DAM, D_w, and D_f from UMFN. Additionally, we investigate the impact of two pivotal loss terms in Eq. (7): L_{id} (identity feature similarity loss) and L_{rec} (reconstruction loss), on the training of UMFN. The results are visualized in Fig. 11, and our key observations are as follows: (1) Removing the DAM results in the retention of original domain information in cross-domain prototypes, highlighting its role in domain adaptation; (2) Removing D_w leads to incomplete faces with poorly generated visual effects, underscoring its importance in ensuring global consistency; (3) Removing D_f causes a loss of local details in generated prototypes, demonstrating its

TABLE VI
ABLATION EXPERIMENTS OF THE CONTRASTIVE LEARNING MECHANISM AND FEATURE FUSION IN THE HFR NETWORK

Feature fusion	Contrastive	Dataset			
		CASIA NIR-VIS 2.0		LAMP-HQ	
		Rank-1	@FAR=0.001	Rank-1	@FAR=0.001
		95.7	95.1	91.2	87.6
✓		97.5	96.7	94.7	90.5
	✓	98.7	98.2	97.3	95.3
✓	✓	**99.9**	**99.7**	**99.1**	**98.7**

contribution to fine-grained feature preservation; (4) Without L_{id}, the generated prototypes fail to preserve identity characteristics, emphasizing its role in maintaining identity consistency; (5) The absence of L_{rec} causes the control masks to fail, preventing the generation of specified domain prototypes, which confirms its essential role in the reconstruction process.

Furthermore, Table VI quantitatively analyzes the impact of the contrastive learning mechanism and feature concatenation on the HFR network training. Results show that contrastive learning significantly enhances the network's utilization of the generated prototypes, allowing it to extract key features similar to real prototypes while reducing the effects of noise and feature loss. Besides, the fusion of domain-invariant and identity-discriminative features improves classification performance.

V. CONCLUSION

This paper has proposed the UMFN model that adeptly overcomes the challenges posed by diverse facial variations and cross-domain adaptability. By reconstructing frontal, neutral-expression facial prototypes and facilitating unsupervised domain adaptation, our approach addresses these limitations with remarkable efficacy. Through joint prototype and feature learning, along with a well-designed HFR network, our method significantly enhances identity recognition accuracy across various domains. Comprehensive experiments demonstrate that the UMFN achieves superior performance in both single-modal and cross-modal face normalization tasks, while the HFR network excels in heterogeneous face recognition.

REFERENCES

[1] M. Pang, B. Wang, S. Huang, Y.-M. Cheung, and B. Wen, "A unified framework for bidirectional prototype learning from contaminated faces across heterogeneous domains," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1544–1557, 2022.

[2] D. Liu, X. Gao, C. Peng, N. Wang, and J. Li, "Universal heterogeneous face analysis via multi-domain feature disentanglement," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 735–747, 2024.

[3] J. Xin, Z. Wei, N. Wang, J. Li, and X. Gao, "Large pose face recognition via facial representation learning," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 934–946, 2024.

[4] J. Xin, Z. Wei, N. Wang, J. Li, X. Wang, and X. Gao, "Learning a high fidelity identity representation for face frontalization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6952–6964, Nov. 2023.

[5] W. Carver and I. Nwogu, "Facial expression neutralization with StoicNet," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jun. 2021, pp. 201–208.

[6] D. Guo et al., "Face illumination normalization based on generative adversarial network," *Natural Comput.*, vol. 22, no. 1, pp. 105–117, Mar. 2023.

[7] W. Xu, X. Xie, and J. Lai, "RelightGAN: Instance-level generative adversarial network for face illumination transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 3450–3460, 2021.

[8] N. Zhang, N. Liu, J. Han, K. Wan, and L. Shao, "Face de-occlusion with deep cascade guidance learning," *IEEE Trans. Multimedia*, vol. 25, pp. 3217–3229, 2022.

[9] M. Pang, B. Wang, Y. Cheung, Y. Chen, and B. Wen, "VD-GAN: A unified framework for joint prototype and representation learning from contaminated single sample per person," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2246–2259, 2021.

[10] J. Xu et al., "Personalized face inpainting with diffusion models by parallel visual attention," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5420–5430.

[11] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, 2023.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[13] M. Pang, B. Wang, M. Ye, Y.-M. Cheung, Y. Chen, and B. Wen, "DisP+V: A unified framework for disentangling prototype and variation from single sample per person," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 867–881, Feb. 2023.

[14] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014, pp. 2672–2680.

[15] M. Sun, J. Wang, J. Liu, J. Li, T. Chen, and Z. Sun, "A unified framework for biphasic facial age translation with noisy-semantic guided generative adversarial networks," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1513–1527, 2022.

[16] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.

[17] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.

[18] M. Pang, Y.-M. Cheung, Q. Shi, and M. Li, "Iterative dynamic generic learning for face recognition from a contaminated single-sample per person," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1560–1574, Apr. 2021.

[19] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proc. 26th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2018, pp. 627–635.

[20] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8398–8406.

[21] X. Luan, J. Zheng, and W. Li, "Learning unsupervised face normalization through frontal view reconstruction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5201–5212, Aug. 2022.

[22] C. Li, S. Ge, D. Zhang, and J. Li, "Look through masks: Towards masked face recognition with de-occlusion distillation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3016–3024.

[23] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee, "Complete face recovery GAN: Unsupervised joint face rotation and de-occlusion from a single-view image," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1173–1183.

[24] 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.*, Oct. 2017, pp. 2223–2232.

[25] B. Cao, N. Wang, J. Li, Q. Hu, and X. Gao, "Face photo-sketch synthesis via full-scale identity supervision," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108446.

[26] C. Liang, M. Zhu, N. Wang, H. Yang, and X. Gao, "PMSGAN: Parallel multistage GANs for face image translation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 1–14, Jul. 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10014017

[27] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "BBDM: Image-to-image translation with Brownian bridge diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1952–1961.

[28] B. Luo, Z. Wang, J. Wang, J. Zhu, X. Zhao, and Y. Gao, "Multi-energy guided image translation with stochastic differential equations for near-infrared facial expression recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Mar. 2024, vol. 38, no. 1, pp. 565–573.

[29] Y. Que, L. Xiong, W. Wan, X. Xia, and Z. Liu, "Denoising diffusion probabilistic model for face sketch-to-photo synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 10424–10436, Oct. 2024.

[30] X. Qi et al., "Biphasic face photo-sketch synthesis via semantic-driven generative adversarial network with graph representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2182–2195, Feb. 2025.
[31] W. Hu, Y. Yang, and H. Hu, "Pseudo label association and prototype-based invariant learning for semi-supervised NIR-VIS face recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 1448–1463, 2024.
[32] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2079–2089, May 2017.
[33] S. Liu, D. Yi, Z. Lei, and S. Z. Li, "Heterogeneous face image matching using multi-scale features," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 79–84.
[34] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–7.
[35] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1979–1993, Aug. 2017.
[36] 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.
[37] C. Peng, N. Wang, J. Li, and X. Gao, "Soft semantic representation for cross-domain face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 346–360, 2021.
[38] W. Hu and H. Hu, "Domain-private factor detachment network for NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1435–1449, 2022.
[39] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, "Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 4, pp. 886–896, Apr. 2019.
[40] 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.
[41] W. Hu, W. Yan, and H. Hu, "Dual face alignment learning network for NIR-VIS face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2411–2424, Apr. 2022.
[42] Y. Yang, W. Hu, H. Lin, and H. Hu, "Robust cross-domain pseudo-labeling and contrastive learning for unsupervised domain adaptation NIR-VIS face recognition," *IEEE Trans. Image Process.*, vol. 32, pp. 5231–5244, 2023.
[43] W. Hu, B. Liu, H. Zeng, Y. Hou, and H. Hu, "Adversarial decoupling and modality-invariant representation learning for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5095–5109, Aug. 2022.
[44] 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.
[45] J. Yu, J. Cao, Y. Li, X. Jia, and R. He, "Pose-preserving cross spectral face hallucination," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2019, pp. 1018–1024.
[46] Z. Yang, J. Liang, C. Fu, M. Luo, and X.-Y. Zhang, "Heterogeneous face recognition via face synthesis with identity-attribute disentanglement," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1344–1358, 2022.
[47] 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.*, vol. 44, no. 6, pp. 2938–2952, Jun. 2022.
[48] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Dec. 2017, pp. 5769–5779.
[49] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, Dec. 2014, pp. 3320–3328.
[50] A. Yu et al., "LAMP-HQ: A large-scale multi-pose high-quality database and benchmark for NIR-VIS face recognition," *Int. J. Comput. Vis.*, vol. 129, pp. 1467–1483, Feb. 2021.
[51] 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.
[52] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," Dept. School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001, 2012, vol. 3, no. 3, p. 8.
[53] H. Kiani Galoogahi and T. Sim, "Face sketch recognition by local radon binary pattern: LRBP," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1837–1840.
[54] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[55] B. F. Klare et al., "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.
[56] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life'Images, Detection, Alignment, Recognit.*, 2008, pp. 1–15.
[57] Y. Yang, W. Hu, and H. Hu, "Neutral face learning and progressive fusion synthesis network for NIR-VIS face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5750–5763, Oct. 2024.
[58] 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.* Cham, Switzerland: Springer, 2016, pp. 87–102.
[59] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.
[60] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9851–9858.
[61] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[63] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5911–5920.
[64] J. Liao, T. Guha, and V. Sanchez, "Self-supervised random mask attention GAN in tackling pose-invariant face recognition," *Pattern Recognit.*, vol. 159, Mar. 2025, Art. no. 111112.
[65] J. R. A. Moniz, C. Beckham, S. Rajotte, S. Honari, and C. Pal, "Unsupervised depth estimation, 3D face rotation and replacement," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2018, pp. 9736–9746.
[66] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Apr. 2018, vol. 32, no. 1, pp. 7355–7362.
[67] 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.
[68] W. Hu and H. Hu, "Adversarial disentanglement spectrum variations and cross-modality attention networks for NIR-VIS face recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 145–160, 2021.
[69] M. Luo, H. Wu, H. Huang, W. He, and R. He, "Memory-modulated transformer network for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2095–2109, 2022.
[70] Y. Yang, W. Hu, and H. Hu, "Unsupervised NIR-VIS face recognition via homogeneous-to-heterogeneous learning and residual-invariant enhancement," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2112–2126, 2024.

Meng Pang (Member, IEEE) received the B.Sc. and M.Sc. degrees in software engineering from Dalian University of Technology, Dalian, China, in 2013 and 2016, respectively, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China, in 2019. He was a Post-Doctoral Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, from 2020 to 2022. He is currently a Distinguished Professor with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China. His research interests include image processing, artificial intelligence security, and artificial intelligence medical. He serves as an Associate Editor for *Expert Systems*.

Wenjun Zhang received the B.Eng. degree in computer science and technology from Jiangxi Normal University, Nanchang, China, in 2023. He is currently pursuing the M.Eng. degree with Nanchang University. His research interests include text-to-image generation and image editing.

Yang Lu (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in software engineering from the University of Macau, Macau, China, in 2012 and 2014, respectively, and the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, China, in 2019. He is currently an Assistant Professor with the Department of Computer Science and Technology, School of Informatics, Xiamen University, Xiamen, China. His current research interests include open-world robust deep learning, such as long-tail learning, federated learning, label-noise learning, and continual learning.

Yiu-ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China. His research interests include machine learning, pattern recognition, visual computing, and optimization. He is an IET Fellow, a BCS Fellow, a RSA Fellow, and a IETI Distinguished Fellow. He serves as an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and *Pattern Recognition*.

Nanrun Zhou received the Ph.D. degree in communication and information systems from Shanghai Jiao Tong University, in 2005. Since 2006, he has been a Faculty Member with the Department of Electronic Information Engineering, Nanchang University, where he has been a Professor, from November 2010 to June 2022, and a Gang Jiang Distinguished Professor, since 2014. Since July 2022, he has been a Faculty Member with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science. He has published over 280 papers in refereed international conferences and journals. He has been selected in the first or second rank of Jiangxi Province Baiqianwan Talent for the New Century Programme, the Young Scientist of Jiangxi Province (Jinggang Star), and the Ganpo Programme 555 for Outstanding Talent, and the Major Academic Discipline and Technical Leader of Jiangxi Province, leading a team of researchers carrying out cutting-edge research in the field of information security. He is currently an Associate Editor of *IET Optoelectronics* and *Frontiers in Physics* and was ever an Editorial Board Member of *China Communications*.