From Modalities to Styles: Rethinking the Domain Gap in Heterogeneous Face Recognition

Anjith George, *Member, IEEE,* Sébastien Marcel, *Senior Member, IEEE*

Abstract—Heterogeneous Face Recognition (HFR) focuses on matching faces from different domains, for instance, thermal to visible images, making Face Recognition (FR) systems more versatile for challenging scenarios. However, the domain gap between these domains and the limited large-scale datasets in the target HFR modalities make it challenging to develop robust HFR models from scratch. In our work, we view different modalities as distinct styles and propose a method to modulate feature maps of the target modality to address the domain gap. We present a new Conditional Adaptive Instance Modulation (CAIM) module that seamlessly fits into existing FR networks, turning them into HFR-ready systems. The CAIM block modulates intermediate feature maps, efficiently adapting to the style of the source modality and bridging the domain gap. Our method enables end-to-end training using a small set of paired samples. We extensively evaluate the proposed approach on various challenging HFR benchmarks, showing that it outperforms state-of-the-art methods. The source code and protocols for reproducing the findings will be made publicly available.

Index Terms—Face Recognition, Heterogeneous Face Recognition, Style transfer, Instance Normalization, Biometrics.

✦

1 INTRODUCTION

Facial recognition (FR) technology has gained popularity in the field of access control due to its high efficiency and user-friendly nature. Most state-of-the-art FR methods achieve excellent performance in "in the wild" conditions and even reach a level comparable to human performance in recognizing faces [1], thanks to the advancement of convolutional neural networks (CNN). Typically, FR systems are designed to work within a homogeneous domain, meaning that both the enrollment and matching phases are conducted using the same type of data, usually facial images captured with an RGB camera. Nonetheless, there are scenarios where performing matching in a heterogeneous setting could be beneficial. For instance, near-infrared (NIR) cameras, commonly found in smartphones and security cameras, offer superior performance across various lighting conditions and exhibit resilience to spoofing attacks [2], [3]. Despite these advantages, developing an FR system tailored for NIR imagery requires an extensive collection of annotated training data, which is often scarce.

Heterogeneous Face Recognition (*HFR*) systems are designed to facilitate cross-domain matching (Fig. 1), enabling the comparison of enrolled RGB images with NIR (or other types of) images without necessitating the enrollment of separate modalities [5]. This approach proves to be invaluable, especially in conditions where acquiring visible images is challenging. For example, thermal images can be utilized for recognition purposes regardless of the lighting conditions, making face recognition feasible day and night, and even at considerable distances. *HFR* systems are versatile and capable of processing and matching facial images from diverse sources and modalities, which significantly broadens
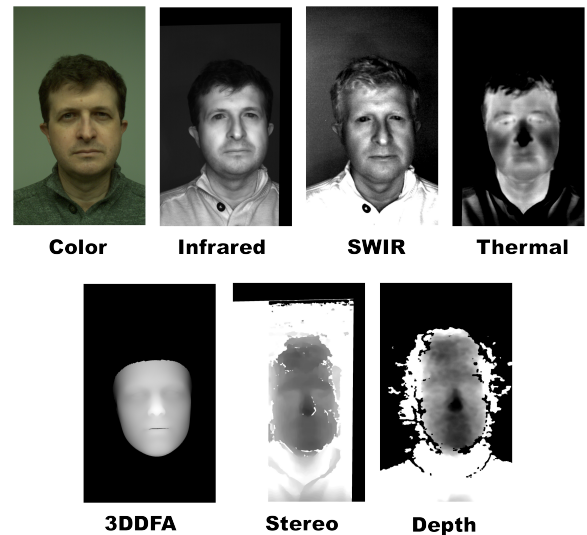

Fig. 1. This figure shows the facial images of the same individual acquired using distinct imaging modalities (Images taken from MCXFace dataset [4]). The task in *HFR* is to facilitate cross-domain matching while overcoming the challenges posed by the domain gap.

the potential applications and utility of face recognition systems across various challenging scenarios.

HFR extends the use of Face Recognition (FR) systems to challenging scenarios, such as those involving low-lighting or long-range, by capitalizing on the specific characteristics of imaging modalities. HFR approaches effectively mitigate certain constraints, broadening the scope and applicability of FR systems. Despite its usefulness, developing a Heterogeneous Face Recognition (*HFR*) system comes with its own challenges. Cross-domain matching is challenging primarily because of the domain gap. This

• *All authors are with Idiap Research Institute, Martigny, Switzerland. Sébastien Marcel is also affiliated with Université de Lausanne (UNIL), Lausanne, Switzerland. E-mail: {anjith.george, sebastien.marcel}@idiap.ch*

gap can lead to a drop in performance when face recognition (FR) networks, which are typically trained on visible-light images, are applied to images from different sensing modalities [6]. Moreover, creating models that are robust to both visible and other modalities is challenging, exacerbated by the limited availability of large-scale multimodal datasets. Collecting large-scale paired datasets for these additional modalities is not only challenging but can also incur significant costs. Hence, it is essential to devise an HFR framework that requires only a limited set of labeled samples for training the models.

In our approach, we build upon face recognition networks pre-trained with a large dataset of faces from the visible spectrum, using it as our foundational network. We address the challenge of different modalities by conceptualizing them as unique *styles*. Our proposed framework is designed to bridge the domain gap by adapting the network's intermediate feature maps to align with these styles. The core of our method is the introduction of a new module, which we refer to as the Conditional Adaptive Instance Modulation (CAIM) [7]. The CAIM module can be integrated seamlessly into the face recognition network's intermediate stages. This trainable module is capable of being trained from scratch, transforming a standard face recognition system into an HFR network capable of handling a variety of modalities, all while requiring a minimal number of training sample pairs.

The main contributions of this work are as follows:

- We conceptualize the domain gap in Heterogeneous Face Recognition (HFR) as a manifestation of distinct *styles* from different imaging modalities, and address this domain gap as a style modulation problem.
- A new trainable component called Conditional Adaptive Instance Modulation (CAIM) is introduced, which can transform a pre-trained FR network into a heterogeneous face recognition network, requiring only a limited number of paired samples for training.
- We implemented our approach with two different face recognition models to evaluate the generalization of our approach.
- We demonstrate the robustness and effectiveness of our proposed method through extensive evaluation on various challenging HFR benchmarks.

Finally, the protocols and source codes will be made available publicly [1].

The structure of the rest of the paper is organized in the following manner: In Section 2, we review the previous literature in Heterogeneous Face Recognition (HFR). The specifics of the CAIM approach are elaborated in Section 3. Section 4 and 5 provide a thorough evaluation of the CAIM method, including comparative analyses with state-of-the-art methods, followed by in-depth discussions. Finally, Section 6 concludes the paper with a summary of our findings and proposes directions for future research.

2 RELATED WORK

The objective of Heterogeneous Face Recognition (HFR) methods is to accurately match faces across images captured by different sensing modalities. Yet, the discrepancy between these domains, known as the domain gap, can impair the efficacy of face recognition networks when performing a direct comparison of multi-modal images. Therefore, it's crucial for HFR methodologies to

effectively close this modality gap. In this section, we review recent literature on strategies proposed for addressing the domain gap.

2.1 Invariant feature-based methods

Various strategies have been developed for Heterogeneous Face Recognition (HFR) with the goal of extracting features that remain consistent across different imaging modalities. Liao *et al.* [8] introduced a technique that relies on the Difference of Gaussian (DoG) filters combined with multi-scale block Local Binary Patterns (MB-LBP) to capture invariant features. Klare *et al.* [9] proposed a method employing Local Feature-based Discriminant Analysis (LFDA), which utilizes Scale-Invariant Feature Transform (SIFT) and Multi-Scale Local Binary Pattern (MLBP) as feature descriptors. Zhang *et al.* [10] proposed the Coupled Information-Theoretic Encoding (CITE) approach, which seeks to maximize the mutual information across modalities within quantized feature spaces. Approaches based on Convolutional Neural Networks (CNNs) have also been applied to HFR, demonstrating the versatility of deep learning models in this context [6], [11]. Roy *et al.* [12] proposed a method termed Local Maximum Quotient (LMQ), specifically designed to identify invariant features in cross-modal facial imagery. In [13], authors introduced a feature-based approach for HFR for composite sketch recognition. This approach involved extracting features using the Scale-Invariant Feature Transform (SIFT) and the Histogram of Oriented Gradient (HOG) from different facial components. These features were then integrated at the score level, where the facial components were combined using a linear function.

2.2 Common-space projection methods

Common-space projection methods aim to learn a transformation that projects facial images from various modalities into a unified subspace, thereby reducing the domain gap [11], [14]. Lin and Tang [15] devised a method known as common discriminant feature extraction to extract features from cross-modal images and align them within a shared feature space. Yi *et al.* [16] utilized Canonical Correlation Analysis (CCA) to correlate Near-Infrared (NIR) and Visible Spectrum (VIS) face images. Lei *et al.* [17], [18] introduced regression-based techniques to establish mapping functions that bridge the gap between different modalities. Sharma and Jacobs [19] developed a method based on Partial Least Squares (PLS) to learn a linear mapping that maximizes the covariance between face images across modalities. Klare and Jain [5] proposed a method for representing face images by their similarities to a predefined set of prototype faces, followed by projecting these representations onto a linear discriminant analysis subspace for recognition purposes. In [20], authors suggested that the high-level features in convolutional neural networks, when trained on visible light spectra, are actually domain-agnostic and can be used to encode images from other sensing modalities. They adapt the initial layers of a pre-trained FR model, termed Domain-Specific Units (DSUs), to minimize the domain gap, training the entire system in a contrastive learning setting. A challenge with this approach is determining the exact number of layers to adapt, which requires thorough experimentation to optimize. Liu *et al.* [21] proposed a novel method known as Coupled Attribute Learning for HFR (CAL-HFR), which uniquely does not require manual labeling of facial attributes. This method utilizes deep convolutional networks to map face images from heterogeneous scenarios into a shared
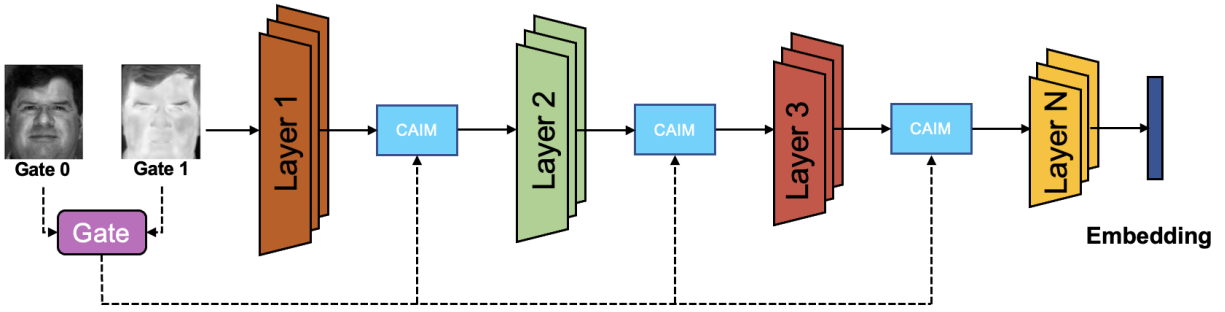

Fig. 2. Schematic diagram of the proposed framework: Layer 1 to Layer N represent the frozen blocks of layers from a pretrained Face Recognition (FR) model. The CAIM module is inserted between the initial few blocks.

space. Additionally, they introduced the Coupled Attribute Guided Triplet Loss (CAGTL), a specially designed loss function aimed at addressing the issues of inaccurately estimated attributes in end-to-end training. Recently, Liu *et al.* [22] proposed a semi-supervised learning approach for modality-independent heterogeneous face recognition (HFR) representation, termed as Modality-Agnostic Augmented Multi-Collaboration representation for Heterogeneous Face Recognition (MAMCO-HFR). This method introduces a multi-collaborative face representation that leverages interactions across various network depths to extract potent discriminative information for identity recognition. Additionally, they proposed a modality-agnostic augmentation technique that creates adversarial disturbances to effectively map unlabeled faces into a modality-agnostic domain.

2.3 Synthesis based methods

Synthesis-based approaches in Heterogeneous Face Recognition (HFR) [23], [24] focus on creating images in the source domain from those in the target modality. This synthetic generation facilitates the use of standard face recognition networks for biometric identification. Wang *et al.* [25] explored a patch-based synthetic method utilizing Multi-scale Markov Random Fields, and Liu *et al.* [26] employed Locally Linear Embedding (LLE) for establishing a pixel-wise correspondence between visible (VIS) images and viewed sketches. The use of CycleGAN, as presented in [27], for unpaired image-to-image translation has paved the way for transforming target domain images to match the source domain [28]. Furthermore, Zhang *et al.* [29] introduced a method using Generative Adversarial Networks (GANs) to create photo-realistic VIS images from polarimetric thermal images through GAN-based Visible Face Synthesis (GAN-VFS). Several recent approaches have been proposed using GANs for the synthesis of VIS images from another modality, such as the Dual Variational Generation (DVG-Face) framework [24], which achieved state-of-the-art results in many challenging *HFR* benchmarks. Liu *et al.* [30] introduced the Heterogeneous Face Interpretable Disentangled Representation (HFIDR), a novel approach capable of explicitly interpreting the dimensions of face representation. This method focuses on extracting latent identity information for cross-modality recognition and employs a technique to transform the modality factor, enabling the synthesis of cross-modality faces. In [31], authors proposed the Memory-Modulated Transformer Network (MMTN) for HFR, treating the problem as an unsupervised, reference-based "one-to-many" generation problem. The MMTN incorporates a memory module to capture prototypical

style patterns and a style transformer module to blend the styles of input and reference images at a local level. Recently, George *et al.* [4] introduced the concept of Prepended Domain Transformers (PDT), which prepends a trainable neural network module to a pre-trained FR network, converting it into an HFR network. This module translates feature representations to align cross-domain embeddings in the feature space, without the need for explicit generation of source domain images.

2.4 Challenges in HFR

In recent literature, Heterogeneous Face Recognition (HFR) methods, particularly those utilizing Generative Adversarial Networks (GANs), have gained prominence for their synthesis-based approaches. These methods, such as DVG-Face and GAN-VFS [24], [29] achieve reasonable results in generating high-fidelity images. Leveraging a pre-trained FR model in such synthesis-based HFR methods obviates the requirement for a vast amount of training data to develop a new FR model. Nonetheless, the synthesis step introduces a significant computational burden, which may hinder its practical deployment in real-life scenarios. We propose a different perspective: treating the domain gap between visible images and images from other modalities as a variation in "styles". By adopting this viewpoint, we can address the domain gap directly within the feature space through modulation of the feature maps. This strategy eliminates the computational and memory-intensive process of synthesizing images in the source modality.

3 PROPOSED METHOD

We follow the notations consistent with recent literature [4], [20], [32], [33] to formally define the *HFR* task.

3.1 Formal definition of *HFR*

Consider a domain $\mathcal{D}$ that includes a set of samples $X \in \mathbb{R}^d$ and a marginal distribution $P(X)$ (of dimensionality d). The goal of a face recognition (FR) system, $\mathcal{T}^{fr}$, can be characterized by a label space Y with a conditional probability $P(Y|X, \Theta)$, where X and Y represent random variables, and Θ denotes the parameters of the model. In the training stage of an FR system, the conditional probability $P(Y|X, \Theta)$ is typically determined through supervised learning using a face dataset $X = x_1, x_2, ..., x_n$ and their corresponding identity labels $Y = y_1, y_2, ..., y_n$.

In the heterogeneous face recognition (*HFR*) problem, we assume the presence of two domains: a source domain $\mathcal{D}^s = X^s, P(X^s)$ and a target domain $\mathcal{D}^t = X^t, P(X^t)$, both sharing

the labels Y. The objective of the *HFR* problem, $\mathcal{T}^{hfr}$, is to estimate a $\hat{\Theta}$ such that $P(Y|X^s, \Theta) = P(Y|X^t, \hat{\Theta})$.

3.2 Proposed Framework

In our proposed approach, we consider face images from various modalities as separate *styles*, considering the domain discrepancy in the *HFR* challenge to be a manifestation of these style variations. We propose that by addressing the domain-specific *style*, we can reduce the domain gap. To accomplish this, we employ conditional modulation on the intermediate feature maps within a pre-trained face recognition network.

Using the parameters Θ_{FR} from a pre-trained face recognition (FR) model developed on the source domain $\mathcal{D}^s$, our strategy does not alter the model's original weights. Instead, we inject a set of trainable network modules between the frozen layers of the FR network, named CAIM, which are designed to modulate the intermediate feature maps. The CAIM modules perform normalization and style modulation on feature maps from the target modality, to align the embeddings of corresponding samples from both modalities in the embedding space. Figure 2 illustrates the overall design of our proposed system. We incorporate the CAIM modules primarily within the initial blocks of the network, as these are more closely related to modality-specific characteristics. An external gating mechanism is deployed to enable the CAIM modules solely for the target modality data while allowing the source modality data to pass unaffectedly, thereby mitigating the risk of catastrophic forgetting.

The *HFR* task can be mathematically formulated as:

$$P(Y|X_t, \hat{\Theta}) = P(Y|X_t, [\Theta_{FR}, \theta_{CAIM_{i,i \in (1,2,..,K)}}]) \qquad (1)$$

Where, θ_{CAIM_i} denotes the ith CAIM block out of K blocks.

The CAIM blocks, specified by the parameters θ_{CAIM_i}, are the only trainable components and can be fine-tuned in a supervised manner. When the system is trained, the CAIM module acts as a pass-through for images from the source domain (X^s), effectively allowing the network to produce the reference embeddings through Θ_{FR}. However, for images from the target domain (X^t), the processing involves both the frozen network layers (Θ_{FR}) and the newly introduced CAIM blocks. The training utilizes a contrastive loss function, as described by [34], to align the embeddings in the shared representational space. The contrastive loss is given as:

$$\mathcal{L}_{Contrastive}(\hat{\Theta}, Y_p, X_s, X_t) = (1 - Y_p)\frac{1}{2}D_W^2$$
$$+ Y_p\frac{1}{2}max(0, m - D_W)^2 \qquad (2)$$

Where $\hat{\Theta}$ represents the network's weights together with the frozen weights, X_s and X_t denote heterogeneous image pairs. The label Y_p indicates whether the pairs share the same identity. The margin in the contrastive loss function is denoted by m, while D_W represents the metric used to compute the distance between the embeddings of the two images in a pair. The chosen distance measure D_W could be the Euclidean distance or the cosine distance, depending on which is used to compare the feature representations produced by the network. Further details on the CAIM block's design are provided in subsequent subsections.

3.3 Architecture of the CAIM block

Figure 2 illustrates the components of the proposed Conditional Adaptive Instance Modulation framework. The CAIM blocks are inserted between the frozen layers of the pre-trained face recognition network. The detailed design of the CAIM block is shown in Fig. 3. This block takes an input feature map along with a global gating signal and outputs a feature map with the same dimensions. The first component in the CAIM block is an Instance Normalization (IN) layer that normalizes each feature map individually, without trainable affine parameters. Following this, a parallel Convolutional Neural Network (CNN) module takes the original, un-normalized feature map to extract a shared representation. This CNN module consists of two sets of 3×3 convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation function, and then a Global Average Pooling (GAP) layer. To generate the scaling and shifting parameters for the normalized feature maps, two dense (fully connected) layers are appended to the shared representation. These dense layers are tasked with calculating the parameters that modulate the normalized feature maps. Furthermore, a residual connection is incorporated into the network. When the global gate signal is set to zero, the CAIM block acts as an identity function, obtaining the same embeddings from the original pre-trained face recognition network for the reference modality, effectively bypassing the modulation process.

3.4 Style Modulation for HFR

In this section, we discuss using style modulation as a strategy to bridge the domain gap between visible and other modalities, starting with the application of Instance Normalization [35].

3.4.1 Instance normalization

Previous works have demonstrated that the statistical properties of feature maps in deep neural networks (DNNs) effectively capture the style of images [36]. Ulyanov *et al.* [35] showed that substituting batch normalization layers with Instance Normalization (IN) significantly enhances style transfer. Instance Normalization layer normalizes the feature maps, and this process can be represented as follows:

$$\text{IN}(x) = \gamma \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \beta \qquad (3)$$

Here, $\gamma, \beta \in \mathbb{R}^C$ represent affine parameters learned from the data, while $\mu(x)$ and $\sigma(x)$ are calculated across spatial dimensions for each individual sample, as opposed to across mini-batches in BatchNorm.

Dumoulin *et al.* [37] subsequently extended Instance Normalization to Conditional Instance Normalization (CIN). In this approach, a set of parameters, γ^s and β^s, can be learned for a predefined set of styles, denoted by s.

In [38], Huang *et al.* introduced a network module named Adaptive Instance Normalization (AdaIN), specifically designed for image style transfer. This module is designed to align the mean and variance of content features with those of style features in the context of image style transfer. The authors suggest that Instance Normalization facilitates style normalization by normalizing the feature statistics, namely the mean and variance. The AdaIN module operates by taking a content input x and a style input y and aligning the channel-wise mean and variance of x to correspond with those of y. Unlike other methods, AdaIN does not utilize
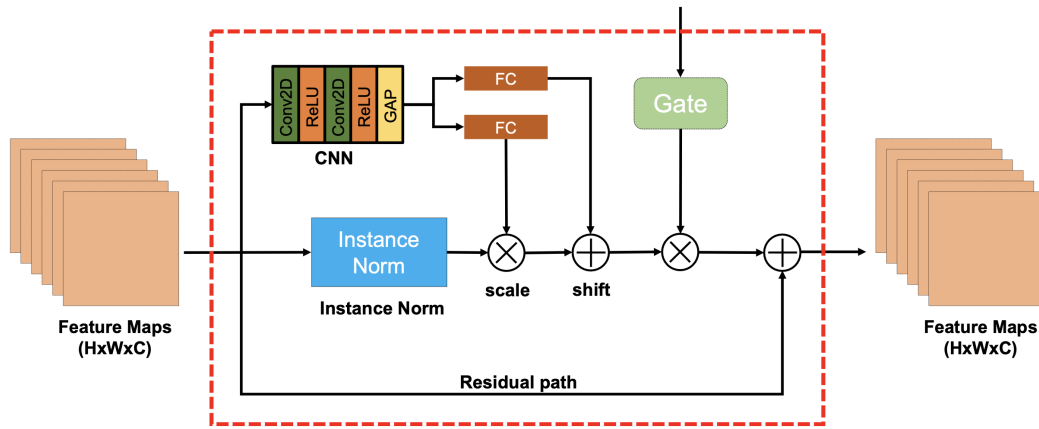
Fig. 3. Architecture of the Conditional Adaptive Instance Modulation (CAIM) block. The global gate signal activates the block. The gate signal becoming zero deactivates this module and the entire module functions as an identity block in this case due to the residual path.

learnable affine parameters; rather, it dynamically computes these parameters based on the style input.

$$\text{AdaIN}(x,y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y) \qquad (4)$$

The normalized content input is scaled by $\sigma(y)$ and shifted by $\mu(y)$. Similar to Instance Normalization, these statistics are computed across spatial locations.

Recent studies [39], have shown that mixing instance-level feature statistics from multiple source domains probabilistically can significantly improve domain generalization. This enhancement is achieved by integrating a variety of styles during the training phase, which leads to the development of a model that is more robust and adaptable across different domains. A critical aspect to note is that this mixing of styles and domains occurs during the initial training phase of the model, where the model is trained from scratch with inputs from various domains. This is a crucial distinction, especially in the context of Heterogeneous Face Recognition (HFR), where there is not enough target domain data to do mixing in the training phase. Hence we often start with a face recognition model that is already pre-trained on the source domain. In the HFR context, we modify this approach by conditionally modulating the feature maps instead of mixing them, and adapting the concept to suit the specific needs of HFR. This conditional modulation in HFR allows the feature statistics of the target modality to adapt, enhancing the model's ability to recognize faces across varied domains without the need for retraining from scratch.

3.4.2 Conditional Adaptive Instance Modulation

The Adaptive Instance Normalization (AdaIN) module, as previously mentioned, is adept at producing images that match or emulate the style of another image. It is particularly useful in creating images that adopt the style characteristics of a reference image. In the context of Heterogeneous Face Recognition (HFR), the objective is to adjust the style of target modality images so that they align with the style of visible spectrum images. This alignment is crucial to ensure that the final image embeddings are consistent across different modalities. This is of particular importance considering that the pre-trained face recognition network

is initially trained on a large dataset of visible spectrum images, making it essential to align the styles between different modalities for effective cross-modal recognition.

Consider an intermediate feature map in the face recognition network, denoted by $F \in \mathbb{R}^{C \times H \times W}$. Here, C, H, and W represent the number of channels, height, and width of the feature map, respectively.

For the target modality, we would like to modulate these feature maps such that the output embedding from the network aligns for the source and target modalities.

To accomplish this, we modulate the intermediate feature map using the CAIM block.

$$\hat{\mathbf{F}} = \text{CAIM}(\mathbf{F}) \qquad (5)$$

The CAIM block's main component is similar to adaptive instance normalization (AdaIN) particularly in its ability to normalize and modify the style of target images. However, unlike AdaIN which relies on an external style input, our approach derives modulation factors directly from the raw input feature maps using a CNN module. Furthermore, we combine this step in a residual fashion while injecting the CAIM block into a pretrained network.

To elaborate further, we first estimate a shared representation from the input feature map by utilizing a shallow CNN network with global average pooling.

$$\xi_{\mathbf{f}} = \text{GAP}\big(\text{CNN}(\mathbf{F})\big) \qquad (6)$$

The σ_f and μ_f parameters are estimated from this shared representation with two fully connected (FC) layers:

$$\sigma_{\mathbf{f}} = \text{FC}_\sigma(\xi_{\mathbf{f}}) \qquad (7)$$
$$\mu_{\mathbf{f}} = \text{FC}_\mu(\xi_{\mathbf{f}}) \qquad (8)$$

The estimated parameters are utilized to scale and shift the normalized feature maps:

$$\text{AIM}(\mathbf{F}) = \sigma_{\mathbf{f}}\left(\frac{\mathbf{F} - \boldsymbol{\mu}(\mathbf{F})}{\boldsymbol{\sigma}(\mathbf{F})}\right) + \mu_{\mathbf{f}} \qquad (9)$$

To ensure stable training, we incorporate a residual connection in the proposed framework. Additionally, when incorporating this

module, a gate is added to activate the module exclusively for the target modality, leaving the feature maps of the source modality unaltered.

The CAIM block can be represented as follows:

$$\text{CAIM}(\mathbf{F}, \mathbf{g}) = \mathbf{g} \cdot \text{AIM}(\mathbf{F}) + \mathbf{F} \qquad (10)$$

Where, g denotes the gate, $\mathbf{g} = 1$ for the target modality, and $\mathbf{g} = 0$ for the source modality (visible images).

3.5 Face Recognition backbone

To ensure reproducibility, we used the publicly available pre-trained *Iresnet100* face recognition model provided by Insight-face [40]. The model was trained on the MS-Celeb-1M dataset [2], which includes over 70,000 identities. The pre-trained face recognition model accepts three-channel images at a resolution of 112×112 pixels. Before passing through the FR network, faces are aligned and cropped to ensure eye center coordinates align with predetermined points. In cases where the input is a single-channel image (like NIR or thermal images), the single channel is duplicated across all three channels, without altering the network's architecture.

3.6 Implementation details

The Conditional Adaptive Instance Modulation (CAIM) block employs a contrastive learning approach, within a Siamese network framework [34]. For all experiments, we set the margin parameter to 2.0. The training used the Adam Optimizer with a learning rate of 0.0001, over 50 epochs, and a batch size of 90. We developed the framework in PyTorch and using the Bob library [41], [42] [3]. In this setup, the frozen layers of the pre-trained face recognition network are shared between the source and target modalities. The CAIM block, inserted between these frozen layers, is operational exclusively for the target modality, activated when the global gate signal is one ($gate = 1$). Conversely, for reference channel images (visible spectrum) with $gate = 0$, the CAIM block essentially acts as a bypass through the residual branch. Only the CAIM block's parameters are updated during training. The experiments are reproducible, and the source code and protocols will be made available publicly.

4 EXPERIMENTS

This section outlines the outcomes of a comprehensive series of experiments carried out using the CAIM framework. Our main objective was to assess the effectiveness of the CAIM method in VIS-Thermal *HFR* , across various challenging datasets. Furthermore, we compared the performance of the CAIM approach against other heterogeneous settings such as VIS-Sketch, VIS-NIR, and VIS-Low Resolution VIS. In all our experiments, we used the standard cosine distance for comparison.

4.1 Databases and Protocols

The following section describes the datasets used (Fig. 4) in the evaluations.

Tufts face dataset: The Tufts Face Database [43] is a collection of face images from various modalities for the *HFR* task.
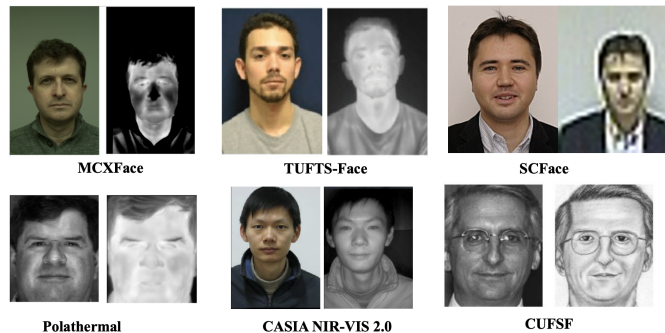
Fig. 4. Sample images from source and target modalities from six different HFR datasets. Images are from MCXFace [4], Tufts Face [43], SCFace [44], Polathermal [45] CASIA NIR-VIS 2.0 [46], CUHK Face Sketch FERET Database (CUFSF) [10] respectively.

For our evaluation of VIS-Thermal *HFR* performance, we use the thermal images provided in the dataset. The dataset comprises 113 identities, consisting of 39 males and 74 females from different demographic regions, and includes images from different modalities for each subject. We adopt the same procedure as in [24], selecting 50 identities at random for the training set and using the remaining subjects for the test set. We report Rank-1 accuracies and Verification rates at false acceptance rates (FAR) of 1% and 0.1% for comparison.

MCXFace Dataset: The MCXFace Dataset [4] includes images of 51 individuals captured in various illumination conditions and three distinct sessions using different channels. The channels available include RGB color, thermal, near-infrared (850 nm), short-wave infrared (1300 nm), Depth, Stereo depth, and depth estimated from RGB images. All channels are spatially and temporally registered across all modalities. Five different folds were created for each of the protocols by randomly dividing the subjects into $train$ and dev partitions. Annotations for the left and right eye centers for all images are also included in the dataset. We have performed the evaluations on the challenging "VIS-Thermal" protocols of this dataset.

Polathermal dataset: The Polathermal dataset [45] is an *HFR* dataset collected by the U.S. Army Research Laboratory (ARL). It contains polarimetric LWIR imagery together with color images for 60 subjects. The dataset has conventional thermal images and polarimetric images for each subject. For our experiments, we use conventional thermal images and follow the five-fold partitions introduced in [20]. Specifically, 25 identities are used for training, while the remaining 35 identities are used for testing. We report the average Rank-1 identification rate from the evaluation set of the five folds.

SCFace dataset: The SCFace dataset [44] consists of high-quality enrollment images for face recognition, while the probe samples are low-quality images from various surveillance scenarios captured by different cameras. There are four different protocols in the dataset, based on the quality and distance of the probe samples: close, medium, combined, and far, with the "far" protocol being the most challenging. In total, the dataset contains 4,160 static images from 130 subjects (captured in both visible and infrared spectra).

CUFSF dataset: The CUHK Face Sketch FERET Database (CUFSF) [10] consists of 1194 faces from the FERET dataset [47], where each face image has a corresponding sketch drawn by an artist. Due to the exaggerations in the sketches, this dataset poses a

challenge for the *HFR* task. Following [48], we use 250 identities for training the model and reserve the remaining 944 identities for testing. The Rank-1 accuracies are reported for comparison.

CASIA NIR-VIS 2.0 dataset: The CASIA NIR-VIS 2.0 Face Database [46], contains images taken under both the visible spectrum and near-infrared lighting conditions, with 725 distinct individuals. For every person in the dataset, there are 1-22 visible spectrum photos and 5-50 near-infrared (NIR) photos. The given experimental protocols utilize a 10-fold cross-validation method, wherein 360 identities are set aside for training. The evaluation's gallery and probe set comprise 358 distinct individuals. The training and testing sets have entirely separate identities. Experiments are carried out in each fold and the mean and standard deviation of the performance metrics are reported.

4.2 Metrics

We evaluate the models using various performance metrics that are commonly used in previous literature, including Area Under the Curve (AUC), Equal Error Rate (EER), Rank-1 identification rate, and Verification Rate at different false acceptance rates (0.01%, 0.1%, 1%, and 5%).

4.3 Experimental results

The experiments performed in the different datasets and the results are discussed in this section. For comparison, we compared the results of CAIM against the paper baselines reported in [4].

4.3.1 Experiments with Tufts face dataset

The performance of the CAIM method and other state-of-the-art techniques in the VIS-Thermal protocol of the Tufts face dataset is presented in Table 1. This dataset is very challenging due to variations in pose and other factors. The extreme yaw angles present in the dataset cause a decline in the performance of even visible spectrum face recognition systems, along with a similar decline in *HFR* performance. Despite this challenge, the CAIM approach achieves the best verification rate and ranks second in Rank-1 accuracy (73.07%), following DVG-Face [24]. These results demonstrate the effectiveness of the proposed method.

TABLE 1
Experimental results on VIS-Thermal protocol of the Tufts Face dataset.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
LightCNN [49]	29.4	23.0	5.3
DVG [50]	56.1	44.3	17.1
DVG-Face [24]	**75.7**	68.5	36.5
DSU-Iresnet100 [4]	49.7	49.8	28.3
PDT [4]	65.71	69.4	45.5
MAMCO-HFR [22]	-	68.8	-
CAIM (Proposed)	73.07	**76.81**	**46.94**

4.3.2 Experiments with MCXFace dataset

Table 2 presents the average performance across five folds for the VIS-Thermal protocols in the MCXFace dataset. The reported values are the mean of the five folds in the dataset. The baseline model shown corresponds to the performance of the pretrained *Iresnet100* FR model directly on the thermal images. It can be seen that the proposed CAIM approach achieves the best performance compared to other methods with an average Rank-1 accuracy of 87.24 %.

TABLE 2
Performance of the proposed approach in the VIS-Thermal protocol of MCXFace dataset, the Baseline is a pre-trained *Iresnet100* model.

Method	AUC	EER	Rank-1
Baseline	84.45 ± 3.70	22.07 ± 2.81	47.23 ± 3.93
DSU-Iresnet100 [4]	98.12 ± 0.75	6.58 ± 1.35	83.43 ± 5.47
PDT [4]	98.43 ± 0.78	6.52 ± 1.45	84.52 ± 5.36
CAIM (Proposed)	**98.97 ± 0.24**	**5.05 ± 0.91**	**87.24±2.75**

4.3.3 Experiments with Polathermal dataset

We have performed experiments in the thermal to visible recognition scenarios in the Polathermal dataset and the results are presented in Table 3. The table shows the average Rank-1 identification rate in the five protocols of the Polathermal 'thermal to visible protocols' (using the reproducible protocols in [20]). The proposed CAIM approach achieves an average Rank-1 accuracy of 95.00% with a standard deviation of (1.63%), only second to the PDT approach [4].

TABLE 3
Pola Thermal - Average Rank-1 recognition rate

Method	Mean (Std. Dev.)
DPM in [45]	75.31 % (-)
CpNN in [45]	78.72 % (-)
PLS in [45]	53.05% (-)
LBPs + DoG in [8]	36.8% (3.5)
ISV in [51]	23.5% (1.1)
GFK in [52]	34.1% (2.9)
DSU(Best Result) [20]	76.3% (2.1)
DSU-Iresnet100 [4]	88.2% (5.8)
PDT [4]	**97.1% (1.3)**
CAIM (Proposed)	95.00% (1.63)

4.3.4 Experiments with SCFace dataset

We conducted a series of experiments on the SCFace dataset to evaluate the performance of the proposed approach using the visible images protocol. The dataset presents a heterogeneity challenge due to the quality disparity between the gallery (high-resolution mugshots) and probe (low-resolution surveillance camera) images. The results are presented in Table 4 and are based on the evaluation set of the standard protocols. The baseline model employed in this experiment is a pre-trained *Iresnet100* model, while the proposed CAIM model is trained using contrastive training. It can be seen that the performance of the baseline model improves with the proposed approach in most of the cases. In particular, the improvement is more significant in the "far" protocol where the quality of the probe images is very low. The CAIM module helps in adapting the intermediate feature map so that the *HFR* framework is invariant to quality and resolution, leading to improved results compared to the baseline. The proposed method achieves comparable performance to the PDT approach in this dataset.

4.3.5 Experiments with CUFSF dataset

In this section, we present experiments on the challenging task of sketch to photo recognition. We report the Rank-1 accuracies

TABLE 4
Performance of the proposed approach in the SCFace dataset, the
Baseline is a pretrained *Iresnet100* model.

Protocol	Method	AUC	EER	Rank-1	VR@ FAR=0.1%
Close	Baseline	100.0	0.00	**100.0**	100.0
	DSU-Iresnet100 [4]	100.0	0.00	**100.0**	100.0
	PDT [4]	100.0	0.00	**100.0**	100.0
	CAIM (Proposed)	100.0	0.01	**100.0**	100.0
Medium	Baseline	99.81	2.33	98.60	92.09
	DSU-Iresnet100 [4]	99.95	1.39	98.98	93.25
	PDT [4]	99.96	0.93	**99.07**	95.81
	CAIM (Proposed)	99.92	1.86	98.60	94.88
Combined	Baseline	98.59	6.67	91.01	77.67
	DSU-Iresnet100 [4]	98.91	4.96	92.71	80.93
	PDT [4]	99.06	4.50	93.18	82.02
	CAIM (Proposed)	99.58	3.24	**94.57**	84.65
Far	Baseline	96.59	9.37	74.42	49.77
	DSU-Iresnet100 [4]	97.18	8.37	79.53	58.26
	PDT [4]	98.31	6.98	84.19	60.00
	CAIM (Proposed)	98.81	5.09	**86.05**	61.86

obtained with the baseline and other methods in Table 5 using the protocols outlined in [48]. The proposed approach achieves a Rank-1 accuracy of 76.38%, which is the best among the compared methods. However, the absolute accuracy in sketch to photo recognition is low compared to other modalities. The CUFSF dataset contains viewed hand-drawn sketch images [53] that appear holistically similar to the original subjects for humans. Unlike other imaging modalities such as thermal, near-infrared, and SWIR, sketch images may not preserve the discriminative information that a face recognition network seeks, as they contain exaggerations depending on the artist, making them more challenging for *HFR* . Nevertheless, the proposed CAIM approach improves the performance significantly.

TABLE 5
CUFSF: Rank-1 recognition in sketch to photo recognition

Method	Rank-1
Baseline	56.57
IACycleGAN [48]	64.94
DSU-Iresnet100 [4]	67.06
PDT [4]	71.08
CAIM (Proposed)	**76.38**

4.3.6 *Experiments with CASIA-VIS-NIR 2.0 dataset*

We conducted experiments using the CASIA-VIS-NIR 2.0 dataset to demonstrate our proposed method's efficiency in various heterogeneous situations, particularly in VIS-NIR recognition. Observing the baselines, there's a smaller domain gap in this case, with some pre-trained FR models trained in the VIS modality achieving reasonable results. Given this, we employ stricter evaluation thresholds, using VR@FAR=0.1% and VR@FAR=0.01% for comparisons. The dataset contains 10 sub-protocols, and we report the average and standard deviation across these ten folds. The findings, shown in Tab. 6, reveal that our proposed strategy outperforms other state-of-the-art methods. These results showcase the adaptability of our framework across diverse heterogeneous scenarios.

TABLE 6
Experimental results on CASIA NIR-VIS 2.0.

Method	Rank-1	VR@FAR=0.1%	VR@FAR=0.01%
IDNet [54]	87.1±0.9	74.5	-
HFR-CNN [55]	85.9±0.9	78.0	-
Hallucination [56]	89.6±0.9	-	-
TRIVET [57]	95.7±0.5	91.0±1.3	74.5±0.7
W-CNN [58]	98.7±0.3	98.4±0.4	94.3±0.4
PACH [59]	98.9±0.2	98.3±0.2	-
RCN [60]	99.3±0.2	98.7±0.2	-
MC-CNN [61]	99.4±0.1	99.3±0.1	-
DVR [62]	99.7±0.1	99.6±0.3	98.6±0.3
DVG [50]	99.8±0.1	99.8±0.1	98.8±0.2
DVG-Face [24]	99.9±0.1	99.9±0.0	99.2±0.1
PDT [4]	99.95±0.04	99.94±0.03	99.77±0.09
MAMCO-HFR [22]	99.9±0.1	99.8±0.1	-
CAIM (Proposed)	99.96±0.02	99.95±0.02	99.79±0.11

TABLE 7
Performance with different number of CAIM blocks. 1-5 indicates the
CAIM module is inserted in all blocks from first to fifth layers.
Experiment performed in Tufts face dataset.

Layers	AUC	EER	Rank-1	VR(0.1% FAR)	VR(1% FAR)
1	91.28	17.10	49.19	3.34	49.17
1-2	94.91	11.35	64.45	39.70	67.72
1-3	**97.01**	**8.53**	**73.07**	**46.94**	**76.81**
1-4	96.18	9.28	68.76	45.08	72.36
1-5	95.73	10.76	69.30	33.40	71.61

4.4 Ablation Studies

In this subsection, we conduct a series of ablation studies to better understand the efficacy of various components and to assess the generalizability of the CAIM approach.

4.4.1 *Effect of number of CAIM blocks*

To understand the effect of having a different number of CAIM blocks, we performed a set of experiments in the Tufts face dataset by inserting a different number of CAIM blocks in the pre-trained FR network. We start by placing only one CAIM block after the first block of the pre-trained FR layer. Then we increased the number of CAIM blocks from one to 5. The results of this experiment are presented in Table 7. Our analysis reveals that adapting lower layers is effective in minimizing the domain gap, as the high-level facial structure is consistent across various modalities. In this case, adding three CAIM blocks achieved the best performance (this setting is used in all other experiments). Conversely, adapting more layers does not bring significant improvements as they are more task-specific. In our case, the task is face recognition which is the same for both source and target modalities.

The optimal number of layers to adapt can vary depending on the specific modality and architecture, but we have observed that adapting layers "1-3" generally yields satisfactory results across a diverse range of modalities. Consequently, we have consistently applied these settings in all our experiments, though they may not be optimal. Conducting additional experiments to determine the optimal number of layers for each dataset and architecture could potentially enhance performance. To evaluate this, we have evaluated our approach on the CUFSF dataset with a different number of layers, and the results are shown in Table 8. The results indicate a modest improvement in rank-1 accuracy when layers "1-4" are adapted. However, modifying additional layers poses risks of overfitting and increased computational overhead (Table 12).

TABLE 8
Performance with different number of CAIM blocks. 1-5 indicates the CAIM module is inserted in all blocks from the first to fifth layers. The experiment was performed in CUFSF face dataset for iresnet100 model.

Layers	AUC	EER	Rank-1	VR(0.1% FAR)	VR(1% FAR)
1	99.22	4.13	65.89	69.49	89.30
1-2	99.53	3.07	69.28	72.78	90.57
1-3	99.78	2.01	76.38	81.25	95.55
1-4	99.72	2.33	76.69	80.72	95.55
1-5	99.72	2.63	76.17	79.34	94.28

Additionally, we conducted experiments using the ElasticFace model on the Tufts face dataset to determine the optimal number of layers. In this scenario, adapting layers "1-4" proved more effective than just "1-3". This suggests that the number of layers to tune can be further optimized for separate models and datasets. Nonetheless, adapting "1-3" layers provides a reasonable trade-off in terms of performance and computational overhead.

TABLE 9
Performance with different number of CAIM blocks. 1-5 indicates the CAIM module is inserted in all blocks from the first to fifth layers. The experiment was performed in Tufts face dataset for ElasticFace model.

Layers	AUC	EER	Rank-1	VR(0.1% FAR)	VR(1% FAR)
1	87.50	20.41	48.11	25.23	46.57
1-2	93.59	13.54	61.76	31.54	59.18
1-3	95.24	10.39	73.43	50.65	73.65
1-4	96.04	10.20	71.81	56.77	74.03
1-5	94.34	13.36	59.25	29.50	60.30

4.4.2 Effectiveness of components of CAIM block

Further to understand the effectiveness of the conditional operation, we conducted experiments using the **AIM** and Instance Norm (IN) modules in an unconditional manner. These experiments were conducted using the Tufts-face dataset, with the results shown in Table 10. The conditional path in CAIM keeps the original performance on the source modality intact and prevents catastrophic forgetting when adapted to an extra modality. It can be seen that an unconditional integration of the block violates this premise and leads to inferior performance. These results underline both the effectiveness and necessity of the conditional operation.

TABLE 10
Ablation experiments on Tufts Face dataset with unconditional block.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
AIM	6.82	3.71	0.19
IN	36.27	17.44	3.53
CAIM (Proposed)	73.07	**76.81**	**46.94**

4.4.3 Experiment with another FR model

To evaluate the effectiveness of the approach with models trained with different loss functions, we have performed experiments with models trained with ArcFace [63] and ElasticFace [64] loss functions. We use the same *iresnet100* architecture for both of these models. The performance of the models is shown in Table 11. It can be seen that both models perform comparably, showcasing the

effectiveness of the approach. Also, it is to be noted that, despite the original models being trained using different loss functions, the learning phase of the CAIM module is the same as described in the previous section.

TABLE 11
Ablation experiments on Tufts Face dataset with ArcFace and ElasticFace.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
ElasticFace [64] + CAIM	**73.43**	73.65	**50.65**
ArcFace [63] + CAIM	73.07	**76.81**	46.94

4.4.4 Computational Complexity

We have evaluated the computational load of the CAIM approach applied to the iresnet100 architecture, particularly measuring the overall computation in terms of floating point operations (GFLOPS) and the number of parameters (expressed in millions of parameters - MPARAMS). The results presented in Table 12 show that the additional computational load and parameters required by the CAIM approach are minimal. To be more specific, adapting layers "1-3" results in a mere 0.6% increase in the number of parameters and an 8.6% rise in computational requirements, while converting the FR model to an HFR one.

TABLE 12
Computational complexity of the CAIM approach with different number of layers in terms of floating point operations (GFLOPS) and the number of parameters (expressed in millions of parameters - MPARAMS).

	GFLOPS	MPARAMS
iresnet100	2.42	65.15
iresnet100 + CAIM(1)	2.56	65.22
iresnet100 + CAIM(1-2)	2.59	65.30
iresnet100 + CAIM(1-3)	2.63	65.58
iresnet100 + CAIM(1-4)	2.66	66.73
iresnet100 + CAIM(1-5)	2.69	71.32

5 DISCUSSIONS

We introduced a new strategy that adapts feature maps of the target modality to align with the style of visible images, thereby effectively reducing the domain gap between different image modalities. To achieve this, we introduce a novel module called CAIM that can be inserted into a pre-trained FR model, which enables the conversion of a face recognition model to an *HFR* model. Our experimental results demonstrate the effectiveness and robustness of our proposed approach, with state-of-the-art performance achieved in various *HFR* benchmarks. In five out of six datasets, the proposed approach outperforms all other approaches compared. Our method shows superior adaptability in the feature space compared to PDT [4], whose transformations are constrained by the PDT block's receptive field, making our framework more flexible. Our approach can convert an FR model to an HFR model with less than 10% additional compute. The proposed approach can be extended to newer FR architectures, and can also be improved by better training methods.

6 CONCLUSIONS

In this work, we introduce a novel framework for heterogeneous face recognition by considering different imaging modalities as distinct "styles". Our proposed strategy transforms a conventional face recognition (FR) model into a heterogeneous face recognition (HFR) model by aligning the style of the target modality feature maps with that of visible images. To accomplish this, we introduce a novel network module named "CAIM", which can be seamlessly integrated between the frozen layers of a pre-trained FR network. This new CAIM module is trained for HFR in a contrastive learning setup. Our experimental results showcase our method's state-of-the-art performance across several challenging benchmarks. Our approach is versatile and compatible with face recognition models trained using different loss functions. To encourage further research and extensions of our work, we will make the source codes and protocols available publicly.

REFERENCES

[1] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," *Advances in face detection and facial image analysis*, vol. 1, pp. 189–248, 2016.

[2] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 4, pp. 627–639, 2007.

[3] A. George, D. Geissbuhler, and S. Marcel, "A comprehensive evaluation on multi-channel biometric face presentation attack detection," *arXiv preprint arXiv:2202.10286*, 2022.

[4] A. George, A. Mohammadi, and S. Marcel, "Prepended domain transformer: Heterogeneous face recognition without bells and whistles," *IEEE Transactions on Information Forensics and Security*, 2022.

[5] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1410–1422, 2012.

[6] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein cnn: Learning invariant features for nir-vis face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1761–1773, 2018.

[7] A. George and S. Marcel, "Bridging the gap: Heterogeneous face recognition with conditional adaptive instance modulation," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2023, p. 1.

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

[9] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 639–646, 2010.

[10] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *CVPR 2011*. IEEE, 2011, pp. 513–520.

[11] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for nir-vis face recognition," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[12] H. Roy and D. Bhattacharjee, "A novel quaternary pattern of local maximum quotient for heterogeneous face recognition," *Pattern Recognition Letters*, vol. 113, pp. 19–28, 2018.

[13] D. Liu, J. Li, N. Wang, C. Peng, and X. Gao, "Composite components-based face sketch recognition," *Neurocomputing*, vol. 302, pp. 46–54, 2018.

[14] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 188–194, 2015.

[15] D. Lin and X. Tang, "Inter-modality face recognition," in *European conference on computer vision*. Springer, 2006, pp. 13–26.

[16] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li, "Face matching between near infrared and visible light images," in *International Conference on Biometrics*. Springer, 2007, pp. 523–530.

[17] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1123–1128.

[18] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1707–1716, 2012.

[19] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *CVPR 2011*. IEEE, 2011, pp. 593–600.

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

[21] D. Liu, X. Gao, N. Wang, J. Li, and C. Peng, "Coupled attribute learning for heterogeneous face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4699–4712, 2020.

[22] D. Liu, W. Yang, C. Peng, N. Wang, R. Hu, and X. Gao, "Modality-agnostic augmented multi-collaboration representation for semi-supervised heterogenous face recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4647–4656.

[23] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 687–694.

[24] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "Dvg-face: Dual variational generation for heterogeneous face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[25] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 11, pp. 1955–1967, 2008.

[26] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *2005 IEEE Computer Society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 1005–1010.

[27] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *arXiv:1703.10593 [cs]*, Mar. 2017.

[28] H. B. Bae, T. Jeon, Y. Lee, S. Jang, and S. Lee, "Non-visual to visual translation for cross-domain face recognition," *IEEE Access*, vol. 8, pp. 50 452–50 464, 2020.

[29] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu, "Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 100–107.

[30] D. Liu, X. Gao, C. Peng, N. Wang, and J. Li, "Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 10, pp. 5611–5625, 2021.

[31] M. Luo, H. Wu, H. Huang, W. He, and R. He, "Memory-modulated transformer network for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2095–2109, 2022.

[32] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[33] A. George and S. Marcel, "Heterogeneous face recognition using domain invariant units," in *ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[34] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[35] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6924–6932.

[36] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[37] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016.

[38] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[39] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=6xHJ37MVxxp

[40] "Pytorch insightface," Sep 2021. [Online]. Available: https://github.com/nizhib/pytorch-insightface

[41] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel, "Continuously reproducing toolchains in pattern recognition and machine learning experiments," in *International Conference on Machine Learning (ICML)*, Aug. 2017. [Online]. Available: http://publications.idiap.ch/downloads/papers/2017/Anjos_ICML2017-2_2017.pdf

[42] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*, Oct. 2012. [Online]. Available: https://publications.idiap.ch/downloads/papers/2012/Anjos_Bob_ACMMM12.pdf

[43] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

[44] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[45] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan, "A polarimetric thermal database for face recognition research," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 119–126.

[46] S. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 348–353.

[47] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[48] Y. Fang, W. Deng, J. Du, and J. Hu, "Identity-aware cyclegan for face photo-sketch synthesis and recognition," *Pattern Recognition*, vol. 102, p. 107249, 2020.

[49] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[50] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "Dual variational generation for low shot heterogeneous face recognition," in *Advances in Neural Information Processing Systems*, 2019.

[51] T. de Freitas Pereira and S. Marcel, "Heterogeneous face recognition using inter-session variability modelling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 111–118.

[52] A. F. Sequeira, L. Chen, J. Ferryman, P. Wild, F. Alonso-Fernandez, J. Bigun, K. B. Raja, R. Raghavendra, C. Busch, T. de Freitas Pereira *et al.*, "Cross-eyed 2017: Cross-spectral iris/periocular recognition competition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 725–732.

[53] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The facesketchid system: Matching facial composites to mugshots," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2248–2263, 2014.

[54] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[55] S. Saxena and J. Verbeek, "Heterogeneous face recognition with cnns," in *European Conference on Computer Vision*, 2016.

[56] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[57] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in *International Conference on Biometrics*, 2016.

[58] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein cnn: Learning invariant features for nir-vis face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1761–1773, 2018.

[59] B. Duan, C. Fu, Y. Li, X. Song, and R. He, "Pose agnostic cross-spectral hallucination via disentangling independent factors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[60] Z. Deng, X. Peng, and Y. Qiao, "Residual compensation networks for heterogeneous face recognition," in *AAAI Conference on Artificial Intelligence*, 2019.

[61] Z. Deng, X. Peng, Z. Li, and Y. Qiao, "Mutual component convolutional neural networks for heterogeneous face recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 3102–3114, 2019.

[62] X. Wu, H. Huang, V. M. Patel, R. He, and Z. Sun, "Disentangled variational representation for heterogeneous face recognition," in *AAAI Conference on Artificial Intelligence*, 2019.

[63] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[64] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1578–1587.

Anjith George has received his Ph.D. and M-Tech degree from the Department of Electrical Engineering, Indian Institute of Technology (IIT) Kharagpur, India in 2012 and 2018 respectively. After Ph.D, he worked in Samsung Research Institute as a machine learning researcher. Currently, he is a research associate in the biometric security and privacy group at Idiap Research Institute, focusing on developing face recognition and presentation attack detection algorithms. His research interests are real-time signal and image processing, embedded systems, computer vision, machine learning with a special focus on Biometrics.

Sébastien Marcel heads the Biometrics Security and Privacy group at Idiap Research Institute (Switzerland) and conducts research on face recognition, speaker recognition, vein recognition, attack detection (presentation attacks, morphing attacks, deepfakes) and template protection. He received his Ph.D. degree in signal processing from Université de Rennes I in France (2000) at CNET, the research center of France Telecom (now Orange Labs). He is Professor at the University of Lausanne (School of Criminal Justice) and a lecturer at the École Polytechnique Fédérale de Lausanne. He is also the Director of the Swiss Center for Biometrics Research and Testing, which conducts certifications of biometric products.