Heterogeneous Face Recognition with Prepended Domain Transformers

Anjith George[1*] and Sebastien Marcel[1]

[1*] Idiap Research Institute, Rue Marconi 19, Martigny, 1920, Valais, Switzerland.

*Corresponding author(s). E-mail(s): anjith.george@idiap.ch; Contributing authors: sebastien.marcel@idiap.ch;

Abstract

Face recognition (FR) has become a very popular method for biometric authentication thanks to its non-contact nature and high accuracy. State-of-the-art face recognition systems are obtaining human parity even in unconstrained scenarios, thanks to the deep neural network architectures and the large datasets available for training them. However, there are several other types of face imaging modalities such as infrared, thermal, depth and so on which can boost the performance of FR systems even further. The main challenge in using these new modalities is the lack of availability of large-scale labeled datasets to train FR models. Heterogeneous face recognition provides a solution to this issue by leveraging the large sets of training data available for visible spectrum data. Heterogeneous Face Recognition (HFR) involves matching facial images from different domains, such as thermal to visible images (VIS), sketches to visible images, and near-infrared to visible images. This process is especially beneficial for aligning visible spectrum images with those from other modalities. However, HFR poses significant challenges due to the domain gap between the source and target images, compounded by the lack of large-scale, paired heterogeneous face image datasets for training HFR models. In this chapter, we introduce a lightweight and effective method for cross-modality face image matching. The core idea in our approach is to integrate a neural network component, known as the prepended domain transformer (PDT), at the beginning of an existing face recognition (FR) model to bridge the domain gap. By retraining this prepended module with a small number of paired samples in a contrastive learning framework, we achieved state-of-the-art results in various HFR benchmarks. The PDT blocks are versatile and can be retrained for different source-target combinations using our framework. This approach is compatible with any pre-trained FR model due to its architecture-agnostic nature. Additionally, its modular design allows for training with a limited set of paired samples, making it easier for integration and deployment. The source code and protocols for reproducing the results are publicly available.

Keywords: Heterogeneous Face Recognition, Convolutional Neural Network, Biometrics, Face Recognition, Cross-Modal Face Recognition.

1 Introduction

Face recognition (FR) offers a convenient method for access control. The majority of advanced FR methods achieve remarkable results in real-world conditions and are even comparable to human-level accuracy in recognizing faces, largely due to the use of deep neural networks [1]. Typically, FR systems operate within a homogeneous domain, meaning both the enrollment and matching processes use the same type of data, usually images captured by an RGB camera in the visible light spectrum. However, there are instances where performing matching in a heterogeneous scenario can be beneficial [2–5]. For instance, near-infrared (NIR) cameras, which are widely found in smartphones and surveillance systems, demonstrate superior performance in various lighting conditions [6, 7]. Additionally, NIR imagery is highly effective in resisting presentation attacks [7, 8]. In contrast, an FR system limited to a homogeneous setting would necessitate the use of enrollment samples captured with the same NIR camera as the probe samples making it difficult to integrate in practical scenarios.

Heterogeneous face recognition (HFR) systems are designed to address the issue of cross-domain matching. In such systems, for instance, enrolled RGB images can be matched with near-infrared (NIR) images, removing the requirement for enrolling in different modalities [9]. This flexibility allows the use of other imaging modalities, such as thermal [10, 11] and shortwave-infrared [12–14], without necessitating enrollment samples in those specific modalities. Figure 1 illustrates examples of the same individual captured using various sensing technologies. Thermal imaging, which doesn't rely on active illumination, is effective in both daytime and nighttime conditions. Overall, the HFR approach is highly advantageous in practical applications and offers a way to enhance face recognition capabilities in challenging and uncontrolled settings [15].

While HFR (Heterogeneous Face Recognition) offers significant benefits compared to FR, it faces several challenges. A primary issue is the substantial difference between images captured by various sensors, known as the domain gap. Performance of networks trained on RGB images often deteriorates when applied to images from different sensing modalities [16]. Additionally, there is a

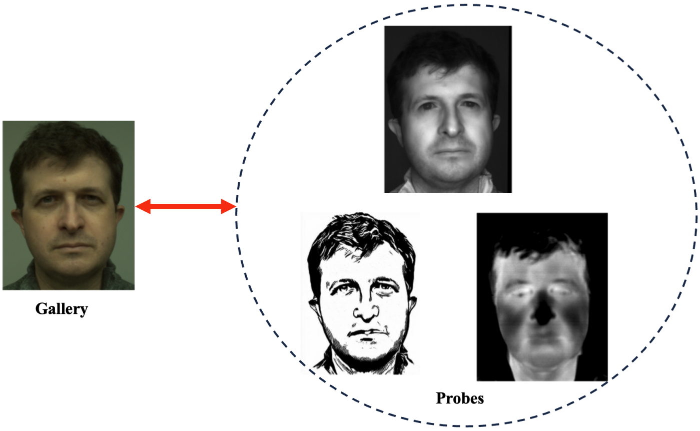

Fig. 1 This figure shows the face of the same individual captured through different modalities such as RGB image, sketch, thermal, and near-infrared. The objective of HFR (Heterogeneous Face Recognition) is to perform cross-modal face recognition by using RGB images as a reference and comparing them with probe images from these other modalities.

scarcity of large-scale heterogeneous datasets suitable for training HFR models [17]. With a limited amount of paired data, a promising approach is to utilize pre-trained FR (Face Recognition) models that have been developed using extensive facial recognition datasets. Moreover, given the high cost and limited availability of heterogeneous data, it's important to develop HFR approaches that are effective even with a small amount of paired data.

The main challenge in Heterogeneous Face Recognition (HFR) is bridging the domain gap between source and target modalities. Most of the HFR approaches found in literature generally fall into three main categories based on their strategies to mitigate this gap. 1) Common-Space Projection Methods [18, 19]: These techniques focus on mapping different facial modalities into a unified, shared subspace, thereby reducing the domain gap. 2) Invariant Feature-Based Methods [9, 16]: The goal of these methods is to extract features from face images that are consistent across different modalities, facilitating the matching of heterogeneous facial images. 3) Synthesis-Based Methods [20, 21]: These approaches involve generating images in the source domain (often visible spectrum images) from the target modality. After this synthesis stage, standard pre-trained Face Recognition models can be applied for performing the matching.

Most existing datasets for heterogeneous face recognition (HFR) are limited in size, posing challenges for training HFR models from scratch. Hence, it is more advantageous to utilize pre-trained face recognition (FR) models that have been developed using extensive, large-scale face datasets. In this chapter, we introduce a surprisingly simple method for Heterogeneous Face Recognition (HFR). Our strategy leverages a pre-trained Face Recognition (FR) model as a key component in our framework. Our approach, named PDT , is independent of the specific architecture of the FR model, offering flexibility in various deployment scenarios. We achieve this by attaching a novel network module, referred to as Prepended Domain Transformer (PDT), in front of a pre-trained FR module. This module is specifically designed to adapt target domain images [22] to match source domain images. The PDT is the only component with learnable parameters in our framework, and it is efficient in terms of parameters, achieving state-of-the-art results even with a limited number of paired samples. The proposed approach is practical for deployment, as it simply involves adding a new module to an existing FR system to transform it into an HFR system. The minimal parameter and computational overhead introduced by our framework make it suitable for real-time applications. The approach is versatile and can be easily retrained for different pairs of heterogeneous modalities. Through extensive evaluation, we demonstrate that this simple enhancement consistently achieves state-of-the-art performance across several challenging HFR datasets. We have deliberately kept a simple design for the framework to not only showcase its effectiveness but also to facilitate future enhancements.

The key contributions of the PDT framework are outlined as follows:

- We introduce a novel heterogeneous face recognition framework that integrates a pre-trained FR model with a Prepended Domain Transformer block. This method is adaptable to various architectures and is suitable for a wide variety of heterogeneous face recognition scenarios. This approach has low computational demands and a minimal number of learnable parameters, facilitating ease of implementation and deployment.
- A comprehensive set of experiments on public HFR datasets is conducted to demonstrate the efficacy of the PDT approach across diverse heterogeneous face recognition scenarios.
- We also present the multi-channel heterogeneous face recognition (MCXFace) dataset, consisting of both homogeneous and heterogeneous protocols with modalities such as color, thermal, depth, stereo, infrared, short-wave infrared, and synthesized 3D maps. We also provide standard protocols and baselines for the heterogeneous scenarios of this dataset.

Additionally, we have made the source codes publicly available to encourage further development and extension of this work [1].

The structure of the chapter is as follows: Section 2 reviews recent literature on Heterogeneous Face Recognition. The details of the PDT approach are elaborated in Section 3. Comprehensive evaluations of the PDT method, including comparisons with state-of-the-art techniques and in-depth discussions, are presented in Sections 4 and 5. Finally, conclusions and potential future research directions are discussed in Section 6.

2 Related work

The objective of Heterogeneous Face Recognition (HFR) involves cross-modal matching of face images captured through various sensing modalities. The main

[1] https://gitlab.idiap.ch/bob/bob.paper.tifs2022_hfr_prepended_domain_transformer

difficulty in HFR arises from the "domain gap", where the same individual's image significantly differs across different modalities. This variation increases the intra-class variance, making the matching process difficult. Directly comparing multi-modal images leads to reduced performance. Several strategies have been proposed in the prevailing literature to address this domain gap.

Table 1 shows some recent *HFR* approaches reported in the literature.

Table 1 Literature table with recent *HFR* approaches

Reference	Method	Dataset	Domain	Accuracy	Open-source
DSU [23]	Domain specific units	Polathermal, CUFS, NIVL, CASIA NIR-VIS	Thermal, Sketch, Infrared	Low	✓
Di *et al.* [24]	Synthesis	ARL-polarimetric	Thermal	Low	✗
Fondje *et al.* [25]	Identity and invariance losses	Polathermal	Thermal	Low	✓
GANVFS [26]	Visible Face Synthesis	Polarimetric thermal	Thermal	Low	✗
DVG-Face [21]	Dual-generation	CASIA NIR-VIS 2.0, BUAA-VisNir, IIIT-D Sketch Viewed, Tufts Face	Near-infrared, Thermal, Sketch	High	✓
PDT	Prepended Domain Transformer	Polathermal, SCFace, ARL-VTF, Tufts face, MCXFace, CASIA NIR-VIS 2.0	Near-infrared, Thermal, Low-resolution, Shortwave infrared	High	✓

2.1 Common-space projection approaches

Common-space projection methods [18, 19], have the goal of learning a mapping to project different facial representations into a shared subspace with the aim to mitigate the differences across various domains. Lin and Tang introduced a common discriminant feature extraction method for cross-modal image features, projecting them into a common feature space [27]. Yi *et al.* proposed Canonical Correlation Analysis (CCA) as a means to align face images from Near-Infrared (NIR) and visible (VIS) domains [28]. Authors in [29, 30] explored regression-based approaches to establish mapping functions connecting cross-modality domains and common spaces. Sharma and Jacobs presented a method based on Partial Least Squares (PLS) to learn a linear mapping for face images across different modalities, maximizing mutual covariance [31]. Klare and Jain introduced a prototype-based face representation approach, projecting faces into a linear discriminant subspace for recognition [9]. In [23], authors proposed the Domain-Specific Units (DSU) approach to adapt convolutional neural network low-level features for various sensing modalities, thereby reducing the domain gap [23]. However, the hyper-parameter for determining the number of layers to adapt requires extensive experimentation, and adaptation is needed for different pre-trained FR model architectures. Recently, Cheema *et al.* presented a Cross-Modality Discriminator Network (CMDN) for Heterogeneous Face Recognition (HFR) [32]. CMDN follows a standard ResNet50 model with a Squeeze-and-Excitation (SENet-50) module [33]. The CMDN employs a Deep Relational Discriminator (DRD) module, which is essentially a Multi-Layer Perceptron (MLP) supervised by binary cross-entropy loss (BCE), to facilitate cross-domain matching. CMDN's learning is guided by the unit class loss, a combination of triplet loss and a modified version that incorporates class means. CMDN is initialized from a pre-trained face recognition backbone trained on the VGGFace2 dataset [34] and then undergoes further training on the Visible-Thermal face dataset (IRIS face dataset) [35]. For the HFR task, CMDN is fine-tuned with a variant of triplet loss (Unit class loss). The embeddings from two modalities (gallery and probe) are concatenated for positive and negative pairs. An MLP model (DRD module) is trained on top of these concatenated embeddings using BCE loss. Scoring is performed with probe-gallery pairs using three different strategies: 1) utilizing embeddings and cosine loss, 2) utilizing the output of the DRD module, and 3) combining scores from strategies 1 and 2. Notably, the evaluation strategy employs pairs of samples rather than comparing probes against the gallery, and it was reported that the fusion model achieved superior results.

2.2 Invariant feature-based approaches

Invariant feature-based techniques are designed to extract features that remain consistent across different modalities, facilitating the matching of heterogeneous face images. Liao *et al.* [36] proposed using the Difference of Gaussian (DoG) filters to emphasize image structure. Subsequently, they employed multi-scale block local binary patterns (MB-LBP) [37] as features and developed a subspace-based face recognition system that jointly incorporates samples from both source and target domains. Klare *et al.* [38] proposed a local feature-based discriminant analysis (LFDA) approach for Heterogeneous Face Recognition. Their method involved extracting scale-invariant feature transform (SIFT) [39] and multi-scale local binary pattern (MLBP) [40] descriptors from patch units of sketch and visible (VIS) images to address the HFR task. Authors in [41] proposed a coupled information-theoretic encoding (CITE) extraction method aimed at maximizing mutual information between heterogeneous modalities within quantized feature spaces. Roy *et al.* [42] introduced the local maximum quotient (LMQ) method to extract invariant features for HFR. Recently, several studies have explored the use of convolutional neural network (CNN)-based methods to extract invariant features for Heterogeneous Face Recognition (HFR) [16, 19].

2.3 Synthesis based approaches

Synthesis-based approaches for Heterogeneous Face Recognition (HFR) [20, 21] focus on generating the source domain (typically visible or VIS) from the target modality, enabling subsequent biometric matching using face recognition networks trained for homogeneous settings. In a patch-based synthesis approach proposed by Wang *et al.* [43], Multi-scale Markov Random Fields were employed to synthesize visible images from sketches and vice versa. Various face recognition methods, including Eigenfaces, Fisherfaces, and dual space LDA, among others, were evaluated for the HFR task in their approach. Liu *et* *al.* [44] utilized Locally Linear Embedding (LLE) to learn pixel-level mappings between VIS images and viewed sketches. In [45] CycleGAN [46] was used to transform target domain images into the source domain. During CycleGAN training, a Siamese network with a contrastive loss was integrated to preserve face identity. Pre-processing with methods from [47] was also employed to further reduce the domain gap. Zhang *et al.* [26] introduced a Generative Adversarial Network-based Visible Face Synthesis (GAN-VFS) method, synthesizing photo-realistic visible face images from polarimetric images. Their approach incorporated identity loss and perceptual loss during training and utilized a VGG network to extract embeddings. The method was evaluated on the Polathermal dataset, achieving an average Equal Error Rate of 34.58%. With advancements in GAN training and architectures, recent approaches have employed GANs to synthesize VIS images from other modalities. For instance, Fu *et al.* [21] introduced the Dual Variational Generation (DVG-Face) framework, treating HFR as a dual generation problem. They designed a dual generator to learn joint distributions of heterogeneous pairs, addressing data scarcity in HFR. A pairwise identity-preserving loss ensured identity consistency, and the generated images were used to train the HFR network in a contrastive setting, achieving state-of-the-art results in various HFR benchmarks. George *et al.* [48] treated different modalities as distinct styles and proposed a method to align them using a lightweight conditional adaptive instance modulation (CAIM) module trained in a contrastive fashion.

2.4 Limitations and challenges with existing approaches

The majority of recent methods for HFR described in the literature, such as those presented in [21, 26], predominantly rely on synthesis-based approaches. One reason for the popularity of synthesis methods in HFR is the widespread adoption of Generative Adversarial Network (GAN)–based techniques. GANs have gained prominence due to their capacity to produce high-quality images, coupled with significant advancements in their training methodologies, as demonstrated in [49]. Furthermore, synthesis-based HFR approaches can take advantage of pre-trained Face Recognition (FR) models, obviating the need for extensive retraining with large datasets.

However, this approach has limitations when applied to practical scenarios. Synthesis-based HFR necessitates the initial generation of an RGB image from the target modality image before passing it through an FR model for embedding extraction during inference. This synthesis step adds a substantial computational burden, potentially limiting its practical deployment. Moreover, synthesis-based algorithms, primarily designed for generating high-quality RGB images that satisfy perceptual quality and diversity, may not necessarily preserve identity-related information when the images are used in the context of HFR tasks. In other words, it is crucial to preserve discriminative features that align with those of the source class rather than focusing solely on

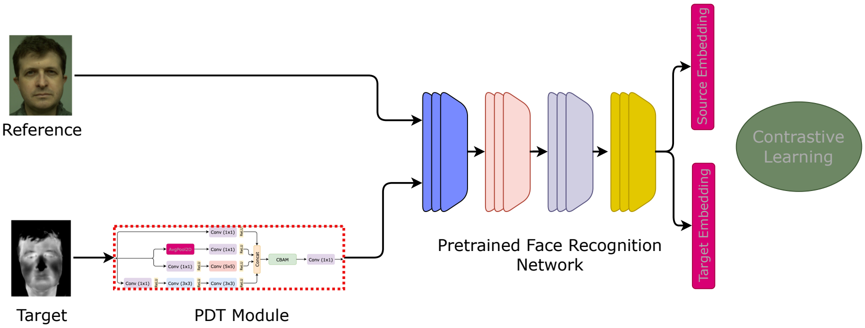

Fig. 2 Schematic diagram of the PDT framework. Images from the target domain undergo transformation through the introduced Prepended Domain Transformer (PDT) block. The PDT block, which is attached at the beginning of the pre-trained Face Recognition (FR) model, is trained in a Siamese setting using contrastive loss.

generating high-quality images. This becomes particularly critical when dealing with a limited amount of paired training data, as solving the generation problem under such constraints can be considerably more challenging.

3 Proposed Method

We follow the definitions in [23, 50] to formalize the*HFR* task.

3.1 Definition of HFR problem

Let's define a domain $\mathcal{D}$, consisting of samples $X \in \mathbb{R}^d$ and their corresponding marginal distribution $P(X)$, where d represents the dimensionality. The function of a Face Recognition (FR) system, denoted as $\mathcal{T}^{fr}$, is characterized by a label space Y. The relationship between X and Y is expressed through the conditional probability $P(Y|X,\Theta)$, with X and Y being random variables and Θ denote the model parameters. During the training phase of an FR system, the model learns $P(Y|X,\Theta)$ in a supervised manner using a dataset of facial images $X = \{x_1, x_2, ..., x_n\}$ and their corresponding identities $Y = \{y_1, y_2, ..., y_n\}$.

Now, consider the Heterogeneous Face Recognition (HFR) problem, in HFR, we consider two distinct domains: the source domain $\mathcal{D}^s = \{X^s, P(X^s)\}$ and the target domain $\mathcal{D}^t = \{X^t, P(X^t)\}$, both linked by the shared label space Y.

The main challenge in the HFR problem, denoted as $\mathcal{T}^{hfr}$, is to identify a parameter set $\hat{\Theta}$ that satisfies the condition $P(Y|X^s,\Theta) = P(Y|X^t,\hat{\Theta})$. The methodology for estimating $\hat{\Theta}$ varies across different HFR approaches.

3.2 PDT framework

We initially consider that we have access to sample sets from two domains: $X_s = \{x_1, x_2, ..., x_n\}$ from $\mathcal{D}^s$ and $X_t = \{x_1, x_2, ..., x_n\}$ from $\mathcal{D}^t$. Both sets are associated with the same set of labels $Y = \{y_1, y_2, ..., y_n\}$. We also assume the availability of the parameters of a Facial Recognition (FR) model, denoted as Θ_{FR}, which in our context, specifically refers to a pre-trained FR model trained using visible spectrum images. Building on the concept of synthesis-based HFR, we propose that a module with a learnable parameter set θ_{PDT} can convert the target domain image into a new representation ($\hat{X^t} = \mathcal{F}_{\mathcal{PDT}}(X^t)$). This transformation aims to minimize the domain disparity while preserving essential discriminative features. This new representation ($\hat{X^t}$) can then be utilized in conjunction with a pre-trained FR model to effectively perform the Heterogeneous Face Recognition (HFR) task. Note that, this transformed representation is not optimized for perceptual quality but to preserve and transform discriminative features.

To achieve this objective, we propose to introduce a compact network module named "Prepended Domain Transformer " (PDT) in front of a pre-trained Facial Recognition (FR) model. The proposed framework is illustrated in a schematic diagram shown in Fig. 2. Essentially, this module functions as a transformer for images from the target modality, producing a *transformed* image represented as ($\mathcal{F}_{\mathcal{PDT}}(X^t)$).

This approach can be seen as an expansion of the DSU [23] method. Instead of adapting the lower layers as DSU does, we introduce a neural network block, named Prepended Domain Transformer (PDT), in front of a pre-trained Facial Recognition (FR) model to specifically address domain-specific information. A significant limitation of DSU, however, is its dependency on the initial layers of the FR model. This means that the adaptation is limited to either adapting or freezing layers that are already a part of the FR network. Moreover, the architecture of the low-level layers in the FR network is tailored for optimal performance in visible spectrum face recognition, which might not be ideally suited for the Heterogeneous Face Recognition (HFR) task. This creates a bottleneck in learning potential. In contrast, Prepended Domain Transformer offers greater flexibility. By altering the architecture of the PDT block, it's possible to even modify the local receptive field. One could also employ a neural architecture search [51] to fine-tune the architecture of the PDT block for specific heterogeneous applications. The design of the PDT is more generic, fitting a broader range of heterogeneous tasks, making PDT more versatile compared to DSU. Furthermore, in PDT, the transformation occurs directly in the pixel space, allowing this method to be integrated into various FR architectures as a plug-in module. For comparison, we have also re-implemented the DSU heterogeneous face recognition approach using the recent Iresnet100 pre-trained model as an additional baseline (DSU-Iresnet100).

The *transformed* image is then passed into a pre-trained Facial Recognition (FR) model to obtain the embeddings necessary for the Heterogeneous Face Recognition (HFR) task. Our proposed method allows us to formulate the HFR problem as follows:

$$P(Y|X_t, \hat{\Theta}) = P(Y|X_t, [\theta_{PDT}, \Theta_{FR}]) \tag{1}$$

The parameters of the PDT block (θ_{PDT}) can be optimized in a supervised manner using back-propagation. During the forward pass for a pair of images (X^s, X^t), the X^s image is directly processed through the shared pre-trained FR network to produce its embedding. Conversely, the target image (X^t) first goes through the PDT module $(\hat{X^t} = \mathcal{F}_{\mathcal{PDT}}(X^t))$, and subsequently, this *transformed* image is passed through the same pre-trained FR model to generate its embedding. For the training phase, contrastive loss [52] is utilized as the loss function. This loss formulation tries to minimize the distance between these cross-modal embeddings when the identities are identical and to maximize the distance when the identities differ. The formula for Contrastive loss is as follows:

$$\begin{aligned}\mathcal{L}_{Contrastive}(\Theta, Y_p, X_s, X_t) =& (1 - Y_p)\frac{1}{2}D_W^2 \\ &+ Y_p\frac{1}{2}max(0, m - D_W)^2\end{aligned} \tag{2}$$

Where Θ represents the weights of the network, while X_s and X_t denote the heterogeneous image pairs. The label Y_p indicates whether these pairs belong to the same identity or not. The margin is denoted by m, and D_W is the function used to calculate the distance between the embeddings of the two samples. Specifically, $Y_p = 0$ when the identities in X_s and X_t are identical, and $Y_p = 1$ when they are different. The distance function D_W is typically calculated as the Euclidean distance between the features extracted by the network.

During the training phase, the parameters of the shared Facial Recognition (FR) model are kept frozen, and only the parameters of the PDT module are adapted in the backward pass. Upon completion of the training, the model that shows the lowest validation loss is chosen for further evaluation.

3.3 Architecture of the Prepended Domain Transformer (PDT) block

The Prepended Domain Transformer block is designed to be both parameter-efficient and versatile, enabling its application across various Heterogeneous Face Recognition (HFR) scenarios. It is designed to accept and produce "three-channel" images of the same size for both its input and output. This design facilitates straightforward visualization of the output from the PDT module and seamless integration of the transformed images into pre-trained Facial Recognition (FR) models during the inference stage. Moreover, this module can be conveniently "plugged" into any existing pre-trained FR system, effectively transforming it into an HFR pipeline.

The architecture of the PDT module is shown in Fig. 3. The initial stage of the PDT module draws inspiration from the inception architecture [53], emphasizing the concept of multi-scale processing through parallel branches with

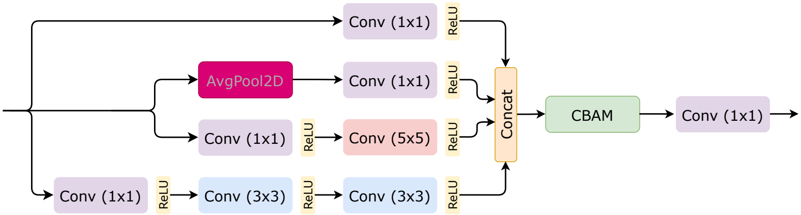

Fig. 3 Architecture of the Prepended Domain Transformer (PDT) block.

different kernel sizes. The inclusion of parallel branches is essential because the required receptive field varies in various heterogeneous settings, and incorporating multi-scale features at the input level results in a generic design with minimal computational complexity. Specifically, four parallel paths start from the input image: 1) a 1×1 filter, a 3×3 branch with two sequential filters, a 5×5 branch, and an average pooling branch. Within each branch, 1×1 convolutions are applied to reduce the number of output channels, and ReLU activation is applied after each convolution operation. The feature map is constructed by concatenating the outputs from these branches, which contain features obtained using filters with varying receptive fields. To enhance the network's ability to determine "what" and "where" to focus, an attention mechanism is introduced. We incorporate a Convolutional Block Attention Module (CBAM) [54], which efficiently attends both channel and spatial dimensions within the feature map. The attention maps generated by the CBAM block are multiplied by the input feature map, facilitating a selective emphasis on meaningful features along both channel and spatial dimensions. This addition of the CBAM module renders the proposed architecture robust across a wide range of Heterogeneous Face Recognition (HFR) scenarios. Following the CBAM block, the channel dimension of the output feature map is still relatively high, so a 1×1 convolutional layer is introduced to reduce the channel dimension to three.

The total number of parameters to be learned in this module is a mere 1.4K. This minimalist design allows the network to prioritize essential features with minimal parameter overhead. It's important to note that this module can be further fine-tuned and optimized for specific heterogeneous scenarios.

3.4 Face Recognition backbone

As mentioned in the previous sections, the PDT module can be seamlessly integrated into any pre-trained Facial Recognition (FR) model. While most of our experiments were conducted using the *Iresnet100* model [55], this approach can be extended to a variety of publicly available pre-trained FR models. To ensure reproducibility, we employed publicly accessible pre-trained face recognition models sourced from [56] [2], which were trained on the MS-Celeb-1M-v1c dataset [3]. This dataset comprises 72,778 unique identities and approximately 3.28 million images.

[2] https://github.com/JDAI-CV/FaceX-Zoo

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

In most instances, these pre-trained FR models are designed to process three-channel images with a resolution of 112 × 112. To accommodate single-channel inputs, such as Near-Infrared (NIR) or thermal images, we replicate the single channel to create three-channel images without altering the network architecture. This adjustment was necessary since the pre-trained networks were specifically engineered to handle three-channel RGB images, and we wanted to preserve the layers and weights of the FR network to maintain its performance on RGB images without any modifications.

3.5 Implementation details

The PDT framework is trained using Contrastive loss within a conventional Siamese network setup [52]. In all experiments, a margin parameter of 2.0 is used. We employed the Adam Optimizer with a learning rate of 0.001 and trained the model for 20 epochs, utilizing a batch size of 90. The implementation of the framework was done using PyTorch and the Bob library [57, 58] [4]. In the Siamese network, the entire pre-trained Facial Recognition (FR) model is shared between the source and target modalities, except for the newly added PDT module in the target channel branch. During training, only the parameters of the PDT module are adapted, while the weights of the FR model remain fixed. It's worth noting that our experiments are designed to be reproducible, and the source code and protocols are publicly available [5].

The PDT method is versatile and applicable to a range of Heterogeneous Face Recognition (HFR) scenarios, including VIS-Thermal, VIS-SWIR, VIS-Low resolution VIS, and more. In the experimental section, we showcase results from experiments conducted on VIS-NIR, VIS-THERMAL, VIS-SKETCH, VIS-Low resolution, and VIS-DEPTH modalities. Furthermore, we have intentionally kept the components of the proposed framework and the training procedure simple to emphasize the effectiveness of our approach.

4 Experiments

This section presents a comprehensive series of experiments and ablation studies with the PDT approach. Our primary focus has been on assessing the performance of VIS-Thermal *HFR* across four distinct datasets. Additionally, we benchmark the PDT method against other heterogeneous settings like VIS-NIR and VIS-SWIR. Further investigations include analyzing the amount of paired data needed for model training and analyzing performance variations across different FR architectures. For all the evaluations, the standard cosine distance metric is employed for matching the reference and target embeddings.

4.1 Databases and Protocols

The following section details the datasets used for our evaluations.

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

[5] https://gitlab.idiap.ch/bob/bob.paper.tifs2022_hfr_prepended_domain_transformer

Polathermal dataset: The Pola Thermal dataset [59] – Polarimetric and Thermal Database, is an *HFR* dataset compiled by the U.S. Army Research Laboratory (ARL). This dataset encompasses polarimetric LWIR (long-wave infrared) imagery and simultaneously captured color images of 60 subjects. It includes both conventional thermal images and polarimetric images. In our study, we employ the conventional thermal images, adhering to the reproducible protocols introduced in [23]. We maintain the same five-fold partitions where the 60 subjects are divided into two groups: 25 identities for the training set and 35 for testing. For comparing different methodologies, we report the average Rank-1 identification rate computed from the evaluation set of the five folds.

Tufts face dataset: The Tufts Face Database [60], provides a diverse collection of face images across different modalities for the *HFR* task. In particular, we utilize the thermal images from this dataset to evaluate VIS-Thermal *HFR* performance. The dataset encompasses a total of 113 individuals, including 39 males and 74 females from varied demographic backgrounds. Each subject is represented through images in multiple modalities. For comparative analysis, we adopt the method used by the authors in [21], where 50 identities are randomly chosen as the training set, and the rest serve as the test set. Our results include Rank-1 accuracies and Verification rates at false acceptance rates (FAR) of both 1% and 0.1%.

ARL-VTF dataset: The DEVCOM Army Research Laboratory Visible-Thermal Face Dataset (ARL-VTF), as presented in [17], offers a diverse collection of data from 395 subjects. This dataset features over 500,000 images captured using three visible spectrum cameras and one thermal (long-wave infrared - LWIR) camera. It includes variations in expressions, poses, and eyewear. We adhere to the original evaluation protocols provided with the dataset, which also includes annotations for facial landmarks. Additionally, the dataset offers several protocols for assessing the impact of pose, expressions, and eyewear. The test set for each scenario is standardized to facilitate direct comparisons with state-of-the-art methods.

Each protocol within the dataset is named according to the following conventions: Gallery and Probe protocols are designated "**G**" and "**P**" respectively. "**V**" and "**T**" denote the visible and thermal images. Categories such as "**B**", "**E**", and "**P**" denote baseline, expression, and pose sequences. Presence of the "$*$" symbol indicates any or all sequence categories in the protocol. A subject who does not possess glasses has the tag **0**, and **-** and $+$ tags are present for subjects who have their glasses removed or worn respectively.

The protocol names in the dataset are structured as follows:

$$\begin{aligned}
\langle\text{set}\rangle &::= \text{“}\mathbf{G}\text{”} \mid \text{“}\mathbf{P}\text{”}; \\
\langle\text{modality}\rangle &::= \text{“}\mathbf{V}\text{”} \mid \text{“}\mathbf{T}\text{”}; \\
\langle\text{sequence}\rangle &::= \text{“}\mathbf{B}\text{”} \mid \text{“}\mathbf{E}\text{”} \mid \text{“}\mathbf{P}\text{”} \mid \text{“}*\text{”}; \\
\langle\text{eyewear}\rangle &::= \text{“}\mathbf{0}\text{”} \mid \text{“}\mathbf{-}\text{”} \mid \text{“}+\text{”};
\end{aligned}$$

$$\langle\text{protocol}\rangle ::= \langle\text{set}\rangle, \text{“\_”}, \langle\text{modality}\rangle, \langle\text{sequence}\rangle, [\langle\text{eyewear}\rangle+];$$

For a comprehensive explanation of these protocols, refer to [17].

CASIA NIR-VIS 2.0 dataset: The CASIA NIR-VIS 2.0 Face Database [61], consists of images of subjects captured in both visible and near-infrared (NIR) spectrums, featuring a total of 725 identities. In this dataset, each subject is represented by 1-22 visible images and 5-50 NIR images. For experimental purposes, a 10-fold cross-validation protocol is employed, with 360 identities designated for training. The evaluation is conducted using a gallery and probe set comprising 358 identities, ensuring that the train and test sets consist of non-overlapping identities. Experiments are conducted in each fold, and the results are presented as the mean and standard deviation of the performance metrics.

SCFace dataset: The SCFace dataset[62], consists of facial high-quality mugshot images for enrollment. The probe samples in this dataset simulate surveillance scenarios and are captured from various cameras, typically being of lower quality. The dataset is structured into four distinct protocols based on the distance and quality of the probe samples: close, medium, combined, and far, with the “far” protocol being the most challenging due to greater distances and lower image quality. Overall, the SCFace dataset includes 4,160 static images, covering both visible and infrared spectrums, from 130 subjects.

CUFSF dataset: The CUHK Face Sketch FERET Database (CUFSF) [41], includes 1,194 faces from the FERET database [63]. Each face in the FERET dataset is paired with a corresponding sketch image created by an artist. This dataset presents a unique challenge as the sketches often exhibit more pronounced shape exaggerations compared to the original photos. Following the protocol outlined in [64], we use 250 identities for training the model, while the remaining 944 identities form the test set. Rank-1 accuracies are reported to facilitate comparisons.

MCXFace Dataset: This section introduces the Multi-Channel Heterogeneous Face Recognition (MCXFace) *HFR* dataset, derived from the previously created HQ-WMCA dataset [65, 66]. The MCXFace dataset features images of 51 subjects captured across various channels, in three different sessions, and under multiple lighting conditions. The available channels include color (RGB), thermal, near-infrared (850 nm), short-wave infrared (1300 nm), Depth, Stereo depth, and depth estimated from RGB images using the 3DDFA method [67]. All channels are spatially and temporally aligned across all modalities. Detailed information about the sensors and data collection sessions is available in our earlier works [65, 66]. The MCXFace dataset exclusively contains bonafide samples and is divided into ’train’ and ’dev’ sets with distinct identities, enabling experiments in various homogeneous and heterogeneous settings. Each protocol is named in the format $< SOURCE > - < TARGET > - < split >$. Additionally, annotations for the left and right eye centers are provided for all images. The dataset is publicly accessible at the following link [6].

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

Each file is a ".jpg" file with a resolution of 1920 × 1200. A sample with different modalities is shown in Fig. 4.

Protocols: There are several homogeneous as well as heterogeneous experimental protocols in the MCXFace dataset. For example, the protocol name "VIS-THERMAL-split3" indicates that the source or enrollment is RGB (VIS), and the probes are from the THERMAL channel, meaning it's a heterogeneous experiment. Also, it corresponds to "split3" in the VIS-THERMAL five-fold splits. SOURCE and TARGET will be the same for homogeneous experiments. For each set, there is a *train* set which contains images of different identities with both SOURCE and TARGET images. There is also a *dev* set, which contains the SOURCE (modality) images for enrollment (gallery), and TARGET (modality) images for probing (probes). Training and model selection are done by partitioning the *train* set. The score files from this *dev* set are used for comparison, and are not to be used in training or model selection. For comparing results with different methods experiments need to be done using all five folds in each of the source-target combinations and mean and standard deviations can be reported.

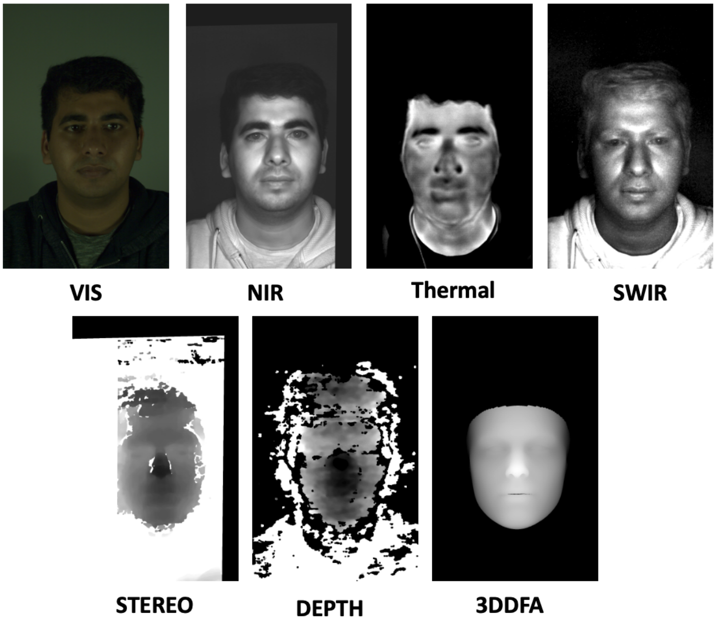

Fig. 4 Face images of a subject in the MCXFace dataset. In the first row, images are from color, near-infrared, short-wave infrared, and thermal channels. The second row shows different versions of depth: depth synthesized from RGB images using the 3DDFA method, depth computed from stereo cameras, and depth computed by a commercial depth camera.

4.2 Metrics

For model evaluation, we employ multiple metrics in line with those used in previous literature. The metrics we've selected include a subset from the following key performance indicators: Area Under the Curve (AUC), Equal Error Rate (EER), Rank-1 identification rate, and Verification Rate at various false acceptance rates (0.01%, 0.1%, 1%, and 5%).

4.3 Experimental Results

This section details the experiments conducted across various datasets and discusses the results.

4.3.1 Experiments with Polathermal dataset

We conducted experiments focusing on thermal to visible recognition scenarios. The results, as shown in Table. 2, shows the average Rank-1 identification rate for the five protocols within the Polathermal 'thermal to visible protocols', based on the reproducible protocols outlined in [23, 68]. Our re-implementation of the DSU-Iresnet100 baseline outperforms the original model from [23], suggesting that employing a superior pre-trained model enhances the results. Notably, the PDT approach achieves an average Rank-1 accuracy of 97.1% with a standard deviation of 1.3%, significantly outperforming the results of other baselines in the literature.

Table 2 Pola Thermal - Mean Rank-1 recognition rate

Method	Mean (Std. Dev.)	Info
DPM in [59]	75.31 % (-)	Paper baseline
CpNN in [59]	78.72 % (-)	Paper baseline
PLS in [59]	53.05% (-)	Paper baseline
LBPs + DoG features in [36]	36.8% (3.5)	Reproducible baseline
ISV in [69]	23.5% (1.1)	Reproducible baseline
GFK in [70]	34.1% (2.9)	Reproducible baseline
DSU(Best Result) [23]	76.3% (2.1)	Reproducible
DSU-Iresnet100	88.2% (5.8)	Reproducible
PDT (Proposed)	**97.1% (1.3)**	Reproducible

4.3.2 Experiments with Tufts face datasets

Table 3 presents the performance of the PDT approach compared to other state-of-the-art methods on the VIS-Thermal protocol of the Tufts face dataset. This dataset poses significant challenges due to variations in pose and other factors. It is known that the performance of visible spectrum face recognition systems diminishes at extreme yaw angles; thus, a similar decline is expected in *HFR* performance. Despite these challenges, it is evident that the proposed approach achieves the highest verification rate and ranks second in Rank-1 accuracy, just behind DVG-Face [21]. These results demonstrate the effectiveness of the proposed method in handling the complexities of the Tufts face dataset.

Table 3 Experimental results on VIS-Thermal protocol of the Tufts Face dataset.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
LightCNN [71]	29.4	23.0	5.3
DVG [72]	56.1	44.3	17.1
DVG-Face [21]	**75.7**	68.5	36.5
DSU-Iresnet100	49.7	49.8	28.3
PDT (Proposed)	65.71	**69.39**	**45.45**

4.3.3 Experiments with CASIA-VIS-NIR 2.0 dataset

While the primary focus of our approach has been on challenging Thermal-VIS *HFR*, we also conducted experiments on the CASIA-VIS-NIR 2.0 dataset to demonstrate the versatility of our proposed approach in different heterogeneous scenarios, particularly in NIR-VIS recognition. The baseline results indicate that the domain gap in this scenario is relatively smaller, and some pre-trained FR models, originally trained using VIS modality, already achieve reasonable performance. Consequently, we implement tighter thresholds for evaluation, using VR@FAR=0.1% and VR@FAR=0.01% as our comparative metrics. The dataset comprises 10 sub-protocols, and we report the mean and standard deviation across all ten folds for a comprehensive comparison. The comparative results are detailed in Tab. 4. It is evident from these results that our proposed approach outperforms other state-of-the-art methods, highlighting its superior performance and generalizability across various heterogeneous recognition scenarios.

Table 4 Experimental results on CASIA NIR-VIS 2.0.

Method	Rank-1	VR@FAR=0.1%	VR@FAR=0.01%
IDNet [73]	87.1±0.9	74.5	-
HFR-CNN [74]	85.9±0.9	78.0	-
Hallucination [75]	89.6±0.9	-	-
TRIVET [76]	95.7±0.5	91.0±1.3	74.5±0.7
W-CNN [77]	98.7±0.3	98.4±0.4	94.3±0.4
PACH [78]	98.9±0.2	98.3±0.2	-
RCN [79]	99.3±0.2	98.7±0.2	-
MC-CNN [80]	99.4±0.1	99.3±0.1	-
DVR [81]	99.7±0.1	99.6±0.3	98.6±0.3
DVG [72]	99.8±0.1	99.8±0.1	98.8±0.2
DVG-Face [21]	99.9±0.1	99.9±0.0	99.2±0.1
PDT (Proposed)	**99.95**±0.04	**99.94**±0.03	**99.77**±0.09

4.3.4 Experiments with SCFace dataset

We conducted a series of experiments on the SCFace dataset, following the protocols specified for visible images within the dataset. The heterogeneity in this dataset stems from the disparity in quality between the gallery and probe images: the gallery images are high-resolution mugshots, whereas the probe images are low-resolution images from surveillance cameras. The results from these experiments are compiled in Table 5, with the reported values of the "evaluation" set of the protocols. Our baseline model utilizes a pre-trained *Iresnet100* model, consistent with other experiments, and the rows marked with PDT indicate the results from our proposed model, where the PDT model is trained using contrastive training. Additionally, we include results from the re-implemented DSU-Iresnet100 model for comparative analysis. The results demonstrate that our proposed approach enhances the performance of the baseline model. This improvement is particularly noticeable in the "far" protocol, where the quality of the probe images is significantly poor. The PDT module in this scenario appears to be effective in learning features that are invariant to quality and resolution, thereby improving the overall results.

Table 5 The performance of the proposed approach on the SCFace dataset is evaluated using two models: the baseline, which is a pre-trained *Iresnet100* model, and the PDT model, which incorporates the proposed approach.

Protocol	Method	AUC	EER	Rank-1	VR@ FAR=0.1%
Close	Baseline	100.0	0.00	100.0	100.0
	DSU-Iresnet100	100.0	0.00	100.0	100.0
	PDT	100.0	0.00	100.0	100.0
Medium	Baseline	99.81	2.33	98.60	92.09
	DSU-Iresnet100	99.95	1.39	98.98	93.25
	PDT	99.96	0.93	99.07	95.81
Combined	Baseline	98.59	6.67	91.01	77.67
	DSU-Iresnet100	98.91	4.96	92.71	80.93
	PDT	99.06	4.50	93.18	82.02
Far	Baseline	96.59	9.37	74.42	49.77
	DSU-Iresnet100	97.18	8.37	79.53	58.26
	PDT	98.31	6.98	84.19	60.00

4.3.5 Experiments with MCXFace dataset

The experiments performed on the MCXFace dataset provide a valuable opportunity to assess the performance of models across a variety of heterogeneous scenarios, including VIS-Thermal, VIS-Depth, VIS-SWIR, and VIS-NIR, among others. The results, which show the average performance across five folds, are presented in Table 6. For each modality, such as VIS-Thermal, the reported values represent the aggregate performance over the dataset's five folds. Each protocol includes a baseline evaluation using the pre-trained *Iresnet100* FR model on the respective modality. The lower baseline performance often reflects a significant domain gap. For instance, a standard FR model achieves excellent performance in the NIR and SWIR channels but shows poor results in depth and thermal channels. In Table 6, the rows marked with PDT show the results using our proposed approach. Additionally, results from the re-implemented DSU-Iresnet100 model are also included as a baseline for comparison. A notable improvement is observed in the thermal channel with the proposed approach. However, while there is an enhancement in the depth channel, the final results are still not entirely satisfactory. This suggests that the depth channel may require a different approach, and treating depth data as range images might not be optimal. Alternative representations, such as normals or point clouds, could potentially be more effective for depth modality.

Table 6 Performance of the proposed approach in the MCXFace dataset, the Baseline is a pre-trained *Iresnet100* model, and the PDT is with the proposed approach.

Protocol	Method	AUC	EER	Rank-1	VR@ FAR=0.1%
VIS-Thermal	Baseline	84.45 ± 3.70	22.07 ± 2.81	47.23 ± 3.93	19.76 ± 2.73
	DSU-Iresnet100	98.12 ± 0.75	6.58 ± 1.35	83.43 ± 5.47	52.32 ± 10.06
	PDT	98.43 ± 0.78	6.52 ± 1.45	84.52 ± 5.36	59.05 ± 13.95
VIS-Depth	Baseline	53.33 ± 4.20	48.11 ± 3.40	5.19 ± 1.20	0.00 ± 0.00
	DSU-Iresnet100	52.99 ± 4.74	48.37 ± 0.16	4.91 ± 3.40	0.45 ± 0.62
	PDT	62.16 ± 5.41	41.71 ± 4.31	9.11 ± 3.07	0.70 ± 1.05
VIS-SWIR	Baseline	100.00 ± 0.00	0.03 ± 0.02	100.00 ± 0.00	99.95 ± 0.10
	DSU-Iresnet100	99.99 ± 0.01	0.16 ± 0.20	99.90 ± 0.21	99.65 ± 0.39
	PDT	100.00 ± 0.00	0.06 ± 0.08	99.95 ± 0.12	99.85 ± 0.23
VIS-NIR	Baseline	100.00 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
	DSU-Iresnet100	100.00 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
	PDT	100.00 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

4.3.6 Experiments with ARL-VTF dataset

The ARL-VTF dataset provides a large-scale VIS-Thermal dataset encompassing a variety of variations such as pose, expressions, and eyewear. This diversity enables comprehensive experiments to assess the impact of different factors on recognition performance. In our evaluations, we employed the standard protocols included with the dataset. The test sets for different protocols were fixed, and the best model selected from cross-validation was used for performance evaluation. Table 7 presents the performance of our proposed approach in comparison to state-of-the-art methods. It is evident that our approach achieves top-tier performance in most scenarios. This improvement is particularly noticeable in protocols that involve challenging pose variations (P_TP-). For example, in the $P\_TB0 - G\_VB0-$ protocol, where all images are frontal, our method attains a verification rate of 98.57% for FAR1%. However, in the $P\_TP0 - G\_VB0-$ protocol, which includes samples with pose variations, the verification rate falls to 60.80%, with a corresponding Rank-1 accuracy of 60.23%. The Rank-1 accuracy for purely frontal faces is 99.14%, but this drops to 60.23% when pose variations are introduced. This decline is observed across all methods, suggesting that heterogeneous face recognition (HFR), like homogeneous face recognition, also experiences performance degradation with pose variations.

Table 7 Verification performance comparisons with state-of-the-art methods for different protocols in ARL-VTF dataset.

		Gallery G_VB0-				Gallery G_VB0+			
Probes	Method	AUC	EER	VR@FAR=1%	VR@FAR=5%	AUC	EER	VR@FAR=1%	VR@FAR=5%
P_TB0	Raw	61.37	43.36	3.13	11.28	62.83	42.37	4.19	13.29
	Pix2Pix [82]	71.12	33.80	6.95	21.28	75.22	30.42	8.28	27.63
	GANVFS [83]	97.94	8.14	75.00	88.93	98.58	6.94	79.09	91.04
	Di *et al.* [24]	99.28	3.97	87.95	96.66	99.49	3.38	90.52	97.81
	Fondje *et al.* [25]	99.76	2.30	96.84	98.43	99.87	1.84	97.29	98.80
	PDT (Proposed)	**99.95**	**1.13**	**98.57**	**100.00**	**99.95**	**1.14**	**98.57**	**100.00**
P_TB-	Raw	61.14	41.64	2.77	16.11	57.61	44.73	1.38	6.11
	Pix2Pix [82]	68.77	38.02	6.69	20.28	52.11	48.88	2.22	4.66
	GANVFS [83]	99.36	3.77	84.88	97.66	87.34	18.66	7,00	29.66
	Di *et al.* [24]	99.63	2.66	91.55	98.88	89.24	19.49	16.33	41.22
	Fondje *et al.* [25]	99.83	1.95	96.00	99.48	99.03	4.79	85.56	95.86
	PDT (Proposed)	**99.96**	**1.18**	**98.67**	**100.00**	**99.94**	**1.33**	**98.67**	**100.00**
P_TE0	Raw	61.40	41.96	3.40	12.18	62.50	41.38	4.60	13.25
	Pix2Pix [82]	69.10	35.98	7.01	16.44	73.97	31.87	7.93	19.60
	GANVFS [83]	96.81	10.51	70.41	84.00	97.73	8.90	74.20	86.80
	Di *et al.* [24]	98.46	6.44	81.11	92.49	98.89	5.60	84.23	93.94
	Fondje *et al.* [25]	98.95	3.61	92.61	96.88	99.01	3.57	92.69	96.93
	PDT (Proposed)	**99.90**	**1.72**	**97.43**	**99.77**	**99.90**	**1.72**	**97.43**	**99.77**
P_TE-	Raw	63.26	42.34	4.66	16.28	59.33	43.17	2.04	8.00
	Pix2Pix [82]	68.78	36.24	7.75	18.06	51.05	49.11	2.26	4.95
	GANVFS [83]	98.66	5.93	73.17	92.82	83.68	22.41	6.77	22.13
	Di *et al.* [24]	99.30	3.84	82.55	97.44	86.12	21.68	9.88	31.62
	Fondje *et al.* [25]	99.83	2.27	95.66	99.48	99.48	3.05	89.45	98.07
	PDT (Proposed)	**99.95**	**0.93**	**99.07**	**100.00**	**99.90**	**1.73**	**97.87**	**100.00**
P_TP0	Raw	55.24	46.25	2.23	8.25	55.10	46.34	2.91	8.74
	Pix2Pix [82]	54.86	47.22	3.13	9.78	56.50	46.03	4.01	10.84
	GANVFS [83]	63.70	41.66	16.55	23.73	65.58	40.19	17.95	25.68
	Di *et al.* [24]	65.06	40.24	17.33	24.56	67.13	38.67	18.91	26.46
	Fondje *et al.* [25]	66.26	38.05	22.18	30.72	68.39	36.86	22.64	31.81
	PDT (Proposed)	**87.56**	**20.57**	**60.80**	**68.86**	**87.51**	**20.57**	**60.86**	**68.86**
P_TP-	Raw	55.48	45.98	3.25	8.47	56.82	44.74	2.09	7.57
	Pix2Pix [82]	54.31	47.04	2.93	8.44	50.08	49.67	0.60	4.33
	GANVFS [83]	65.79	40.35	17.84	25.48	59.51	44.04	4.29	15.47
	Di *et al.* [24]	67.27	39.00	18.16	26.02	60.10	43.57	5.77	15.97
	Fondje *et al.* [25]	68.24	37.60	23.09	33.54	63.29	41.79	18.79	27.93
	PDT (Proposed)	**87.78**	**20.40**	**65.33**	**71.20**	**87.30**	**20.65**	**60.00**	**69.87**
P_TB+	Raw	59.52	42.60	4.66	6.00	78.26	29.77	3.88	21.33
	Pix2Pix [82]	59.68	41.72	3.33	3.33	67.08	36.44	2.68	11.11
	GANVFS [83]	87.61	20.16	20.55	44.66	96.82	8.66	46.77	83.00
	Di *et al.* [24]	91.11	17.43	22.33	55.66	97.96	7.21	60.11	88.70
	Fondje *et al.* [25]	99.28	5.32	89.21	94.79	**99.97**	**0.73**	**99.47**	**100.00**
	PDT (Proposed)	**99.48**	**4.11**	**89.33**	**97.33**	99.60	4.00	90.00	97.33

4.3.7 Experiments with CUFSF dataset

Here we perform HFR experiments on challenging sketch-to-photo recognition. Table 8 shows the Rank-1 accuracies achieved using the baseline and other methods, following the protocols described in [64]. Our proposed approach attains a Rank-1 accuracy of 71.08%. However, it's noteworthy that the absolute accuracy in sketch-to-photo recognition remains relatively low. Prior works in the literature show that the performance in sketch recognition can be enhanced using specifically tailored features [84] and specially designed neural network models [85]. This is particularly relevant as the sketch modality differs significantly from other heterogeneous imaging modalities like thermal and near-infrared (NIR). While the CUFSF dataset features viewed hand-drawn sketches [86], which are visually recognizable to humans, there is a notable domain gap in the context of automatic face recognition systems. This is evidenced by the baseline performance, where the pre-trained model achieves a Rank-1 accuracy of 56.57%, highlighting a large domain gap. In contrast to imaging modalities like thermal, NIR, and SWIR, which share a high-level facial representation albeit from different aspects, sketch images incorporate artistic exaggerations and may not optimally preserve discriminative information sought by face recognition networks. This could explain the larger performance gap observed in sketch-photo recognition compared to other imaging modalities, as sketches inherently present a different set of challenges due to their artistic nature and variability.

Table 8 CUFSF: Rank-1 recognition in sketch to photo recognition

Method	Rank-1
Baseline	56.57
IACycleGAN [64]	64.94
DSU-Iresnet100	67.06
PDT (Proposed)	**71.08**

4.4 Analysis of the Framework

To gain deeper insights into the proposed method's effectiveness, additional experiments were carried out on the ARL-VTF dataset, chosen for its substantial subject variety. These experiments aimed to analyze the impact of training data volume, the form of supervision during training, and the performance metrics of various Facial Recognition (FR) architectures. All these experiments used the $G\_VB0 - P\_TB-$ protocol within the ARL-VTF dataset.

4.4.1 Evaluating Performance with Limited Training Data

Training data, especially paired heterogeneous data, is often scarce and costly when it comes to training *HFR* models. To address this, we conducted a series of experiments to examine the impact of training data volume on model performance. The ARL-VTF dataset, notable for its diverse subject range, was utilized for these experiments. We maintained a constant set of test samples, varying only in the amount (or percentage) of training and validation samples. Starting with 100% of the training data, we systematically reduced the volume in decrements of 10%, eventually reaching intervals of 1%. We also recorded the number of subjects in the training set for these scenarios. The outcomes of these experiments are detailed in Table 9. Notably, with as little as 2% of the training data, equivalent to samples from just 4 subjects, our proposed method achieved a Rank-1 accuracy of 94.67%. This high efficiency could be attributed to the parameter efficiency of our approach, where the learnable part of the PDT block contains only about 1.4K parameters, requiring minimal data for effective performance. This finding is particularly significant considering the typically small size of *HFR* datasets.

4.4.2 Investigating Performance with Unpaired Images

Until now, our experiments have involved training networks using contrastive loss in a supervised setting, under the assumption of access to paired heterogeneous samples during training. Here, we simulate an unpaired scenario where

Table 9 Experiment conducted using different subsets of training data, while maintaining the same test set across all experiments.

% of training data	Subjects	AUC	EER	Rank-1	VR@ FAR=0.1%
1%	2	83.25	25.33	20.67	5.33
2%	4	99.15	5.20	94.67	85.33
3%	7	98.46	3.33	93.33	88.00
4%	9	98.91	3.33	93.33	85.33
5%	11	98.55	3.33	96.67	89.33
10%	23	99.39	3.33	96.67	92.00
20%	47	99.73	3.33	97.33	96.67
30%	70	99.77	3.33	96.00	92.67
40%	94	99.77	2.68	97.33	95.33
50%	118	99.95	1.36	99.33	96.67
60%	141	99.9	2.67	96.67	96.67
70%	165	99.68	3.33	96.67	96.00
80%	188	99.67	3.33	96.67	96.00
90%	212	99.8	2.79	96.67	96.00
100%	235	99.96	1.18	99.33	96.67

we lack both paired heterogeneous samples and identity information of the samples. To adapt to this context, we guide our framework to align the feature distributions of the source and target modalities using Maximum Mean Discrepancy (MMD) [87] loss. We experiment using three different settings with MMD loss: 1) applying it to both the *transformed* and source images, aiming to align the inputs of the FR network, 2) using MMD loss on the embeddings generated by the FR network, and 3) applying MMD loss to both the output embeddings and inputs.

The results of these experiments are compiled in Table 10. The first row presents the baseline using a pre-trained *Iresnet100*, while the final row illustrates the outcomes with supervised learning via the PDT method. The intermediate rows show the results with MMD applied in different settings. The findings indicate that our proposed framework yields satisfactory results even in an unpaired setting, with the most favorable outcomes from the simultaneous application of MMD on both inputs and outputs. This suggests that our framework is viable even in the absence of paired samples, achieving reasonable performance through the mere alignment of source and target modality feature distributions. Nonetheless, as discussed earlier, incorporating a small number of labeled samples significantly enhances performance compared to relying solely on unpaired samples during training.

Table 10 Comparison with different types of supervision, the baseline is a pre-trained *Iresnet100*, rows with MMD corresponds to unpaired settings, and the last row is supervised with contrastive loss.

Architecture	AUC	EER	Rank-1	VR@ FAR=0.1%
Baseline	94.55	12.73	31.33	14.00
PDT + MMD (ip)	94.02	13.33	52.67	40.00
PDT + MMD (op)	97.64	6.04	75.33	51.33
PDT + MMD (op + ip)	99.49	3.33	90.00	78.67
PDT + Contrastive	99.96	1.18	99.33	96.67

Table 11 Comparison with different face recognition backbones for the *HFR* Task (These are FaceX-Zoo models)

Architecture	AUC	EER	Rank-1	VR@ FAR=0.1%
EfficientNet (baseline)	94.23	10.05	36.00	26.00
EfficientNet + PDT	99.79	2.73	94.67	84.00
MobileFaceNet (baseline)	91.19	16.62	36.00	20.67
MobileFaceNet + PDT	99.76	2.69	93.33	79.33
ResNeSt (baseline)	97.41	10.00	62.67	36.00
ResNeSt + PDT	99.96	0.63	100.00	88.00
ResNet (baseline)	94.11	14.78	44.67	19.33
ResNet + PDT	99.95	0.74	96.67	86.67
TF-NAS (baseline)	93.93	13.33	38.67	26.67
TF-NAS + PDT	99.94	0.69	99.33	86.67
GhostNet (baseline)	90.67	18.67	33.33	20.00
GhostNet + PDT	99.96	1.22	98.67	94.00
HRNet (baseline)	90.89	16.67	36.00	22.00
HRNet + PDT	99.91	1.97	96.67	88.67
Iresnet100 (baseline)	94.55	12.73	31.33	14.00
Iresnet100 + PDT	99.96	1.18	99.33	96.67

4.4.3 Experiments with different Face Recognition backbones

In the previous section, we discussed using the pre-trained *Iresnet100* Face Recognition (FR) model for all experiments. This section explores the adaptability of the proposed method to different FR architectures. These experiments were conducted on the ARL-VTF dataset, following the $G\_VB0 - P\_TB-$ protocol. We varied the pre-trained FR model within the PDT framework for each test. Besides *Iresnet100*, we utilized a variety of publicly available pre-trained FR models, as indicated in the provided link [7].

Initially, we tested these models without the PDT module to establish a baseline performance. The outcomes with the integrated trained PDT module are presented in Table 11. The table reveals that the proposed method is effective across all the tested architectures. These models differ in complexity and performance, suggesting that our approach is compatible with any FR model architecture, as long as the PDT module is concurrently trained with

[7] https://github.com/JDAI-CV/FaceX-Zoo

Table 12 The table shows Rank-1 Accuracies for cross-testing between various architectures. Rows represent the architecture employed for evaluation (A), while columns indicate the architecture used for training the PDT module (C). The results obtained using the same architecture for both training and testing are emphasized in gray.

	AttentionNet(C)	EfficientNet(C)	GhostNet(C)	HRNet(C)	Iresnet100(C)	MobileFaceNet(C)	ResNeSt(C)	ResNet(C)	TF-NAS(C)
AttentionNet(A)	99.33	91.33	0.00	95.33	88.00	94.00	96.67	96.00	97.33
EfficientNet(A)	70.67	94.67	10.00	83.33	46.67	16.00	16.00	19.33	87.33
GhostNet(A)	88.00	96.00	98.67	92.67	81.33	92.67	89.33	92.00	94.67
HRNet(A)	92.00	88.67	21.33	96.67	82.00	94.00	93.33	90.00	97.33
Iresnet100(A)	92.67	92.00	44.00	92.67	99.33	60.00	80.67	84.67	84.67
MobileFaceNet(A)	87.33	81.33	49.33	86.00	80.67	93.33	78.67	74.67	88.00
ResNeSt(A)	95.33	93.33	1.33	96.67	82.00	92.67	100.00	96.00	100.00
ResNet(A)	99.33	96.67	20.00	97.33	94.00	95.33	98.67	96.67	98.00
TF-NAS(A)	92.00	92.00	42.67	92.67	90.00	85.33	88.00	87.33	99.33

that model. Therefore, the selection of pre-trained FR models for HFR can be based on balancing accuracy and computational complexity.

4.4.4 Generalizability of PDT Weights Across Different Architectures

The preceding sub-section demonstrated the effectiveness of our proposed method across various Face Recognition (FR) backbones, regardless of their distinct architectures and complexities. It's important to note that the PDT module's architecture remained consistent despite the variation in FR architectures. Our further exploration focused on whether the PDT module, trained with one FR architecture, would be effective with another. This line of inquiry could shed light on the transformations learned and assess the feasibility of applying PDT modules to black-box models. For instance, as shown in Table. 12, the pairing of AttentionNet(C) (column) with EfficientNet(A) (row) indicates that the PDT was initially integrated and trained with the AttentionNet model, and subsequently, the PDT weights developed with AttentionNet were evaluated using the EfficientNet architecture. The results from these experiments are detailed in Table. 12 reveal that our method is generally successful across various combinations. However, an exception is observed with GhostNet [88], likely due to its distinctive architecture. GhostNet aims to lessen computational demands by minimizing the filters required for redundant feature maps. This is accomplished through inexpensive linear operations, like depthwise separable convolutions, on output feature maps, generating additional "ghost" features. The original and "ghost" features are then merged in the ghost module. This approach could mean that the PDT, when trained alongside GhostNet, becomes specifically tailored to the filters in its pre-trained backbone, thus hindering the PDT module's performance when paired with other architectures. In summary, while the PDT module is broadly effective, its parameters are inherently linked to a specific architecture, limiting its generalizability with black-box models. However, training the PDT weights using multiple FR models in a distillation framework might enhance performance with black-box FR models. This potential improvement is a topic for future research and falls outside the scope of this current study.

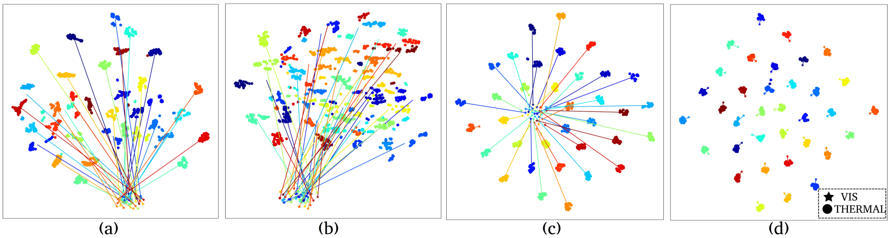

Fig. 5 T-SNE plots for visible and thermal images from various stages, different colors indicate different identities. The lines connect the cluster center of visible and thermal images for each identity. a) shows the plot of images in the original pixel space, b) shows visible images and transformed thermal images in pixel space, c) from an intermediate feature map of the CNN, and d) final embedding space. It can be seen that the identities align in the final embedding space.

4.4.5 Visualization of Intermediate Features

This section presents t-SNE plots depicting the progression of thermal and visible images through different stages of our processing pipeline, as illustrated in Fig. 5. Initially, thermal and visual modalities are clearly distinguishable in these plots. However, even after processing through the PDT stage, the samples from these two modalities continue to exhibit distinct distributions. As the data progresses through the CNN's feature maps, there's an observable alignment between the thermal and visible features. This alignment becomes particularly pronounced in the final layers and the ultimate embedding output, bringing the embeddings of the same identity within the thermal and visible spectrums closer together in the embedding space — a crucial factor for successful HFR.

Notably, even post-PDT block, the feature distributions remain separate. Despite this disparity, the "translated" image generated by the PDT block aids in aligning the embeddings in the latter layers. Since these post-PDT feature maps are also three-channel images of the same resolution, they offer a window into the output of the PDT module. For instance, Fig. 6 shows the thermal, visible, and "translated-thermal" images of the same individual. The face embeddings from these "translated" images and those from the visible spectrum generally show higher match scores. Contrary to what might be intuitively expected, the "translated" image doesn't necessarily visually resemble the VIS image; it primarily needs to retain discriminative features pertinent to the HFR task.

It's also worth noting that the feature maps post-convolutional layers are optimized more for task efficiency than for visual quality. This applies to the PDT output as well. The "translated" image, as seen in the figure, maintains the basic shape and enhances certain features like edges, while the thermal image's features seem more subdued in this representation. This modification assists in drawing the embeddings closer together in the feature space.

Elaborating on this concept, Fig. 7 displays the comparison of match scores between translated and non-translated images. These scores are derived

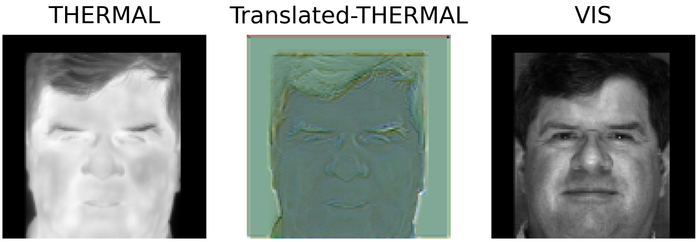

Fig. 6 An illustration of thermal to vis *HFR* scenario in Polathermal dataset, the *Translated-Thermal* is the intermediate output from the PDT module. Although this 'Translated-Thermal' image may not visually resemble the VIS image, the embeddings generated from it align closely with those from the VIS image, as indicated by the high matching scores achieved.

from the cosine similarity calculated using embeddings generated by a pre-trained face recognition network. The results demonstrate an improvement in match scores following the PDT stage for genuine matches, while simultaneously reducing the scores for imposter matches, thereby enhancing the overall performance in HFR.

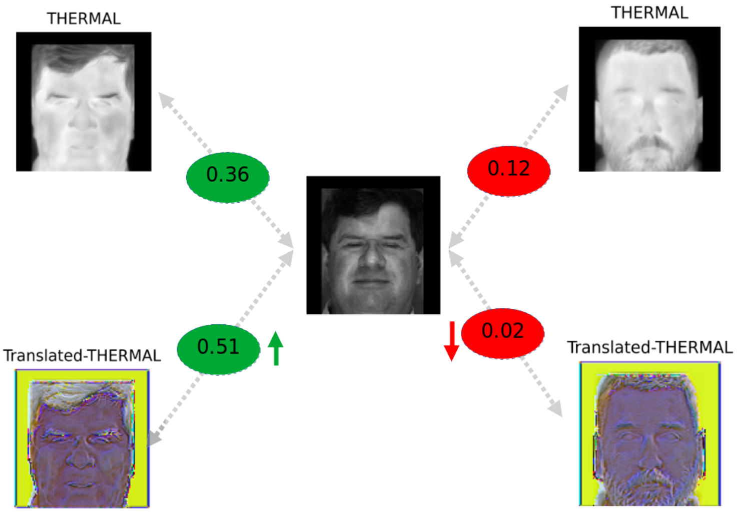

Fig. 7 This figure illustrates the matching scores between Thermal-VIS and Translated-THERMAL-VIS pairings. The scores are calculated by processing the images through a pre-trained face recognition network and measuring the cosine distance between the resulting embeddings. The results clearly show an improvement in the match scores for correct identity matches (indicated in green) and a decrease in scores for mismatched identities (indicated in red) after applying the translation to thermal images

5 Discussions

Contrary to synthesis-based methods, which are computationally intensive, our proposed method offers a straightforward, more efficient solution for heterogeneous face recognition. This innovative module is designed to be easily integrated, requiring only a limited number of samples for training. Its utility even extends to scenarios lacking paired samples. The method's effectiveness lies in its ability to learn a transformation within the pixel domain, adept at bridging the domain gap while preserving discriminative information.

This approach has demonstrated state-of-the-art performance in multiple challenging HFR benchmarks, encompassing a range of thermal to VIS and other diverse heterogeneous protocols. The newly introduced PDT module is not only efficient in terms of parameters but also versatile enough to adapt to various heterogeneous scenarios. As highlighted in Table. 9, the volume of training data needed for our framework is surprisingly small, making it particularly suited for real-world scenarios where data is often limited. Moreover, the method proves to be effective even without paired samples, a significant advantage in heterogeneous contexts where such data is scarce. The framework's modular nature allows for seamless integration with any existing pre-trained FR model. In summary, prepending a learnable neural network module to a pre-trained FR model has been shown to yield state-of-the-art results across a spectrum of challenging HFR settings.

5.1 Limitations and Future directions

The PDT block in our current framework is designed with a general architecture to accommodate a broad spectrum of heterogeneous scenarios. This design, incorporating a multi-branch structure and a CBAM module, offers flexibility in terms of receptive field. Despite its simplicity, this approach achieves performance on par with, and often surpassing, more computationally demanding state-of-the-art models. The receptive field of the branches in the PDT is a key design element, presenting opportunities for further optimization. Tailoring the PDT architecture to specific heterogeneous scenarios could potentially enhance performance even more. The efficacy of our proposed framework largely depends on the performance of the pre-trained network it incorporates. For instance, standard pre-trained FR models often struggle with extreme yaw angles and profile faces, which would adversely affect HFR performance under these conditions, as evidenced in the Tufts face dataset results. However, our approach remains versatile enough to integrate with newer, more robust face recognition models. The PDT approach is particularly suited for heterogeneous "imaging" modalities, which share greater structural similarity with visible face images. Modalities like sketches, which diverge more significantly from this criterion, might be less adaptable within this image domain and could be better served by generative methods. Although PDT blocks trained with one architecture might not perform as well with others, their compatibility with black-box models could be improved using a teacher-student model involving multiple architectures. The approach can be expanded to lightweight face recognition networks through distillation techniques [89], as outlined in works like [90, 91]. Furthermore, our approach can be integrated with GAN-based generation methods and enhanced through the use of advanced loss functions like triplet or quadruplet loss. Potential improvements could also come from further refinement of the PDT architecture, employing strategies like neural architecture search [51], optimized training schedules, data augmentation, and triplet-based training.

6 Conclusions

In this chapter, we present a simple yet highly effective framework for heterogeneous face recognition (HFR). At its core, the method involves augmenting a pre-trained face recognition (FR) model with a novel neural network module tailored to the target modality. This new module, referred to as Prepended Domain Transformer (PDT), is notable for its parameter efficiency and minimal training sample requirements. Its compatibility across various FR architectures makes it a versatile tool for transforming any standard FR model into a heterogeneous one. The design of our framework, including aspects like the training schedule, loss functions, and parameter choices, has been deliberately kept simple to showcase the potential of the proposed method. These elements can be further refined and optimized for even better performance. Our approach has demonstrated superior performance over state-of-the-art methods in several challenging heterogeneous datasets. Additionally, we also describe the MCXFace heterogeneous face recognition dataset, which encompasses multiple modalities and is ideal for HFR evaluations. To foster continued research and development in this area, we are making the source code, protocols, and datasets publicly available, enabling others to build upon and extend our work.

Acknowledgments. The authors would like to thank Innosuisse - Swiss Innovation Agency for supporting the research leading to results published in this manuscript.

Declarations

Some journals require declarations to be submitted in a standardized format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading ‘Declarations’:

- Funding: The authors would like to thank Innosuisse - Swiss Innovation Agency for supporting the research leading to results published in this chapter.
- Conflict of interest/Competing interests: ‘Not applicable’
- Ethics approval: ‘Not applicable’

- Consent to participate: 'Not applicable'
- Consent for publication: 'Not applicable'
- Availability of data and materials: Please see the footnotes in the manuscript.
- Code availability: Please see the footnotes in the manuscript.
- Authors' contributions: : 'Not applicable'

References

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

[2] Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer, Cham, Switzerland (2016). https://doi.org/10.1007/978-3-319-28501-6. https://link.springer.com/book/10.1007/978-3-319-28501-6

[3] Cao, Z., Schmid, N.A., Bourlai, T.: Local operators and measures for heterogeneous face recognition. Face Recognition Across the Imaging Spectrum, 91–115 (2016)

[4] Narang, N., Martin, M., Metaxas, D., Bourlai, T.: Learning deep features for hierarchical classification of mobile phone face datasets in heterogeneous environments. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 186–193 (2017). IEEE

[5] Narang, N., Bourlai, T.: Gender and ethnicity classification using deep learning in heterogeneous face recognition. In: 2016 International Conference on Biometrics (ICB), pp. 1–8 (2016). IEEE

[6] Li, S.Z., Chu, R., Liao, S., Zhang, L.: Illumination invariant face recognition using near-infrared images. IEEE Transactions on pattern analysis and machine intelligence **29**(4), 627–639 (2007)

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

[8] George, A., Marcel, S.: Robust face presentation attack detection with multi-channel neural networks. In: Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment, pp. 261–286. Springer, Singapore (2023)

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

[10] Mokalla, S.R., Bourlai, T.: Utilizing alignment loss to advance eye center detection and face recognition in the LWIR band. IEEE Transactions on Biometrics, Behavior, and Identity Science (2023)

[11] Bourlai, T., Rose, J., Mokalla, S.R., Zabin, A., Hornak, L., Nalty, C., Pari, N., Gleason, J., Castillo, C., Patel, V., Chellappa, R.: Data and algorithms for thermal spectrum face verification. IEEE Transactions on Biometrics, Behavior, and Identity Science (2023)

[12] Kalka, N.D., Bourlai, T., Cukic, B., Hornak, L.: Cross-spectral face recognition in heterogeneous environments: A case study on matching visible to short-wave infrared imagery. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–8 (2011). IEEE

[13] Narang, N., Bourlai, T.: Face recognition in the SWIR band when using single sensor multi-wavelength imaging systems. Image and Vision Computing **33**, 26–43 (2015)

[14] Narang, N., Bourlai, T.: Deep feature learning for classification when using single sensor multi-wavelength based facial recognition systems in SWIR band. Surveillance in Action: Technologies for Civilian, Military and Cyber Surveillance, 147–163 (2018)

[15] Ibsen, M., Rathgeb, C., Brechtel, F., Klepp, R., Pöppelmann, K., George, A., Marcel, S., Busch, C.: Attacking face recognition with T-shirts: Database, vulnerability assessment and detection. IEEE Access (2023)

[16] He, R., Wu, X., Sun, Z., Tan, T.: Wasserstein CNN: Learning invariant features for Nir-Vis face recognition. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1761–1773 (2018)

[17] Poster, D., Thielke, M., Nguyen, R., Rajaraman, S., Di, X., Fondje, C.N., Patel, V.M., Short, N.J., Riggan, B.S., Nasrabadi, N.M., *et al.*: A large-scale, time-synchronized visible and thermal face dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1559–1568 (2021)

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

[19] He, R., Wu, X., Sun, Z., Tan, T.: Learning invariant deep representation for Nir-Vis face recognition. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

[20] Tang, X., Wang, X.: Face sketch synthesis and recognition. In: Proceedings Ninth IEEE International Conference on Computer Vision, pp. 687–694

(2003). IEEE

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

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

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

[24] Di, X., Riggan, B.S., Hu, S., Short, N.J., Patel, V.M.: Polarimetric thermal to visible face verification via self-attention guided synthesis. In: International Conference on Biometrics (ICB), pp. 1–8 (2019). IEEE

[25] Fondje, C.N., Hu, S., Short, N.J., Riggan, B.S.: Cross-domain identification for Thermal-to-Visible face recognition. arXiv preprint arXiv:2008.08473 (2020)

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

[27] Lin, D., Tang, X.: Inter-modality face recognition. In: European Conference on Computer Vision, pp. 13–26 (2006). Springer

[28] Yi, D., Liu, R., Chu, R., Lei, Z., Li, S.Z.: Face matching between near infrared and visible light images. In: International Conference on Biometrics, pp. 523–530 (2007). Springer

[29] Lei, Z., Li, S.Z.: Coupled spectral regression for matching heterogeneous faces. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1123–1128 (2009). IEEE

[30] Lei, Z., Liao, S., Jain, A.K., Li, S.Z.: Coupled discriminant analysis for heterogeneous face recognition. IEEE Transactions on Information Forensics and Security **7**(6), 1707–1716 (2012)

[31] Sharma, A., Jacobs, D.W.: Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In: CVPR 2011, pp. 593–600 (2011). IEEE

[32] Cheema, U., Ahmad, M., Han, D., Moon, S.: Heterogeneous Visible-Thermal and Visible-Infrared face recognition using cross-modality discriminator network and unit-class loss. Computational Intelligence and Neuroscience **2022** (2022)

[33] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[34] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A dataset for recognizing faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74 (2018). IEEE

[35] Zhang, T., Wiliem, A., Yang, S., Lovell, B.: TV-GAN: Generative adversarial network based thermal to visible face recognition. In: 2018 International Conference on Biometrics (ICB), pp. 174–181 (2018). IEEE

[36] Liao, S., Yi, D., Lei, Z., Qin, R., Li, S.Z.: Heterogeneous face recognition from local structures of normalized appearance. In: International Conference on Biometrics, pp. 209–218 (2009). Springer

[37] Liao, S., Zhu, X., Lei, Z., Zhang, L., Li, S.Z.: Learning multi-scale block local binary patterns for face recognition. In: International Conference on Biometrics, pp. 828–837 (2007)

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

[39] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)

[40] Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence **24**(7), 971–987 (2002)

[41] Zhang, W., Wang, X., Tang, X.: Coupled information-theoretic encoding for face photo-sketch recognition. In: CVPR 2011, pp. 513–520 (2011). IEEE

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

[43] Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE

transactions on pattern analysis and machine intelligence **31**(11), 1955–1967 (2008)

[44] Liu, Q., Tang, X., Jin, H., Lu, H., Ma, S.: A nonlinear approach for face sketch synthesis and recognition. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 1005–1010 (2005). IEEE

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

[46] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv:1703.10593 [cs] (2017) https://arxiv.org/abs/1703.10593 [cs]

[47] Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE transactions on image processing **19**(6), 1635–1650 (2010)

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

[49] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training Gans. Advances in neural information processing systems **29**, 2234–2242 (2016)

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

[51] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V.: MnasNet: Platform-aware neural architecture search for mobile. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2820–2828 (2019)

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

[53] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

[54] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block

attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

[55] Pytorch InsightFace (2021). https://github.com/nizhib/pytorch-insightface

[56] Wang, J., Liu, Y., Hu, Y., Shi, H., Mei, T.: FaceX-Zoo: A pytorch toolbox for face recognition. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3779–3782 (2021)

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

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

[59] Hu, S., Short, N.J., Riggan, B.S., Gordon, C., Gurton, K.P., Thielke, M., Gurram, P., Chan, A.L.: A polarimetric thermal database for face recognition research. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 119–126 (2016)

[60] Panetta, K., Wan, Q., Agaian, S., Rajeev, S., Kamath, S., Rajendran, R., Rao, S.P., Kaszowska, A., Taylor, H.A., Samani, A., *et al.*: A comprehensive database for benchmarking imaging systems. IEEE transactions on pattern analysis and machine intelligence **42**(3), 509–520 (2018)

[61] Li, S., Yi, D., Lei, Z., Liao, S.: The CASIA Nir-Vis 2.0 face database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 348–353 (2013)

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

[63] Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image and vision computing **16**(5), 295–306 (1998)

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

[65] Heusch, G., George, A., Geissbühler, D., Mostaani, Z., Marcel, S.: Deep models and shortwave infrared information to detect face presentation attacks. IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM) (2020)

[66] Mostaani, Z., George, A., Heusch, G., Geissenbuhler, D., Marcel, S.: The high-quality wide multi-channel attack (hq-wmca) database. Idiap-RR Idiap-RR-22-2020, Idiap (September 2020)

[67] Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16, pp. 152–168 (2020). Springer

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

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

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

[71] Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13**(11), 2884–2896 (2018)

[72] Fu, C., Wu, X., Hu, Y., Huang, H., He, R.: Dual variational generation for low shot heterogeneous face recognition. In: Advances in Neural Information Processing Systems (2019)

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

[74] Saxena, S., Verbeek, J.: Heterogeneous face recognition with CNNs. In: European Conference on Computer Vision (2016)

[75] Lezama, J., Qiu, Q., Sapiro, G.: Not afraid of the dark: Nir-Vis face recognition via cross-spectral hallucination and low-rank embedding. In:

IEEE Conference on Computer Vision and Pattern Recognition (2017)

[76] Liu, X., Song, L., Wu, X., Tan, T.: Transferring deep representation for Nir-Vis heterogeneous face recognition. In: International Conference on Biometrics (2016)

[77] He, R., Wu, X., Sun, Z., Tan, T.: Wasserstein CNN: Learning invariant features for Nir-Vis face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1761–1773 (2018)

[78] Duan, B., Fu, C., Li, Y., Song, X., He, R.: Pose agnostic cross-spectral hallucination via disentangling independent factors. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)

[79] Deng, Z., Peng, X., Qiao, Y.: Residual compensation networks for heterogeneous face recognition. In: AAAI Conference on Artificial Intelligence (2019)

[80] Deng, Z., Peng, X., Li, Z., Qiao, Y.: Mutual component convolutional neural networks for heterogeneous face recognition. IEEE Transactions on Image Processing **28**(6), 3102–3114 (2019)

[81] Wu, X., Huang, H., Patel, V.M., He, R., Sun, Z.: Disentangled variational representation for heterogeneous face recognition. In: AAAI Conference on Artificial Intelligence (2019)

[82] Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference On (2017)

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

[84] Koley, S., Roy, H., Bhattacharjee, D.: Gammadion binary pattern of shearlet coefficients (GBPSC): An illumination-invariant heterogeneous face descriptor. Pattern Recognition Letters **145**, 30–36 (2021)

[85] Luo, M., Wu, H., Huang, H., He, W., He, R.: Memory-modulated transformer network for heterogeneous face recognition. IEEE Transactions on Information Forensics and Security (2022)

[86] Klum, S.J., Han, H., Klare, B.F., Jain, A.K.: The FaceSketchID system: Matching facial composites to mugshots. IEEE Transactions on Information Forensics and Security **9**(12), 2248–2263 (2014)

[87] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A

kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)

[88] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: GhostNet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580–1589 (2020)

[89] Shahreza, H.O., George, A., Marcel, S.: SynthDistill: Face recognition with knowledge distillation from synthetic data. In: 2023 International Joint Conference on Biometrics (IJCB). IEEE

[90] George, A., Ecabert, C., Shahreza, H.O., Kotwal, K., Marcel, S.: EdgeFace: Efficient face recognition model for edge devices. IEEE Transactions on Biometrics, Behavior, and Identity Science (2024)

[91] Kolf, J.N., Boutros, F., Elliesen, J., Theuerkauf, M., Damer, N., Alansari, M., Hay, O.A., Alansari, S., Javed, S., Werghi, N., *et al.*: EFaR 2023: Efficient face recognition competition. In: 2023 International Joint Conference on Biometrics (IJCB) (2023). IEEE