BRIDGING THE GAP: HETEROGENEOUS FACE RECOGNITION WITH CONDITIONAL ADAPTIVE INSTANCE MODULATION

Anjith George[a]   Sébastien Marcel

Idiap-RR-09-2023

NOVEMBER 2023

---

[a]Idiap Research Institute

Bridging the Gap: Heterogeneous Face Recognition with Conditional Adaptive Instance Modulation

Anjith George and Sébastien Marcel

Idiap Research Institute

Rue Marconi 19, CH - 1920, Martigny, Switzerland

{anjith.george, sebastien.marcel}@idiap.ch

Abstract

Heterogeneous Face Recognition (HFR) aims to match face images across different domains, such as thermal and visible spectra, expanding the applicability of Face Recognition (FR) systems to challenging scenarios. However, the domain gap and limited availability of large-scale datasets in the target domain make training robust and invariant HFR models from scratch difficult. In this work, we treat different modalities as distinct styles and propose a framework to adapt feature maps, bridging the domain gap. We introduce a novel Conditional Adaptive Instance Modulation (CAIM) module that can be integrated into pre-trained FR networks, transforming them into HFR networks. The CAIM block modulates intermediate feature maps, to adapt the style of the target modality effectively bridging the domain gap. Our proposed method allows for end-to-end training with a minimal number of paired samples. We extensively evaluate our approach on multiple challenging benchmarks, demonstrating superior performance compared to state-of-the-art methods. The source code and protocols for reproducing the findings will be made publicly available.

1. Introduction

Face recognition (FR) has become a popular method of access control due to its effectiveness and ease of use. Thanks to convolutional neural networks (CNN), most state of the art FR methods achieve excellent performance in "in the wild" conditions and even "human parity" in face recognition performance [23]. Standard FR systems operate in a homogeneous domain, which means that enrollment and matching are performed using the same modality, typically with face images captured with an RGB camera. However, in some cases, matching in a heterogeneous environment may be advantageous. For example, near-infrared (NIR) cameras, which are common in mobile phones and surveil-
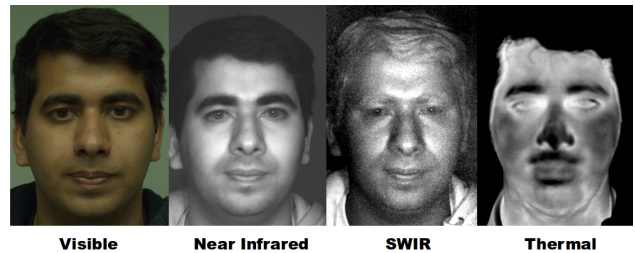

Figure 1. This figure shows the facial images of the same individual acquired using four distinct imaging modalities (Images taken from MCXFace dataset [12]). The task in *HFR* is to facilitate cross-domain matching while overcoming the challenges posed by the domain gap.

lance cameras, provide superior performance regardless of lighting conditions and are robust against presentation attacks [26, 11]. However, training a face recognition system for NIR images will require a large amount of labeled training data which is often not available.

Heterogeneous Face Recognition (*HFR*) systems aim to enable cross-domain matching, allowing enrolled RGB images to be matched with NIR (or any other modality) images and eliminating the need for separate modalities to be enrolled [21]. This approach is particularly valuable in scenarios where visible images are difficult to acquire. For instance, thermal images can be used for recognition regardless of illumination conditions, enabling face recognition during both day and night, and even from long distances. *HFR* systems can handle and match facial images from a variety of sources and modalities, enhancing their applicability across a broad range of situations and applications.

It can be seen that, *HFR* extends the capabilities of Face Recognition (FR) systems in challenging use cases, such as low light or poor visibility, by leveraging the unique properties of different imaging modalities. HFR methods addresses some of the limitations and expands the potential applications of FR systems.

Although highly useful, achieving Heterogeneous Face

Recognition (*HFR*) presents its own challenges. Performing *HFR* using images captured in different modalities is difficult due to the domain gap, which often leads to performance degradation when FR networks trained on visible images are combined with other sensing modalities [16]. Furthermore, training invariant models for both visible and other modalities is challenging due to the limited availability of datasets for new modalities. Collecting large-scale data for these new modalities can also be expensive. Therefore, developing a framework that can be trained with a minimal amount of labeled samples is crucial in this context.

In this work, we utilize a pre-trained face recognition network, trained on a large number of visible spectrum face images, as the base network. We propose to treat different modalities as distinct *styles* and introduce a novel framework to reduce the domain gap. Our method aims to close the gap between visible images and new modalities by modulating the intermediate feature maps of the network. We introduce a new component called Conditional Adaptive Instance Modulation (CAIM) which can be inserted into the intermediate layers of the face recognition network. This learnable component can be trained end-to-end to convert an FR network to an *HFR* network with a minimal amount of training samples.

The main contributions of this work are listed below:

- We formulate the Heterogeneous Face Recognition (*HFR*) problem as a style modulation problem and address the domain gap in *HFR* task using this approach.

- We introduce a new module called the Conditional Adaptive Instance Modulation (CAIM), which can transform a pre-trained FR network into a heterogeneous face recognition network. The framework can be trained with a minimal number of paired samples from the new modality.

- We have evaluated the proposed approach in several standard benchmarks to show the effectiveness of our approach.

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

The remainder of the paper is structured as follows: Section 2 discusses recent literature on Heterogeneous Face Recognition (*HFR*). Details of the CAIM approach are described in Section 3. Extensive evaluation of the CAIM approach, along with comparisons with the state-of-the-art, and discussions, are presented in Section 4 and 5. Lastly, Section 6 presents conclusions and outlines future research directions.

---

[1]`available-upon-acceptance`

2. Related work

The goal of *HFR* methods is to match images obtained from various sensing modalities. However, due to the domain gap, direct image comparison with a face recognition network degrades performance. Hence, it is essential that *HFR* methods should bridge the gap between modalities. In this section, we review recent literature on strategies proposed for bridging this domain gap.

2.1. Invariant feature based methods

Various methods have been proposed for heterogeneous face recognition (*HFR*) to extract invariant features that can match cross-modal images. Liao *et al.* [27] proposed a method based on Difference of Gaussian (DoG) filters and multi-scale block local binary patterns (MB-LBP) features. Klare *et al.* [20] proposed a local feature-based discriminant analysis (LFDA) that used scale-invariant feature transform (SIFT) and multi-scale local binary pattern (MLBP) feature descriptors. Zhang *et al.* [43] introduced the coupled information-theoretic encoding (CITE) method, which maximizes mutual information between the modalities in the quantized feature spaces. Convolutional neural network (CNN) based methods have also been used for *HFR* [15, 16]. Roy *et al.* [33] proposed a novel method called local maximum quotient (LMQ) to extract invariant characteristics in cross-modal facial images.

2.2. Common-space projection methods

The objective of common-space projection methods is to learn a mapping to project different face modalities into a common shared subspace to reduce the domain gap [19, 15]. Lin and Tang [28] developed a common discriminant feature extraction method for extracting features from cross-modal images and projecting them onto a common feature space. Canonical correlation analysis (CCA) was proposed as a way to match face images between NIR and VIS by Yi *et al.* [41]. Regression-based approaches were proposed by Lei *et al.* [24, 25] to learn mapping functions that connect cross-modality domains. Sharma and Jacobs [35] proposed a partial least squares-based method to learn a linear mapping for face image across different modalities so that the mutual covariance is maximized. Klare and Jain [21] proposed a way to represent face images in terms of their similarity to a set of prototype face images. The prototype-based face representation was then projected onto a linear discriminant subspace, which was used to perform the recognition. In [5], the authors proposed a novel approach to the *HFR* task called Domain-Specific Units (DSU). They suggest that high-level features of convolutional neural networks trained on the visible spectrum are domain-independent and can be used to encode images captured in other sensing modalities. The initial layers (DSUs) of a pre-trained FR model are adapted to reduce the domain
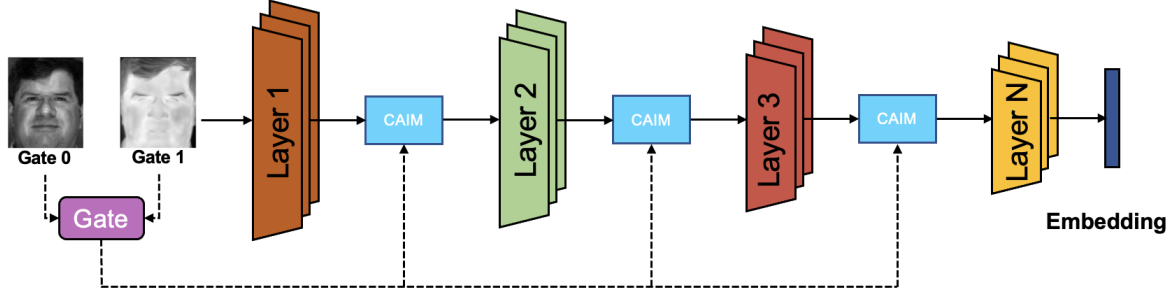
Figure 2. Schematic diagram of the proposed framework: Layer 1 to Layer N represent the frozen blocks of layers from a pretrained Face Recognition (FR) model. The CAIM module is inserted between the initial few blocks.

gap, and the whole pipeline is trained in a contrastive setting. However, the number of layers that need to be adapted is a hyper-parameter that has to be found with an extensive set of experiments.

2.3. Synthesis based methods

Synthesis-based methods for *HFR* [36, 9] attempt to synthesize the source domain images from the target modality, which enables the use of typical face recognition networks for biometric matching. Wang *et al.* [38] proposed a patch-based synthesis approach using Multi-scale Markov Random Fields, while Liu *et al.* [29] used Locally Linear Embedding (LLE) to learn a pixel-level mapping between VIS images and viewed sketches. CycleGAN [45] has been used to transform images from the target domain to the source domain in [4]. Zhang *et al.* [42] proposed a Generative Adversarial Network-based Visible Face Synthesis (GAN-VFS) method to synthesize photo-realistic visible face images from polarimetric images. Many 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 [9], which achieved state-of-the-art results in many challenging *HFR* benchmarks. In [30], authors proposed a Memory-Modulated Transformer Network (MMTN) for heterogeneous face recognition (*HFR*) task. They formulated the *HFR* problem as a synthesis-based approach, specifically, an unsupervised, reference-based "one-to-many" generation task. The MMTN includes a memory module to identify prototypical style patterns and a novel style transformer module for a local fusion of input and reference image styles. Recently, George *et al.* proposed an approach named Prepended Domain Transformers (PDT) [12], which prepends a neural network module to a pretrained FR network to convert it to an *HFR* network. This new block translates the feature representation without explicitly generating the source domain images so that the cross domain embeddings align in the feature space.

2.4. Limitations of current approaches

Recent *HFR* methods in the literature [9, 42] have predominantly utilized synthesis-based methodologies, especially GAN-based synthesis methods due to their ability to generate high-quality images. The use of a pre-trained FR model in synthesis-based *HFR* methods eliminates the need for extensive training data for training a new FR model. However, the synthesis process adds significant computing costs, making it less practical for real-world applications.

We argue that the domain gap between visible images and other modalities can be viewed as different *styles* and the domain gap can be addressed by modulating the feature maps, obviating the compute and memory overhead for synthesizing source modality images.

3. Proposed Method

We follow the notations consistent with recent literature [39, 5, 12] to formally introduce the *HFR* task.

3.1. Formal definition of *HFR*

Let's consider a domain $\mathcal{D}$ consisting of samples $X \in \mathbb{R}^d$ and a marginal distribution $P(X)$ (of dimensionality d). The objective 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 model parameters. During the training phase of an FR system, $P(Y|X, \Theta)$ is generally learned in a supervised manner using a dataset of faces $X = x_1, x_2, ..., x_n$ along with their associated identities $Y = y_1, y_2, ..., y_n$.

Now, let's examine a heterogeneous face recognition (*HFR*) problem. In this scenario, 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.

In general, the goal of the *HFR* problem, $\mathcal{T}^{hfr}$, is to determine a $\hat{\Theta}$ such that $P(Y|X^s, \Theta) = P(Y|X^t, \hat{\Theta})$.

3.2. Proposed Framework

In our approach, we view face images from different modalities as distinct *styles* and attribute the domain gap in the *HFR* problem as a manifestation of different styles. We hypothesize that, by addressing the domain-specific *style*, we can reduce the domain gap. This is achieved by conditionally modulating the intermediate feature maps of a pre-trained face recognition network.

Given the parameters of a pre-trained face recognition (FR) model, denoted as Θ_{FR}, originating from the source domain $\mathcal{D}^s$. Instead of adapting the model's weights, we introduce a new set of network blocks, referred to as CAIM, which essentially modulate the intermediate feature maps. The CAIM module normalizes and adjusts the style of the target modality such that the embeddings of paired samples from both modalities align in the representation space. A schematic diagram of the proposed framework is depicted in Fig. 2. The CAIM block is inserted into the first few blocks, as they are more closely related to the domain. An external gate activates the CAIM module only for the target modality and serves as a pass-through for the reference modality, thus preventing catastrophic forgetting.

Essentially, we can represent the *HFR* task in the following way:

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

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

The only learnable components are θ_{CAIM_i}, which can be trained in a supervised setting. During training, for source images (X^s), the CAIM module acts as a pass-through, allowing the reference embedding to be obtained as they pass only through original network layers (Θ_{FR}). For target domain images (X^t), they pass through both the original network layers (Θ_{FR}) and the CAIM blocks. Contrastive loss [14] is employed as the loss function during the training phase to ensure the embeddings align in the representation space. The contrastive loss is given as:

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

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 is represented by m, and D_W signifies the distance function between the embeddings of the two samples. If the subjects in X_s and X_t have the same identity, the label Y_p is set to 0, and it is set to 1 otherwise. The distance function D_W can be calculated using

the Euclidean distance (or cosine distance) between the representations extracted by the network. The specifics of the CAIM block are elaborated upon in the following subsections.

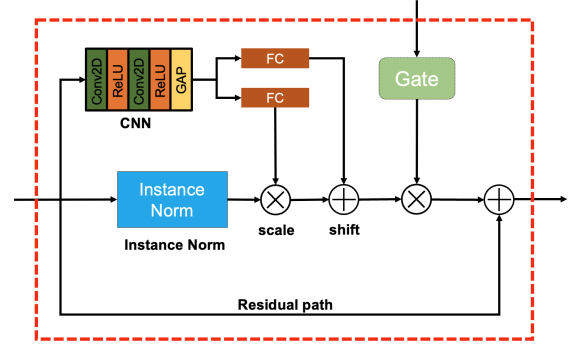## 3.3. Architecture of the CAIM block

Figure 3. Architecture of the Conditional Adaptive Instance Modulation (CAIM) block.

The components of the proposed Conditional Adaptive Instance Modulation module are depicted in Figure 3. The CAIM block is integrated between the frozen layers of a pre-trained facial recognition network. The CAIM block accepts an input feature map and a global gate signal, generating an output feature map of identical dimensions. The initial component of the CAIM block is an Instance Normalization (IN) block without learnable affine parameters, which serves to normalize each feature map at the sample level. Subsequently, a parallel Convolutional Neural Network (CNN) module is employed to derive a shared representation from the un-normalized feature map. The CNN module comprises 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 obtain new scaling and shift parameters, two fully connected layers are introduced. The fully connected layer estimates the parameters utilized for scaling and shifting the normalized feature maps. A residual connection is then incorporated. When the global gate signal is zero, the entire CAIM block functions as an "Identity" function, generating un-altered embeddings as the pre-trained FR model for the reference channel.

3.4. Style Modulation for HFR

In this section, we expand on the utilization of *style* modulation as a means to mitigate the domain gap between visible and additional modalities, starting with the Instance Normalization [37].

3.4.1 Instance normalization

The statistics of feature maps in deep neural networks (DNNs) have been shown to capture the style of images [10]. Ulyanov *et al.* [37] demonstrated that replacing batch normalization layers with Instance Normalization (IN) significantly improves style transfer. The instance normalization layer applies normalization to the feature maps, which can be represented as follows:

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

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

In [18], Huang *et al.* introduced a network module named Adaptive Instance Normalization (AdaIN). 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 argue that Instance Normalization achieves a form of style normalization by normalizing feature statistics, specifically the mean and variance.

The AdaIN module takes a content input x and a style input y, and aligns the channel-wise mean and variance of x to match those of y. AdaIN does not have learnable affine parameters, instead, it adaptively calculates the affine parameters from the style input:

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

Here, 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 [44] have revealed that probabilistically mixing instance-level feature statistics from various source domains can enhance domain generalization. By incorporating diverse styles during training, the model becomes more robust and generalizable across domains. It is important to highlight that, the mixing occurs during the training process of the model incorporating various domains, where the model is trained from scratch. This distinction is crucial, as in the Heterogeneous Face Recognition (HFR) scenario, we begin with a pre-trained face recognition model that has already been trained on the source domain. We extend the idea for *HFR* by conditionally modulating the feature maps instead of mixing them.

3.4.2 Conditional Adaptive Instance Modulation

The AdaIN module is capable of generating or matching images with the style of another image, as discussed above.

Specifically, it enables the creation of images that possess the style of another image. In the *HFR* scenario, the goal is to modulate the style of target modality images to match that of the visible images to ensure the final embeddings align between modalities. This is especially important given that the pretrained face recognition network has been trained on a large dataset of visible spectrum images.

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 this 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.

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

The CAIM block's main component is similar to adaptive instance normalization (AdaIN) as it normalizes and modulates the style of target images. However, instead of utilizing an external image to obtain the style, we estimate the modulation factors from the un-normalized input feature maps using a CNN module. Additionally, we combine this step in a residual fashion while injecting the CAIM block to 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.

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

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

$$\sigma_{\mathbf{f}} = \text{FC}_\sigma(\xi_{\mathbf{f}}) \quad (7)$$

$$\mu_{\mathbf{f}} = \text{FC}_\mu(\xi_{\mathbf{f}}) \quad (8)$$

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

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

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 following:

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

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

3.5. Face Recognition backbone

In the interest of reproducibility, we utilized the publicly accessible pre-trained *Iresnet100* face recognition model provided by Insightface [1]. 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 instances of single-channel input data (e.g., NIR, thermal), the single channel is replicated across all three channels without modifying the network architecture.

3.6. Implementation details

The Conditional Adaptive Instance Modulation (CAIM) block is trained using a contrastive learning paradigm, employing contrastive loss within a Siamese network setting [14]. In all experiments, the margin parameter is set to 2.0. We utilized the Adam Optimizer with a learning rate of 0.0001 and trained the model for 50 epochs using a batch size of 90. The framework was implemented in PyTorch, leveraging the Bob library [2, 3] [3]. The frozen layers of the pre-trained face recognition network are shared between both source and target branches. The CAIM block is integrated between the frozen layers and is only activated for the target modality when the global gate signal is set to one ($gate = 1$). When reference channel images (visible) are processed ($gate = 0$), the entire CAIM block functions as a pass-through via the residual branch. Only the parameters of the CAIM block are updated during training. The experiments are reproducible, and the source code and protocols will be made publicly [4].

4. Experiments

In this section, we present the results of an extensive set of experiments conducted with the CAIM framework. The primary focus of these experiments was to evaluate the performance of the CAIM approach for VIS-Thermal *HFR* across different datasets. Furthermore, we compared the performance of the CAIM approach against other heterogeneous settings such as VIS-Sketch 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 in the evaluations.

Tufts face dataset: The Tufts Face Database [31] is a collection of face images from various modalities for the *HFR* task. 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 [9], 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 [12] 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 [17] 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 the conventional thermal images and follow the five-fold partitions introduced in [5]. 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 [13] 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) [43] consists of 1194 faces from the FERET dataset [32], 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 [7], we use 250 identities for training the model, and reserve the remaining 944 identities for test-

ing. The Rank-1 accuracies are reported for comparison.

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 comparing with state of the art, we compared the results with CAIM against the paper baselines reported in [12].

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 [9]. 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 [40]	29.4	23.0	5.3
DVG [8]	56.1	44.3	17.1
DVG-Face [9]	**75.7**	68.5	36.5
DSU-Iresnet100 [12]	49.7	49.8	28.3
PDT [12]	65.71	69.39	45.45
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 %.

4.3.3 Experiments with Polathermal dataset

We have performed experiments in the thermal to visible recognition scenarios in the Polathermal dataset and the re-

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

Method	AUC	EER	Rank-1
Baseline	84.45 ± 3.70	22.07 ± 2.81	47.23 ± 3.93
DSU-Iresnet100 [12]	98.12 ± 0.75	6.58 ± 1.35	83.43 ± 5.47
PDT [12]	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**

sults 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 [5]). The proposed CAIM approach achieves an average Rank-1 accuracy of 95.00% with a standard deviation of (1.63%), only second to PDT approach [12].

Table 3. Pola Thermal - Average Rank-1 recognition rate

Method	Mean (Std. Dev.)
DPM in [17]	75.31 % (-)
CpNN in [17]	78.72 % (-)
PLS in [17]	53.05% (-)
LBPs + DoG in [27]	36.8% (3.5)
ISV in [6]	23.5% (1.1)
GFK in [34]	34.1% (2.9)
DSU(Best Result) [5]	76.3% (2.1)
DSU-Iresnet100 [12]	88.2% (5.8)
PDT [12]	**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.

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 [12]	100.0	0.00	**100.0**	100.0
	PDT [12]	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 [12]	99.95	1.39	98.98	93.25
	PDT [12]	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 [12]	98.91	4.96	92.71	80.93
	PDT [12]	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 [12]	97.18	8.37	79.53	58.26
	PDT [12]	98.31	6.98	84.19	60.00
	CAIM (Proposed)	98.81	5.09	**86.05**	61.86

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 obtained with the baseline and other methods in Table 5 using the protocols outlined in [7]. 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 [22] 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 [7]	64.94
DSU-Iresnet100 [12]	67.06
PDT [12]	71.08
CAIM (Proposed)	**76.38**

4.4. Ablation Study

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

Table 6. 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

FR layer. Then we increased the number of CAIM blocks from one to 5. The results of this experiment are presented in Table. 6. Our analysis reveals that adapting feature maps in the lower layers of the network is more beneficial, as they tend to be more related to the modality. In this case, adding three CAIM blocks achieved the best performance. 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.

5. Discussions

We address the issue of domain gap in face recognition by treating different domains as different styles. We propose a novel approach that modulates the target modality feature maps to align with the style of visible images, effectively bridging the gap between different 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. By treating the domain gap as different styles, our approach offers a promising solution for addressing the challenges associated with *HFR*. 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 present a novel framework for heterogeneous face recognition by treating different domains as different *styles*. Our proposed approach can convert a face recognition (FR) model to an *HFR* model by modulating the style of target modality feature maps to match that of visible images. To achieve this, we introduce a novel network module called CAIM, which can be inserted between the frozen layers of a pre-trained FR network. This combined module can be trained for *HFR* using contrastive training. Our experimental results demonstrate the state-of-the-art performance of our proposed method in various challenging benchmarks, indicating its effectiveness and robustness.

The source codes and protocols will be made publicly available to facilitate the extension of our work.

References

[1] Pytorch insightface, Sep 2021.

[2] 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.

[3] 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.

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

[5] T. de Freitas Pereira, A. Anjos, and S. Marcel. Heterogeneous face recognition using domain specific units. *IEEE Transactions on Information Forensics and Security*, 14(7):1803–1816, 2018.

[6] 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*, pages 111–118, 2016.

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

[8] 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.

[9] 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.

[10] 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*, pages 2414–2423, 2016.

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

[12] 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.

[13] M. Grgic, K. Delac, and S. Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011.

[14] 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)*, volume 2, pages 1735–1742. IEEE, 2006.

[15] 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.

[16] 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*, 41(7):1761–1773, 2018.

[17] 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*, pages 119–126, 2016.

[18] 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*, pages 1501–1510, 2017.

[19] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):188–194, 2015.

[20] B. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):639–646, 2010.

[21] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1410–1422, 2012.

[22] 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*, 9(12):2248–2263, 2014.

[23] 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*, 1:189–248, 2016.

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

[25] 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*, 7(6):1707–1716, 2012.

[26] 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*, 29(4):627–639, 2007.

[27] 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*, pages 209–218. Springer, 2009.

[28] D. Lin and X. Tang. Inter-modality face recognition. In *European conference on computer vision*, pages 13–26. Springer, 2006.

[29] 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)*, volume 1, pages 1005–1010. IEEE, 2005.

[30] 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*, 17:2095–2109, 2022.

[31] 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*, 42(3):509–520, 2018.

[32] 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*, 16(5):295–306, 1998.

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

[34] 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)*, pages 725–732. IEEE, 2017.

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

[36] X. Tang and X. Wang. Face sketch synthesis and recognition. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 687–694. IEEE, 2003.

[37] 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*, pages 6924–6932, 2017.

[38] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):1955–1967, 2008.

[39] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[40] 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*, 13(11):2884–2896, 2018.

[41] 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*, pages 523–530. Springer, 2007.

[42] 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)*, pages 100–107. IEEE, 2017.

[43] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR 2011*, pages 513–520. IEEE, 2011.

[44] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021.

[45] 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.