Modality Agnostic Heterogeneous Face Recognition with Switch Style Modulators

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) systems aim to enhance the capability of face recognition in challenging cross-modal authentication scenarios. However, the significant domain gap between the source and target modalities poses a considerable challenge for cross-domain matching. Existing literature primarily focuses on developing HFR approaches for specific pairs of face modalities, necessitating the explicit training of models for each source-target combination. In this work, we introduce a novel framework designed to train a modality-agnostic HFR method capable of handling multiple modalities during inference, all without explicit knowledge of the target modality labels. We achieve this by implementing a computationally efficient automatic routing mechanism called Switch Style Modulation Blocks (SSMB) that trains various domain expert modulators which transform the feature maps adaptively reducing the domain gap. Our proposed SSMB can be trained end-to-end and seamlessly integrated into pre-trained face recognition models, transforming them into modality-agnostic HFR models. We have performed extensive evaluations on HFR benchmark datasets to demonstrate its effectiveness. The source code and protocols will be made publicly available.

1. Introduction

Face recognition (FR) technology has become a popular choice for access control due to its efficiency and user-friendliness, often achieving human-level performance [23]. However, FR systems typically operate within a homogeneous domain, using RGB camera-captured facial images for both enrollment and matching. Nevertheless, there are situations where matching in a heterogeneous setting, incorporating different sensing modalities is advantageous. For instance, NIR cameras offer improved performance in varying lighting conditions and spoofing resilience [26, 10], yet developing an NIR-based FR system requires large-

scale annotated data which is often difficult to get. Heterogeneous Face Recognition (*HFR*) systems address this by enabling cross-domain matching without the requirement of enrolling with the target modality, proving valuable in challenging conditions. Heterogeneous face recognition (HFR) refers to the process of recognizing faces across different sensor modalities or image domains. For example, an HFR system can perform biometric matching between a gallery of identities enrolled using RGB images with images captured from another sensor like NIR CCTV cameras, thermal cameras, low-resolution video, and so on. In essence, HFR systems enable cross-domain matching without the requirement of enrollment with each of the domains. Such systems extend FR's utility to scenarios like low-light or long-range recognition using various imaging modalities. However, the domain gap between modalities makes this challenging [17]. Further, the limited availability of paired data makes this even more challenging to train these models.

Most of the existing works in literature try to reduce the domain gap using a pair of source target modality combinations. Several methods have been reported in the literature to address the heterogeneous face recognition task. Invariant Feature-based approaches like Difference of Gaussian (DoG) and scale-invariant feature transform (SIFT) have been proposed to mitigate the domain gap in *HFR* [27, 21]. Common-space projection methods represent another category, focusing on learning mappings that project diverse facial modalities into a shared subspace to bridge the domain gap [20, 16]. Recent advancements in *HFR* methods, such as DVG and generative adversarial network (GAN)-based synthesis, offer high-quality image generation, albeit at increased computational costs, limiting practical utility [9, 45]. A novel approach introduced in [5], termed Domain-Specific Units (DSU), proposes that CNNs' high-level features trained on visible spectrum data can encode images from other modalities, implying their domain-independence. In contrast, Prepended Domain Transformers (PDT) [14] augments a pre-trained FR network with a

dedicated module for the target modality, transforming it into an *HFR* network. In a different approach, outlined in [11], different modalities are treated as distinct styles and cross-modal matching is achieved through the modulation of feature maps in the target modality using conditional adaptive instance modulation (CAIM).

In most of these works, HFR is addressed by using a pair of modalities, a reference modality (often RGB images), and a target modality such as thermal, NIR, SWIR, and so on. Training with a pair of modalities requires one to know what modalities to expect at test time, and the HFR model needs to be trained for that specific pair of modalities. Also, there exist separate computational paths for two modalities, this is problematic in scenarios where this information is not available, for instance in the case of low-resolution and high-resolution images; passing high-resolution images in the low-resolution path may degrade performance. This limitation restricts their scalability as they need to be trained for each pair of modalities, which is impractical when dealing with a wide range of possible modalities. This becomes rather challenging as HFR is already a data-limited problem.

Knowing that there is shared information due to the face structure across modalities it could be possible to develop a model which is modality-agnostic and universal for modalities across the board. In this work, we propose Switch Style Modulation Blocks (SSMB) which introduces modality-agnostic HFR by automatically routing input images, eliminating the need to know the modality during testing. This automatic routing mechanism enhances computational efficiency and broadens the applicability of the model to various face modalities using a single universal model. Moreover, this approach is data-efficient, requiring only a limited number of samples from different modalities for training. By exploiting shared information across face representation in various modalities, the *SSMB* approach addresses key challenges in HFR, making it a versatile and efficient solution for cross-modal face recognition.

The proposed approach is sensor-agnostic and capable of processing input from diverse sources, including CCTV cameras, mobile phones, infrared sensors, and hand-drawn sketches. This versatility is vital in applications like surveillance, where input varies widely. Even in scenarios with low-quality images, such as old databases, sensor-agnostic HFR systems can be useful for recognition tasks. They can also seamlessly integrate with emerging technologies like augmented reality, virtual reality, and robotics, which employ various sensors to capture facial data.

The main contributions of this work can be summarized as follows:

- We introduce a modality-agnostic Heterogeneous Face Recognition (HFR) approach, expanding the applicability of HFR beyond pairs of specific modalities.

- The proposed approach is data and computationally efficient.

- We introduce new protocols to evaluate the performance using the MCXFace dataset.

- We have performed a comprehensive evaluation of the proposed approach across various protocols and datasets demonstrating its effectiveness.

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

2. Related works

Heterogeneous Face Recognition (HFR) methods aim to match faces from images taken with different sensing technologies. However, the domain gap between these modalities can reduce the effectiveness of face recognition systems in directly comparing multi-modal images. Thus, it's essential for HFR approaches to bridge this modality gap. This section reviews recent approaches proposed to tackle the domain gap.

Invariant feature-based methods: In Heterogeneous Face Recognition (HFR), various strategies have been proposed in the literature to extract invariant features across different sensing modalities. Liao et al. [27] presented a method combining the Difference of Gaussian (DoG) filters with multi-scale block Local Binary Patterns (MB-LBP) for invariant feature extraction. Klare et al. [21] employed Local Feature-based Discriminant Analysis (LFDA), incorporating Scale-Invariant Feature Transform (SIFT) and Multi-Scale Local Binary Pattern (MLBP) as descriptors. Zhang et al. [46] introduced Coupled Information-Theoretic Encoding (CITE), aiming to maximize mutual information across modalities in quantized feature spaces. The use of Convolutional Neural Networks (CNNs) has been explored for HFR, demonstrating its effectiveness in this domain [16, 17]. Roy et al. [38] developed a Local Maximum Quotient (LMQ) to identify invariant features in cross-modal facial images. In [31], a method for composite sketch recognition was proposed, using Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradient (HOG) for feature extraction and integrating these at the score level with a linear function to combine facial components.

Common-space projection methods: Common-space projection methods aim to project facial images from diverse modalities into a single subspace, reducing the domain gap [20, 16]. Lin and Tang [28] introduced common discriminant feature extraction for cross-modal image alignment. Yi et al. [44] applied Canonical Correlation Analysis (CCA) to NIR and VIS images. Lei et al.

[24, 25] developed regression-based mappings for reducing the modality gap. Sharma and Jacobs [39] used Partial Least Squares (PLS) for linear mapping across modalities. Klare and Jain [22] projected face images onto a linear discriminant analysis subspace using prototype similarities. De Freitas et al. [5] showed that CNN high-level features are domain-agnostic, utilizing Domain-Specific Units (DSUs) for domain gap minimization. Liu et al. [30] introduced Coupled Attribute Learning for HFR (CAL-HFR) and Coupled Attribute Guided Triplet Loss (CAGTL) for shared space mapping without manual labeling. Recently, Liu et al. [32] proposed a semi-supervised method, Modality-Agnostic Augmented Multi-Collaboration representation for HFR (MAMCO-HFR), leveraging network interactions for discriminative information extraction and introducing a technique for adversarial perturbation-based feature mapping.

Synthesis based methods: Synthesis-based methods in Heterogeneous Face Recognition (HFR) generate synthetic images in the source domain from the target modality, enabling the use of conventional face recognition networks [41, 9]. Wang et al. [42] developed a patch-based synthetic approach using Multi-scale Markov Random Fields. Liu et al. [33] utilized Locally Linear Embedding (LLE) for pixel-wise matching between VIS images and sketches. Cycle-GAN's introduction for unpaired image translation [47] has facilitated transforming target domain images to the source domain [4]. Zhang et al. [45] employed Generative Adversarial Networks (GANs) for generating photo-realistic VIS images from thermal images via GAN-based Visible Face Synthesis (GAN-VFS). The Dual Variational Generation (DVG-Face) framework [9] leverages GANs for VIS image synthesis, showing good performance in *HFR* benchmarks. Liu et al. [29] introduced Heterogeneous Face Interpretable Disentangled Representation (HFIDR) for latent identity information extraction and cross-modality synthesis. Luo et al. [34] introduced a Memory-Modulated Transformer Network (MMTN) which addresses HFR as an unsupervised, reference-based generation problem, combining prototypical style patterns and style blending. Recently, George et al. [14] proposed Prepended Domain Transformers (PDT), a module prepended to a pre-trained FR network for cross-domain feature alignment, eliminating the need for explicit source domain image generation.

Limitations of HFR approaches: Many of the recent HFR methods adopt a synthesis-based approach. However, this requires the synthesis of high-fidelity RGB images from the target modality and is computationally expensive. Further, most of the methods are trained explicitly for a pair of modalities, and they need to be trained for every pair of modalities separately. This makes the problem more challenging as the HFR problem is already data-limited.

3. Proposed method

Most of the approaches proposed in *HFR* specifically focus on cross-domain matching between a pair of modalities such as Visible-Thermal, Visible-Near Infrared, and so on. This requires the *HFR* models to be trained for each pair of modalities explicitly. This also reduces the opportunities to capture generic features and cross-modal consistencies in an already data-scarce domain like HFR. Further, most of the methods available in the literature assume that the label for modality is already known and uses an asymmetric pipeline for processing the data– meaning the dataflow is different for source and target modalities.

In this work, we go beyond the traditional notion of heterogeneous face recognition and design a framework that can handle multiple face imaging modalities at the same time without needing explicit labels for different modalities. We propose to use an efficient mixture of expert-type routing mechanism inside our framework which can automatically infer the path for the specific input modality. The proposed framework could be used for matching between multiple modalities (not just reference and target) without any labels on the modalities. This makes our approach a universal way of doing cross-modality face matching.

3.1. Definition of HFR

In Heterogeneous Face Recognition (*HFR*), we consider a domain $\mathcal{D}$ containing samples $X \in \mathbb{R}^d$ with a marginal distribution $P(X)$ and dimensionality d. A face recognition (FR) system, $\mathcal{T}^{fr}$, aims to associate these samples with labels Y through a model parameterized by Θ, described by the conditional probability $P(Y|X, \Theta)$. During training, this model learns from a dataset of faces $X = x_1, x_2, ..., x_n$ and corresponding identity labels $Y = y_1, y_2, ..., y_n$ via supervised learning.

In *HFR* , we differentiate between a source domain $\mathcal{D}^s = X^s, P(X^s)$ and a target domain $\mathcal{D}^t = X^t, P(X^t)$, both linked by the label space Y. The *HFR* problem involves finding $\hat{\Theta}$ to align conditional probabilities across domains, i.e., $P(Y|X^s, \Theta) = P(Y|X^t, \hat{\Theta})$.

In our framework, we go beyond *HFR* and propose a Modality Agnostic Heterogeneous Face Recognition (MAHFR) which can include multiple modalities of images $\mathcal{D}^t_j = X^t_j, P(X^t_j)$, for different modalities j such as thermal, near infrared and so on.

And we formulate the MAHFR problem as estimating $\hat{\Theta}$ to align conditional probabilities across domains, i.e., $P(Y|X^s, \Theta) = P(Y|X^t_j, \hat{\Theta})$.

3.2. Overall architecture

The architecture of the framework proposed is shown in Fig. 1. We start with a pre-trained face recognition model and insert a trainable module between the frozen layers.
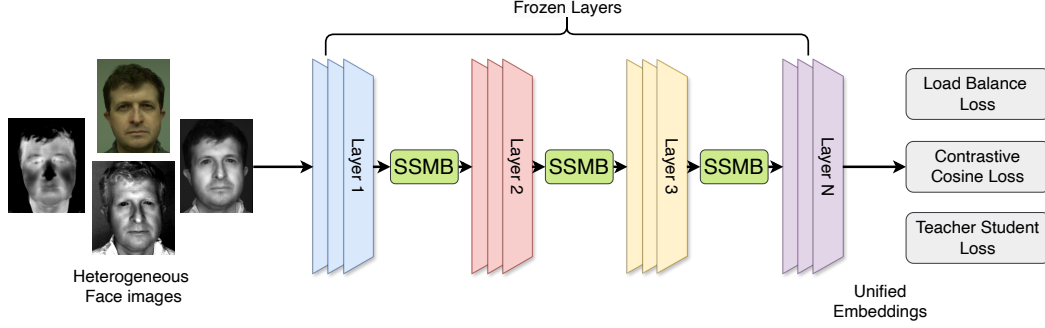
Figure 1. Architecture of the proposed framework, the Switch Style Modulation Blocks adaptively routes the samples (*SSMB*) such that identities for multiple face modalities result in a modality agnostic unified embedding. The embeddings can be used for homogeneous or heterogeneous matching between any pair of modalities.

The objective is that this architecture produces similar embeddings across different face modalities for a specific subject, such that embedding can be used for biometric matching with a cosine similarity metric. The unified embedding obtained is modality agnostic and can be used for both homogeneous and heterogeneous experiments. These *SSMB* modules need to be adaptive in the sense that the operation performed by these blocks should change based on the input sample. The details of the components of this framework are given in the following sub-sections.

3.3. Switch Style Modulation for HFR

The primary challenge in cross-modal face recognition lies in bridging the domain gap between different modalities. Conditional Adaptive Instance Modulation (CAIM), as proposed in the works by George et al. [11, 12], addresses this issue by integrating instance modulation modules within the layers of a pre-trained Face Recognition (FR) network. These modules, placed between the network's frozen layers, are controlled by an external gate, leading to an asymmetric pipeline that necessitates modality labels at the inference stage for selecting the computational path to use.

Building on this, our approach introduces an adaptive mechanism that automatically routes samples from different modalities, eliminating the dependency on an external gating mechanism. This allows for the development of a modality-agnostic representation within a unified framework making HFR more efficient. The architecture of our proposed modulation block, denoted as the *SSMB*, is illustrated in Fig. 2. We simplify the design of the modulators with a single fully connected layer. We use an adaptive routing mechanism based on a Mixture of experts framework for sample-based routing.

Mixture of experts (MoE),[18, 19, 40] enables neural networks to have sample dependant parameters thereby enhancing the abilities of the base model. In the context of large language models (LLMs), MoE [40] routes a token

representation x dynamically to top-k experts, from a pool of $\{E_i(x)\}_{i=1}^{N}$ of N experts. A router module computes logits over the number of experts available with a softmax operation and top-k gate values and experts are selected. The output of the MoE layer can be computed as the weighted combination of the gate values and expert outputs.

$$y = \sum_{i \in \mathcal{T}} p_i(x) E_i(x). \tag{1}$$

where, $\mathcal{T}$ is the set top-k experts selected, and $E_i(x)$ the output of ith expert.

Switch Transformers, as proposed by Fedus et al. [7], introduced a concept of sparsely activated expert models known as switch layers. In this design, for each input sample, the network activates only a subset of its network blocks, thereby maintaining a constant computational requirement. Contrary to other works [40, 37] where the routing is performed to $k > 1$ expert, switch transformer route only to a single expert, referred to as switch routing. The gate value $p_i(x)$ in Equation 1 allows the differentiability of the router module. They also add a load balance loss as an auxiliary loss to balance the routing of samples to different experts.

In this work, rather than establishing distinct models for various segments of the data, we solely utilize modulators within the Mixture-of-Expert (MoE) framework. Specifically, we diverge from the traditional MoE approach and construct a *StyleSwitch* Layer, signifying that modulation is exclusively conducted through routers. We adopt the switch layers where only one expert is activated for a specific sample. This approach allows the model to select sample-dependent modulation parameters at the same time keeps the computational burden small.

For a feature map $\mathbf{F}$, of dimension $C \times H \times W$, we first compute the mean and std deviation of the feature map along channel dimension which are concatenated as the input for the *StyleSwitch* layer.

$$RouterInput = [\boldsymbol{\mu}(\mathbf{F}), \boldsymbol{\sigma}(\mathbf{F})] \tag{2}$$

The router is implemented as a single fully connected layer with N output nodes and a softmax layer, where N is the number of experts.

To estimate the modulation parameters – scale and shift, we use a *StyleExpert* layer, which is implemented as a single fully connected layer. The weight matrix of the *StyleExpert* layer is initialized as an identity matrix so that the training is stable. The *StyleSwitch* layer performs the routing and obtains the final output. Only one expert is activated due to the top-1 routing mechanism. The output of the *StyleSwitch* module has the dimension $2C$, where C is the number of input channels, we split this to get the new modulation parameters.

$$\sigma_{\mathbf{s}}, \mu_{\mathbf{s}} = \mathbf{split}(StyleSwitchOutput) \tag{3}$$

Finally, we scale and shift the instance normalized feature-maps using the output of the *StyleSwitch* layer. We further add a residual connection for stable training. Overall the operation of the *SSMB* can be represented as:

$$\text{SSMB}(\mathbf{F}) = \frac{1}{2}(\sigma_{\mathbf{s}} \left(\frac{\mathbf{F} - \boldsymbol{\mu}(\mathbf{F})}{\boldsymbol{\sigma}(\mathbf{F})} \right) + \mu_{\mathbf{s}} + \mathbf{F}) \tag{4}$$

3.4. Training paradigm

To train this module in an end-to-end fashion we use contrastive loss together with a teacher-student framework similar to the one proposed in [13]. We will denote the network we are training (Fig. 1) as a student network in this context. For training, we need the labels for the reference modality samples (VIS) during training, but explicit labels for non-source modalities are unnecessary as they are grouped as a a single target modality. We create pairs of samples from the source modality and target modality, note that this target modality includes any modality different from the source modality. Pairs with the same identity have a label 1 and pairs with dissimilar identity has a label 0. We use three loss functions in our training.

First, we use a cosine contrastive loss function, denoted by $\mathcal{L}_C$ to align the embeddings of different modalities together, such that the same identity clusters together and different identities are pushed apart:

$$\mathcal{L}_C(e_{S_{s_i}}, e_{S_{t_i}}, y_i) = (1 - y_i) \cdot \max \left(0, \frac{e_{S_{s_i}} \cdot e_{S_{t_i}}}{\|e_{S_{s_i}}\|_2 \|e_{S_{t_i}}\|_2} - m \right)$$
$$+ y_i \cdot \left(1 - \frac{e_{S_{s_i}} \cdot e_{S_{t_i}}}{\|e_{S_{s_i}}\|_2 \|e_{S_{t_i}}\|_2} \right) \tag{5}$$
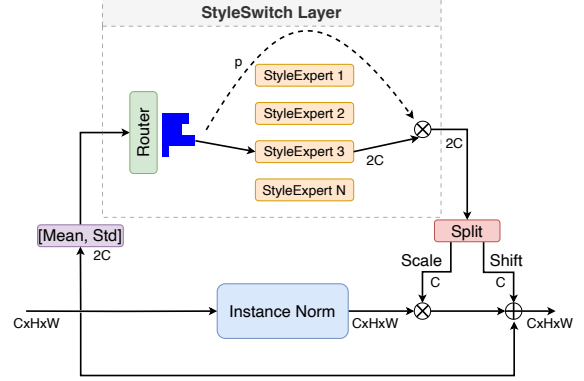

Figure 2. Architecture of the Switch Style modulation block (SSMB). The inputs and outputs of this module are feature-maps of dimension $C \times H \times W$.

Where m denotes the margin, $e_{S_{s_i}}$ and $e_{S_{t_i}}$ denote the embedding for the source modality and target modality from the student network (under training).

Second, due to the limited size of the HFR datasets we use teacher-student supervision using an identity loss computed from the same pretrained face recognition backbone. It is to be noted that this is applied only to the images in the reference modality as the teacher model is to be used for the visible modality [13].

We use the following Teacher Student Identity loss $\mathcal{L}_{TSI}$:

$$\mathcal{L}_{TSI}(e_{T_{s_i}}, e_{S_{s_i}}) = \left(1 - \frac{e_{T_{s_i}} \cdot e_{S_{s_i}}}{\|e_{T_{s_i}}\|_2 \|e_{S_{s_i}}\|_2} \right) \tag{6}$$

where $e_{T_{s_i}}$ denotes the embedding for the source modality image from the pre-trained teacher.

Third, to balance the utilization of different experts we also use a load balancing loss $\mathcal{L}_b$ as reported in [7].

The overall loss function to optimize can be written as:

$$\mathcal{L}(e_{S_{s_i}}, e_{S_{t_i}}, e_{T_{s_i}}, y_i) = (1 - \gamma) \cdot \mathcal{L}_C(e_{S_{s_i}}, e_{S_{t_i}}, y_i) \tag{7}$$
$$+ \gamma \cdot \mathcal{L}_{TSI}(e_{T_{s_i}}, e_{S_{s_i}}) + \alpha \cdot \mathcal{L}_b \tag{8}$$

where γ is a hyper-parameter that influences the relative importance of contrastive loss and teacher-student supervision. Parameter α controls the contribution of load balancing loss in the overall loss.

3.5. Face Recognition backbone

For reproducibility, we utilized the publicly available pre-trained *Iresnet100* face recognition model from Insightface [1], trained on the MS-Celeb-1M dataset featuring over 70,000 identities [2]. This model processes three-channel images of 112×112 pixels. Faces are aligned and cropped to

---

[2]http://trillionpairs.deepglint.com/data

match eye center coordinates with specific reference points before analysis. For single-channel inputs (e.g., NIR or thermal images), the channel is replicated across all three channels without modifying the network structure following [12].

3.6. Implementation details

The proposed approach is trained end-to-end, supervised by three distinct loss functions as described in the previous section. For images in the visible spectrum, we compute the identity loss using both the pre-trained teacher backbone and the student model equipped with *SSMB* layers. Additionally, a load balancing loss and a contrastive cosine loss are incorporated into the optimization process during training. The implementation framework utilizes PyTorch and integrates with the Bob library [2, 3][3]. We employ the Adam Optimizer with an initial learning rate of 0.0001, trained for 50 epochs with a batch size of 48. The margin parameter m is set to zero, and the hyper-parameter γ is set to 0.50. The α parameter is set as 0.01 in all the experiments. We initialize the network layers of our student model from the weights of the pre-trained face recognition network. We use three *SSMB* blocks in the initial layers of the student model. The weights of the *StyleExperts* layer are initialized as an identity matrix so that it produces the same embedding at the beginning of training. Throughout the training phase, only the *SSMB* layers are adapted, while the remainder of the layers remain frozen. During the inference phase, the teacher network is removed, allowing the same student network to process both source and target domain samples without the need for modality labels. This enables the generation of modality-agnostic facial embeddings. Verification and identification of experiments are then performed using cosine distance metric on these unified embeddings.

4. Experiments

Databases used: To evaluate the performance of MAHFR we need to perform experiments on a dataset that contains images of subjects in multiple different modalities. We created new protocols in the MCXFace dataset to facilitate these evaluations.

MCXFace Dataset: The MCXFace Dataset [14] includes images of 51 individuals taken in diverse lighting conditions across three sessions, featuring multiple channels such as RGB, thermal, near-infrared (850 nm), short-wave infrared (1300 nm), depth, stereo depth, and RGB-estimated depth. All channels are synchronized for spatial and temporal consistency. The dataset is divided into five folds per protocol, with subjects randomly divided into training or development groups. Annotations for eye cen-

ters are included for all images. Protocols follow the naming scheme $< SOURCE > - < TARGET > -split < split >$. The dataset is available at [4].

Protocols: The protocols shipped with the dataset include both homogeneous and heterogeneous experimental protocols. For example, $VIS - THERMAL - split4$ denotes a heterogeneous experiment with RGB (VIS) as the source for enrollment and THERMAL as the probe channel, aligning with the third split in the VIS-THERMAL five-fold partitions. We have introduced a new set of protocols, $VIS - UNIVERSAL - split < split >$, to benchmark MAHFR capabilities, where enrollment samples come from the VIS domain and probe samples can be any modality (Thermal, Near Infrared, or Shortwave Infrared). Each protocol features a training set with identities in both SOURCE and TARGET modalities and a development (dev) set with SOURCE images for enrollment and TARGET images for probing. Training and model selection are conducted within the training set, while the dev set scores are used solely for comparison, not training or model selection. For thorough evaluation, experiments should run across all five splits of the protocol, reporting mean values and standard deviations. We additionally compute the modality-wise performance separately to compare different methods.

HFR datasets: In addition to performing experiments in the MCXFace dataset, we also evaluate the performance of the proposed approach in standard HFR benchmarks where only one source target modality pair is present. The Tufts Face Database [35], features face images across various modalities, designed for face recognition tasks. Specifically, for assessing VIS-Thermal face recognition performance, we use the thermal images within this dataset, with 113 identities, following standard protocol in [9]. The SCFace dataset [15], consists of high-quality images for enrollment, and lower-quality samples from surveillance cameras at varying distances as probes. We use the most challenging "far" protocol for reporting the results. The CUHK Face Sketch FERET Database (CUFSF), introduced by Zhang et al. [46], consists of 1194 faces from the FERET dataset [36], each paired with an artist-drawn sketch. These sketches, often exaggerated, present a challenge for the face recognition task. We follow protocols in the previous works for comparison [14].

Baseline methods: We compare the proposed approach against different methods in recent literature specifically the work on domain-specific units DSU [5], and its re-implementation using Iresnet100 [14]. Prepended domain transformer [14] and CAIM [11]. Since these methods were originally designed for a pair of modalities, we maintain the asymmetric nature for a fair comparison– meaning, separate paths for the SOURCE and TARGET modalities. We use the same pre-trained face recognition model and weights in

---

[3]https://www.idiap.ch/software/bob/

[4]https://www.idiap.ch/dataset/mcxface

Table 1. Experimental results on VIS-UNIVERSAL protocol of the MCXFace dataset– aggregated performance.

Modality	AUC	EER	Rank-1	VR@FAR=0.1%	VR@FAR=1%
DSU [14]	95.57±0.80	10.24±0.88	84.21±0.94	67.89±1.05	78.13±1.20
PDT [14]	96.16±1.60	9.60±2.07	80.90±2.49	64.63±5.87	76.30±2.49
CAIM [11]	99.45±0.12	3.67±0.33	90.92±1.30	79.64±2.46	91.58±0.68
SSMB	**99.70±0.08**	**2.59±0.28**	**92.80±0.71**	**84.04±1.71**	**94.50±1.44**

Table 2. Experimental results on VIS-UNIVERSAL protocol of the MCXFace dataset–modality-wise performance breakdown.

Model	Modality	AUC	EER	Rank-1	VR@FAR=0.1%	VR@FAR=1%
DSU [14]	NIR	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
PDT [14]	NIR	99.89±0.15	0.98±1.34	97.72±3.58	91.23±12.48	96.48±4.87
CAIM [11]	NIR	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	**100.00±0.00**
SSMB	NIR	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	**100.00±0.00**
DSU [14]	SWIR	100.00±0.00	0.21±0.10	99.80±0.20	99.38±0.47	99.94±0.13
PDT [14]	SWIR	99.35±0.89	2.57±3.31	90.61±12.82	83.25±22.46	91.66±11.42
CAIM [11]	SWIR	99.99±0.02	0.27±0.25	99.60±0.48	98.83±1.40	99.89±0.25
SSMB	SWIR	100.00±0.00	0.18±0.11	99.86±0.27	99.67±0.26	**100.00±0.00**
DSU [14]	THERMAL	90.05±2.23	17.86±1.97	55.12±3.21	27.92±6.15	51.13±5.48
PDT [14]	THERMAL	91.38±3.80	16.07±4.14	56.55±9.41	30.33±10.34	52.74±11.01
CAIM [11]	THERMAL	98.11±0.39	6.95±0.60	74.51±3.45	38.86±9.70	70.82±4.16
SSMB	THERMAL	98.89±0.27	5.08±0.79	79.41±2.28	43.74±6.51	**78.93±6.67**

all these models and the proposed model.

Metrics: We evaluate the models based on several performance metrics frequently used in recent literature, such as Area Under the Curve (AUC), Equal Error Rate (EER), Rank-1 identification rate, and Verification Rate at specific false acceptance rates (0.01%, 0.1%, 1%, and 5%).

4.1. Experimental results

In this section, we compare the effectiveness of the proposed approach for MAHFR. We further perform evaluations with standard *HFR* datasets to show the effectiveness of the proposed approach.

MAHFR performance on MCXFace dataset: We have performed experiments on the baseline methods and the proposed approach on the MCXFace dataset on the newly introduced $VIS-UNIVERSAL$ protocols. Here the aggregate performance over all the modalities is computed. The results of these protocols are shown in Tab. 1. It can be seen that the proposed approach outperforms other methods by a large margin. For three different probe modalities combined (near-infrared, thermal, and shortwave infrared), the proposed approach achieves an average verification rate of 94.50 %. To analyze the performance further, we further

Table 3. Experimental results on VIS-UNIVERSAL protocol of the MCXFace dataset.

Experts (N)	AUC	EER	Rank-1
1	99.66±0.13	2.85±0.50	91.89±1.17
2	99.65±0.05	2.87±0.32	91.54±0.51
3	99.64±0.10	2.82±0.53	91.54±1.43
4	**99.70±0.08**	**2.59±0.28**	**92.80±0.71**
5	99.68±0.07	2.75±0.41	92.03±1.74

separate the performance of the models in different modalities. The results are shown in Tab. 2. From this table, it can be seen that out of the modalities present, thermal modality is the most challenging one. The proposed approach significantly improves the performance in the thermal channel achieving an average verification accuracy of 78.9% at an FAR (1%). The proposed method achieves 100 % verification accuracy at FAR 1% operating point for both NIR and SWIR modalities. Overall, the proposed approach significantly boosts the unified face recognition performance in the MCXFace dataset.

Effect of number of experts: We have performed a set of experiments to evaluate the change in performance with the number of experts used. The results are shown in Tab. 3, it can be seen that we achieve the best performance when the number of experts is four. Increasing the number of experts does not improve the performance further. It is to be noted that the experts in our framework is a single fully connected layer and the additional computational load required for more experts is minimal.

Experiments with standard HFR datasets: We have performed an extensive evaluation using three standard Heterogeneous Face Recognition (HFR) datasets to benchmark our method against previously proposed approaches in the literature. Specifically, the objective of these evaluations was to compare the performance of our proposed method in scenarios where only a single pair of source and target modalities are present. We have conducted experiments using both one and two experts and have reported the performance for both configurations. The Tufts Face dataset evaluates the performance in the visible-thermal heterogeneous scenario. As seen in the results presented in Ta-

ble 4, our method outperformed other methods in terms of the Verification rate (80.33% at 1% FAR) and achieved the best Rank-1 accuracy with two experts. SCFace dataset addresses heterogeneity in terms of image quality, we specifically focus on the 'far' protocol which has the lowest quality images. The results in Table 5, show that our method attained the highest Rank-1 accuracy at 87.73% and the second-best Equal Error Rate (EER) of 5.91%. Here the best results are obtained with a single expert, this could be because the source and target modality are RGB and the difference is only in the image quality. CUFSF dataset evaluates the performance of heterogeneous face recognition in visible-to-sketch images, and from the results in Tab. 6 it can be seen that the proposed approach achieves s Rank-1 accuracy of 81.67% using two experts significantly outperforming state-of-the-art methods. Overall, the proposed method not only achieves competitive performance but often surpasses existing SOTA methods in the HFR dataset protocols, highlighting its effectiveness.

Table 4. Experimental results on VIS-Thermal protocol of the Tufts Face dataset.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
LightCNN [43]	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 [14]	49.7	49.8	28.3
PDT [14]	65.71	69.39	45.45
CAIM [11]	73.07	76.81	46.94
SSMB (N=1)	75.04	78.29	53.99
SSMB (N=2)	**78.46**	**80.33**	**54.55**

Table 5. Performance of the proposed approach in the SCFace dataset.

Protocol	Method	AUC	EER	Rank-1
Far	DSU-Iresnet100 [14]	97.18	8.37	79.53
	PDT [14]	98.31	6.98	84.19
	CAIM [11]	**98.81**	**5.09**	86.05
	SSMB (N=1)	98.77	5.91	**87.73**
	SSMB (N=2)	98.67	6.36	86.82

5. Discussions

The proposed framework introduces a new way of performing cross-domain face recognition in a modality-agnostic manner. The model trained using our framework can produce embeddings for any modality used in the training. The embeddings obtained share the same embedding space and can be used for heterogeneous or homogeneous matching. To enable this, we utilize a training paradigm that leverages supervision from a teacher network – enabling stable training without overfitting. The addition of multiple modalities at the training time with the adaptive routing enables the network to dynamically adapt the pa-

rameters depending on the input feature maps. It is important to note that specific modality labels are not required at inference time, meaning any face image can be used as the input to compute the representation of face embedding, and the network routes the data automatically to produce the embedding for the specific modality to maximize the matching performance. This could be useful in scenarios where specific source target modality pairs do not need to be trained separately and greatly simplifies the HFR process. It is worth noting that the *StyleExperts* in our framework is lightweight, and the switch routing mechanism can be optimized further to reduce computation at inference time.

Table 6. CUFSF: Rank-1 recognition rate in sketch to photo recognition

Method	Rank-1
IACycleGAN [6]	64.94
DSU-Iresnet100 [14]	67.06
PDT [14]	71.08
CAIM [11]	76.38
SSMB (N=1)	81.14
SSMB (N=2)	**81.67**

6. Conclusion

In this work, we introduce a novel modality-agnostic framework capable of performing cross-domain face recognition without the need for explicit modality labels. This eliminates the need for training separate models for each source-target pairing, allowing for the adoption of a single, unified model. We achieve this by introducing trainable Switch Style Modulation Blocks (SSMB) to dynamically route input feature maps, effectively reducing the domain gap with minimal computational overhead, thereby making our solution both efficient and scalable. Our proposed approach thus facilitates mapping various face modalities into a shared embedding space. Experiments on the MCXFace dataset and standard HFR benchmarks demonstrate the effectiveness of our method. The source code and protocols will be made available publicly.

Acknowledgements

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

[7] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

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

[11] A. George and S. Marcel. Bridging the Gap: Heterogeneous face recognition with conditional adaptive instance modulation. In *2023 International Joint Conference on Biometrics (IJCB)*. IEEE, 2023.

[12] A. George and S. Marcel. From modalities to styles: Rethinking the domain gap in heterogeneous face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[13] A. George and S. Marcel. Heterogeneous face recognition using domain invariant units. ICASSP, 2024.

[14] 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.

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

[16] 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.

[17] 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.

[18] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[19] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[20] 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.

[21] 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.

[22] 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.

[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] 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*, 33(10):5611–5625, 2021.

[30] 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*, 31(11):4699–4712, 2020.

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

[32] 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*, pages 4647–4656, 2023.

[33] 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.

[34] 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*, 2022.

[35] 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.

[36] 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.

[37] P. Ramachandran and Q. V. Le. Diversity and depth in per-example routing models. In *International Conference on Learning Representations*, 2018.

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

[39] 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.

[40] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

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

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

[43] 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.

[44] 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.

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

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

[47] 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.