Prepended Domain Transformer: Heterogeneous Face Recognition without Bells and Whistles

Anjith George, Amir Mohammadi and Sebastien Marcel

***Abstract*—Heterogeneous Face Recognition (*HFR*) refers to matching face images captured in different domains, such as thermal to visible images (VIS), sketches to visible images, near-infrared to visible, and so on. This is particularly useful in matching visible spectrum images to images captured from other modalities. Though highly useful, *HFR* is challenging because of the domain gap between the source and target domain. Often, large-scale paired heterogeneous face image datasets are absent, preventing training models specifically for the heterogeneous task. In this work, we propose a surprisingly simple, yet, very effective method for matching face images across different sensing modalities. The core idea of the proposed approach is to add a novel neural network block called Prepended Domain Transformer (PDT) in front of a pre-trained face recognition (FR) model to address the domain gap. Retraining this new block with few paired samples in a contrastive learning setup was enough to achieve state-of-the-art performance in many *HFR* benchmarks. The PDT blocks can be retrained for several source-target combinations using the proposed general framework. The proposed approach is architecture agnostic, meaning they can be added to any pre-trained FR models. Further, the approach is modular and the new block can be trained with a minimal set of paired samples, making it much easier for practical deployment. The source code and protocols will be made available publicly.**

***Index Terms*—Heterogeneous Face Recognition, Convolutional Neural Network, Biometrics, Face Recognition, Cross-Modal Face Recognition.**

I. Introduction

FACE recognition (FR) systems offer a convenient way for access control. Most of the state-of-the-art FR methods achieve excellent performance in 'in the wild' conditions and human parity in face recognition performance [1], thanks to convolutional neural networks. Typical FR systems operate in the homogeneous domain, meaning the enrollment and matching are performed with the same modality, typically with face images obtained from an RGB camera (visible spectrum).

A. George, A. Mohammadi and S. Marcel are in Idiap Research Institute, Centre du Parc, Rue Marconi 19, CH - 1920, Martigny, Switzerland

Manuscript received September xx, 2021; revised September xx, 2021.

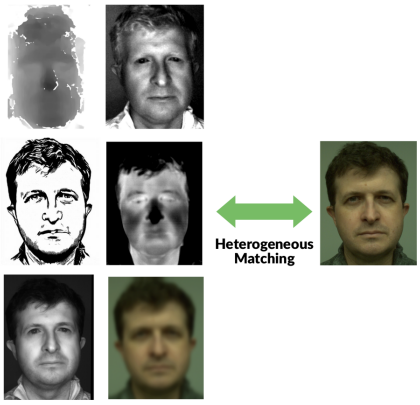

Fig. 1. This figure shows the face of the same person captured with several different modalities such as depth, short-wave infrared, sketch, thermal, near-infrared and blurred faces. The task in *HFR* is to perform cross-modal face recognition given the RGB reference images and the probe image from the new modalities.

However, in several scenarios, matching in a heterogeneous setting could be advantageous. For example, near-infrared (NIR) cameras, ubiquitous in mobile phones and surveillance cameras, offer superior performance irrespective of the illumination conditions [2]. However, an FR system operating in a homogeneous setting will require enrollment samples obtained under the same NIR camera. Heterogeneous face recognition (*HFR*) systems try to alleviate this limitation, by enabling ways to perform cross-domain matching. For example, with an *HFR* system, enrolled RGB images could be matched with NIR images, obviating the requirement for enrollment of different modalities [3]. This also opens up the possibility of using additional channels for recognition, such as thermal and shortwave-infrared, without having enrollment samples in the corresponding modalities. Thermal images can be acquired without using active illumination and hence could be used robustly in day and night conditions. In short, the *HFR* task is very useful in practical applications and could provide a way to extend FR to challenging and uncontrolled environments.

Though *HFR* is very useful, it is very challenging for several reasons. First, there is a large domain gap between the images captured by various sensors. The

performance of the networks trained using RGB images deteriorates when used with images captured with other sensing modalities [4]. Further, there are not many large-scale heterogeneous datasets to train models specifically for *HFR* [5]. Leveraging pre-trained FR models, which are trained using large-scale face recognition datasets, appears to be a reasonable strategy to follow in this limited data domain. Furthermore, special emphasis should be put on developing approaches that work with a limited amount of paired data as this type of heterogeneous data is costly to acquire and not readily available [5].

In this work, we propose a novel, yet surprisingly simple approach for *HFR*. We leverage a pre-trained FR model as one of the key components in our framework. Our approach doesn't depend on the selection of architecture of the FR model giving it maximum flexibility in deployment. We achieve this by prepending a new network module, called Prepended Domain Transformer (PDT), to a pre-trained FR module to transform the target domain images. The only learnable component is the new prepended module, which is very parameter efficient and obtains excellent performance with few paired samples. This method is very practical in deployment scenarios since one just needs to prepend a new module to convert a typical FR pipeline to an *HFR* pipeline. The approach is generic and can be retrained easily for any pair of heterogeneous modalities. Through extensive evaluations, we show that this simple addition achieves state-of-the-art results in many challenging *HFR* datasets. The framework's design is intentionally kept simple to demonstrate the effectiveness of the approach and to allow for future extensions. Moreover, the parameter and computational overhead added by the framework is negligible, making the proposed approach suitable for real-time deployment.

The main contributions of this work are listed below:

- We propose a heterogeneous face recognition framework, leveraging a pre-trained FR model together with a Prepended Domain Transformer block. The approach is architecture agnostic and can be used in many heterogeneous recognition scenarios. The method has minimal computational overhead and a very minimal number of learnable parameters, which makes it easy for deployment.
- We perform an extensive set of experiments on public datasets and validate the effectiveness of the proposed approach in different heterogeneous face recognition settings.
- We introduce a new multi-channel heterogeneous face recognition (MCXFace) dataset which consists of both homogeneous and heterogeneous protocols with channels such as color, thermal, depth, stereo, infrared, short-wave infrared, and synthesized 3D maps. We provide standard protocols and baselines for the heterogeneous protocols in this dataset.

Finally, the source codes are made available publicly to make it easier to extend the work further [1].

The rest of the paper is organized as follows. Section II presents recent literature on *HFR*. Details of the proposed approach are described in Section III. Extensive evaluation of the proposed approach, along with comparisons with the state-of-the-art, and discussions, are presented in Section IV. Conclusions and future directions are described in Section VI.

II. Related work

The task in *HFR* is matching face images collected in different sensing modalities. The challenge lies in the fact that the image of the same subject appears very different in different modalities (domain gap). This difference in appearance increases the within-class variance making matching difficult, and a direct comparison of these heterogeneous images degrade the performance. Several approaches have been proposed in the literature to address this limitation.

A. Common-space projection method

Common-space projection methods [6], [7] aim to learn a mapping to project different face modalities into a common shared subspace in an effort to reduce the domain gap. Lin and Tang [8] developed a common discriminant feature extraction method for extracting features from cross-modal images and projecting them onto a common feature space. Canonical correlation analysis (CCA) was proposed as a way to match face images between NIR and VIS by Yi *et al.* [9]. To learn mapping functions that connect cross-modality domains and common spaces, regression-based approaches were proposed by authors in [10], [11]. Sharma and Jacobs [12] proposed a partial least squares-based method to learn a linear mapping for face image across different modalities so that the mutual covariance is maximized. Klare and Jain [3] proposed a way to represent face images in terms of their similarity to a set of prototype face images. The prototype-based face representation was then projected onto a linear discriminant subspace, which was used to perform the recognition. In [13], the authors proposed a novel approach to the *HFR* task called Domain-Specific Units (DSU). Essentially, they suggest that high-level features of convolutional neural networks trained on the visible spectrum are domain-independent, and they can be used to encode images captured in other sensing modalities. Subsequently, they proposed adapting the initial layers (DSUs) of a pre-trained FR model to make it suitable for different heterogeneous scenarios.

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

Essentially, adapting the lower layers reduces the domain gap, and the whole pipeline is trained in a contrastive setting. However, the number of layers that need to be adapted is a hyper-parameter that has to be found with an extensive set of experiments. Furthermore, this approach requires the model to be adapted for each different architecture of pre-trained FR models. Recently, Cheema et al. [14] proposed a Cross Modality discriminator network (CMDN) for HFR. The architecture of CMDN follows a standard ResNet50 model with the squeeze and excitation module (SENet-50) [15]. Their approach uses a Deep Relational Discriminator (DRD) module to learn cross-domain matching. This DRD module is essentially a multi-layer perceptron (MLP) supervised by binary cross entropy loss (BCE). The learning of the CMDN module is supervised by the unit class loss which is a combination of triplet loss and a modified version of triplet loss that uses class means. The CMDN module (SENet-50) is initialized from a pre-trained face recognition backbone trained on the VGGFace2 dataset [16]. The CMDN network is further trained on another Visible-Thermal face dataset (IRIS face dataset [17]). The tuned CMDN network is further trained for the HFR task using a variant of triplet loss (Unit class loss). Now the embeddings obtained from two modalities (gallery and probe modalities) are concatenated for positive and negative pairs and an MLP model (DRD module) is trained on top of these concatenated embeddings with the BCE loss. The scoring is performed with probe-gallery pairs, and three different strategies were used for scoring 1) Using the embeddings and cosine loss, 2) using the output of the DRD module, and 3) score fusion of 1 and 2. Their evaluation strategy uses pairs of samples instead of probes against the gallery. They reported that the fusion model achieves better results.

B. Invariant feature based methods

Invariant feature-based methods aim to extract modality invariant features to match heterogeneous face images. Liao *et al.* [18] propose to use Difference of Gaussian (DoG) filters to highlight the structure of the images. Then, they use multi-scale block local binary patterns (MB-LBP) [19] as features and train a subspace-based face recognition system jointly on the samples of source and target domains. In [20], Klare *et al.* proposed a local feature-based discriminant analysis (LFDA) for the *HFR* task. They extracted scale-invariant feature transform (SIFT) [21] and multi-scale local binary pattern (MLBP) [22] feature descriptors as a patch unit from the sketch and VIS images for the *HFR* task. Zhang *et al.* [23] proposed a coupled information-theoretic encoding (CITE) extraction method to maximize the mutual information between the heterogeneous modalities in the quantized feature spaces. The local maximum quotient (LMQ) was proposed to extract invariant characteristics in cross-modality facial images in [24]. Several works have also used convolutional neural network (CNN) based methods [4], [7] to extract invariant features for the *HFR* task.

C. Synthesis based methods

Synthesis-based *HFR* methods [25], [26] attempt to synthesize the source domain (VIS in most of the cases) from the target modality, after synthesizing the source images typical face recognition networks can be used to perform the biometric matching. Authors in [27] proposed a patch-based synthesis approach to generate VIS to sketches and reverse using Multi-scale Markov Random Fields. The approach was evaluated using several face recognition methods such as Eigenfaces, Fisherfaces, dual space LDA, and so on. The work in [28] used Locally Linear Embedding (LLE) to learn a pixel-level mapping between VIS images and viewed sketches. Authors in [29] use CycleGAN [30] to transform images from the target domain to the source domain. Contrastive loss using a Siamese network is added during the training of CycleGAN to preserve the identity of faces. Moreover, the images are pre-processed before inputting to CycleGAN using [31] to reduce the domain gap further. Authors in [32], proposed a Generative Adversarial Network-based Visible Face Synthesis (GAN-VFS) method to synthesize photo-realistic visible face images from polarimetric images. Identity loss was combined with a perceptual loss in the training process. The synthesized visible images were further used by a VGG network to extract the embeddings. Their method was evaluated on the Polathermal dataset and achieved an average Equal Error Rate of 34.58%. With the advancement in the development of stable methods to train GANs, several recent approaches have been proposed using GANs for the synthesis of VIS images from another modality. The work in [26] treated *HFR* as a dual generation problem and proposed a Dual Variational Generation (DVG-Face) framework. A Dual generator was designed to learn the joint distribution of heterogeneous pairs and to generate heterogeneous pairs to address the lack of adequate data to train the *HFR* model. A pairwise identity preserving loss on the generated images was employed to ensure identity consistency. The generated images are used to train the *HFR* network in a contrastive setting. This approach achieved state-of-the-art results in many challenging *HFR* benchmarks.

D. Limitations of current approaches

The majority of recent *HFR* methods proposed in the literature [26], [32] utilize synthesis-based methodologies. GAN-based synthesis methods have become popular due to their ability to generate high-quality

images combined with breakthroughs in training them [33]. Furthermore, synthesis-based *HFR* approaches can benefit from the use of a pre-trained FR model, which eliminates the requirement to train the model with a huge amount of training data.

However, this technique falls short when it comes to actual use cases. The synthesize-based *HFR* must first generate an RGB image from the target modality image before passing it through a FR model to retrieve the embeddings at inference time. The synthesis process adds a significant amount of computing cost, which may restrict its usefulness in practical deployment scenarios. Moreover, synthesis-based (generative) algorithms are often trained to generate RGB images that are optimized for both perceptual and identity loss [26] (along with several other metrics). However, because the generated pictures are utilized in conjunction with a neural network, the perceptual quality of the generated images may not be significant in the context of the *HFR* task. In other words, it is critical to retain discriminative traits that potentially match those in the source class rather than generating high-fidelity images. This is particularly important as the generation process is often a much harder problem to solve when the amount of paired training data is limited.

III. Proposed Method

We follow the definitions in [13], [34] to formalize the*HFR* task.

A. Formal definition of HFR

Consider a domain $\mathcal{D}$ with samples $X \in \mathbb{R}^d$ and a marginal distribution $P(X)$ (with dimensionality-d). The task of an FR system $\mathcal{T}^{fr}$ can be defined by a label space Y whose conditional probability is $P(Y|X,\Theta)$, where X and Y are random variables and Θ defines the model parameters. In the training phase of an FR system, $P(Y|X,\Theta)$ is typically learnt in a supervised fashion given a dataset of faces $X = \{x_1, x_2, ..., x_n\}$ together with their identities $Y = \{y_1, y_2, ..., y_n\}$.

Now consider the following heterogeneous face recognition (*HFR*) problem. Here we assume that we have two domains, source domain $\mathcal{D}^s = \{X^s, P(X^s)\}$ and target domain $\mathcal{D}^t = \{X^t, P(X^t)\}$ sharing the labels Y.

In a broad sense, the task in the *HFR* problem $\mathcal{T}^{hfr}$ is to find a $\hat{\Theta}$, where $P(Y|X^s,\Theta) = P(Y|X^t,\hat{\Theta})$. The form and the nature in which $\hat{\Theta}$ is estimated varies with different *HFR* approaches.

B. Proposed approach

In the proposed approach, let us first assume that the samples from both domains, $X_s = \{x_1, x_2, ..., x_n\}$ and $X_t = \{x_1, x_2, ..., x_n\}$ from $\mathcal{D}^s$ and $\mathcal{D}^t$ with the shared set of labels $Y = \{y_1, y_2, ..., y_n\}$ are available. Also, assume that the parameters of an FR model Θ (we denote it as Θ_{FR} in the following discussion) for the (VIS) model is available from $\mathcal{D}^s$. In our case, Θ_{FR} is essentially the parameters of a pre-trained FR model trained using visible spectrum images. Following the discussion on the synthesis based HFR, we hypothesize that a module with a learnable set of parameters θ_{PDT} can transform the target domain image to a new representation ($\hat{X^t} = \mathcal{F}_{\mathcal{PDT}}(X^t)$) to reduce the domain gap while retaining discriminative information. This new representation ($\hat{X^t}$) can be used together with a pre-trained FR model to achieve the *HFR* task.

To accomplish this task, we propose to prepend a small network module called "Prepended Domain Transformer " (PDT) to a pre-trained FR model. A schematic diagram of the proposed framework is depicted in Fig. 2. Essentially, we apply this module as a transformation to the target modality images, which generates a *transformed* ($\mathcal{F}_{\mathcal{PDT}}(X^t)$) image.

This can also be viewed as an extension of the DSU [13] approach. Instead of adapting the lower layers as in DSU, we prepend a neural network block in front of a pre-trained FR model to handle domain-specific features. However, one of the main restrictions of DSU is that the local features that could be modified are restricted by the design of the face recognition (FR) network. One can only try to adapt or freeze layers already present in the FR network. It should also be noted that the architecture of the low-level layers in the FR network is designed for optimizing the performance in visible spectrum face recognition, and as such the lower layers (and their architecture) may not be optimal for the *HFR* task. This creates a bottleneck that limits the possibility of learning. In Prepended Domain Transformer there is more flexibility by changing the architecture of the PDT block, one can even change the local receptive field by changing the architecture of the PDT block. One could even perform a neural architecture search [35] to optimize the architecture of the PDT block for a specific heterogeneous use case. The architecture we use is more general which fits a larger set of heterogeneous tasks. In this sense, PDT is more flexible compared to DSU. Moreover, in PDT, the transformation is performed in the pixel space itself, meaning this approach can be applied to several FR architectures as a plug-in module. We have reimplemented DSU heterogeneous face recognition approach with the recent Iresnet100 pre-trained model as an additional baseline (DSU-Iresnet100).

The *transformed* image can then be passed to a pre-trained FR model to get the embeddings for the *HFR* task. With the proposed approach, we can express the *HFR* problem in the following way:

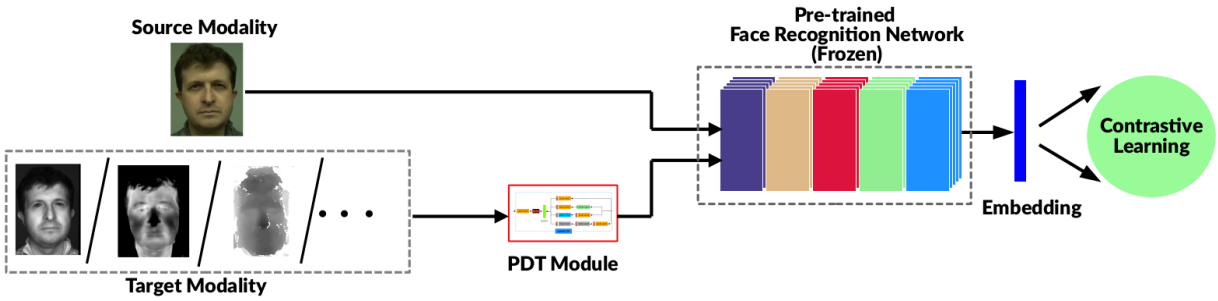

Fig. 2. Schematic diagram of the proposed framework. Target domain images are transformed using the proposed Prepended Domain Transformer (PDT) block. The PDT block prepended to the frozen FR model is trained in a standard Siamese setting with contrastive loss.

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

The parameters of PDT block (θ_{PDT}) can be learned in a supervised setting using back-propagation. In the forward pass for a tuple (X^s, X^t), the X^s image directly passes through the shared pre-trained FR network to produce the embedding. The target image (X^t) first passes through the PDT module ($\hat{X^t} = \mathcal{F}_{\mathcal{PDT}}(X^t)$), and then the *transformed* image passes through the shared pre-trained FR model to generate the embedding. Contrastive loss [36] is employed as the loss function in the training phase, to reduce the distance between these cross-modal embeddings when the identities are the same and increase the distance when the identities are different. The Contrastive loss is given as:

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

Where Θ denotes the weights of the network, X_s, X_t denote the heterogeneous pairs and Y_p the label of the pair, i.e., whether they belong to the same identity or not, m is the margin, and D_W is the distance function between the embeddings of the two samples. The label $Y_p = 0$, when the identities of subjects in X_s and X_t are the same, and $Y_p = 1$ otherwise. The distance function D_W can be computed as the Euclidean distance between the features extracted by the network.

The parameters of the shared FR model are kept frozen during the training and only the parameters of the PDT module are updated in the backward pass. At the end of the training, the model corresponding to minimum validation loss is selected which is used for the evaluations.

C. Architecture of the Prepended Domain Transformer (PDT) block

The Prepended Domain Transformer block is parameter-efficient and generic, allowing it to be applied to a wide range of *HFR* scenarios. The input and output of the PDT block are 'three-channel' images with the same size. This makes it easy to visualize the output of the proposed PDT module and to pass the modified images on to pre-trained FR models during inference. This module can also be readily "plugged in" to any pre-trained FR pipeline to convert it to an *HFR* pipeline.

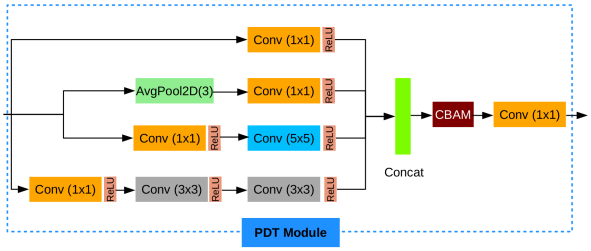

Fig. 3. Architecture of the Prepended Domain Transformer (PDT) block.

The architecture of the proposed PDT module is shown in Fig. 3. The initial part of the PDT bock utilizes the design principles inspired from the inception architecture [37]. We follow the idea of multi-scale processing by using different parallel branches with different kernel sizes. Parallel branches were necessary since the receptive field required for various heterogeneous settings differs, and having multi-scale features at the input level aids in a generic design with minimal computational complexity. There are four parallel paths 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. In each of these branches, 1×1 convolutions are used to reduce the number of output channels. We use ReLU activation after each of the convolution operations. A feature map is formed by concatenating

the outputs from each of these branches, which consist of features obtained using filters with different receptive fields. The addition of an attention mechanism helps the network in deciding "what" and "where" to focus. A Convolutional Block Attention Module (CBAM) [38] attention module was added which achieves this in a simple and parameter efficient way. The CBAM block acts on a feature map along the channel as well as the spatial dimension in a sequential manner. The attention maps obtained are multiplied by the input feature map. The addition of the CBAM module helps in focusing on meaningful features along the channel and spatial dimensions making the proposed architecture robust to a wide variety of *HFR* scenarios. After the CBAM block, the channel dimension of the output feature map is still high and a 1×1 convolutional layer is added to reduce the channel dimension to three.

Overall the number of parameters to learn is merely 1.4K. The minimal design enables the network to focus on important features with a minimal parameter overhead. It is to be noted that, this module can be further optimized for specific heterogeneous scenarios.

D. Pre-trained FR backbone

As described earlier, the PDT module can be prepended to any pre-trained FR model. Though we perform most of the experiments with *Iresnet100* model [39], this can be extended to many publicly available pre-trained FR models. For the sake of reproducibility, we used publicly available pre-trained face recognition models available from [40] [2]. These models were trained on MS-Celeb-1M-v1c [3] with 72,778 identities and about 3.28M images. In most cases, the pre-trained FR model accepts three-channel images with a resolution of 112×112. Faces are first aligned and cropped ensuring eye center coordinates fall on pre-fixed points. In the case of single-channel inputs (such as NIR, thermal, etc.), we replicate the same channel to three channels without making any changes to the network architecture. This was necessary since the pre-trained networks were designed to accept three-channel RGB images, and we didn't want to alter the layers/weights of the FR network that would change the performance of the pre-trained network in the RGB images in any manner.

E. Implementation details

The proposed framework is trained with Contrastive loss in a standard Siamese network setting [36]. The margin parameter is set as 2.0 in all the experiments. We used Adam Optimizer with a learning rate of 0.001 and trained the model for 20 epochs with a batch size of 90. The framework was implemented in PyTorch using the Bob library [41]. Except for the new PDT module added to the target channel branch, the whole pre-trained FR model is shared between source and target modalities in the Siamese network. Only the parameters of the PDT module are adapted during training, while the weights of the FR model remain frozen. The experiments are reproducible and the source code and the protocols will be made available publicly [4].

The proposed method may be applied to a variety of *HFR* scenarios, including VIS-Thermal, VIS-SWIR, and VIS-Low resolution VIS, and so on. Furthermore, components of the proposed framework and the training routine are kept simple to demonstrate the efficacy of the proposed approach.

IV. Experiments

An extensive set of experiments and ablation studies with the proposed approach are presented in this section. Primarily, we have evaluated VIS-Thermal *HFR* performance in four different datasets. We also compare the performance of the proposed approach against other heterogeneous settings such as VIS-NIR, VIS-SWIR, and so on. We further perform several studies evaluating the amount of data required for training the models and also the performance with different FR architectures.

A. Databases and Protocols

The datasets used in the evaluations are described in the following section.

Polathermal dataset: Pola Thermal dataset [42] – Polarimetric and Thermal Database is an *HFR* dataset collected by the U.S. Army Research Laboratory (ARL). The dataset contains polarimetric LWIR (long-wave infrared) imagery together with color images collected synchronously for 60 subjects. The dataset contains thermal imagery collected for conventional thermal images as well as polarimetric images. In this work, we use the conventional thermal images for our experiments using the reproducible protocols introduced in [13]. We follow the same five-fold partitions in which 60 subjects were split into a training set with 25 identities and 35 identities for testing. To compare different methods, the average Rank-1 identification rate is reported from the evaluation set of the five folds.

Tufts face dataset: The Tufts Face Database [43] provides face images captured with different modalities for the *HFR* task. Specifically, we use the thermal images provided in the dataset to evaluate VIS-Thermal *HFR* performance. Overall, there are a total of 113 identities

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

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

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

comprising of 39 males and 74 females from different demographic regions. For each subject, images from different modalities are available. For comparison purposes, we follow the procedure followed by authors in [26], 50 identities are randomly selected from the data as the training set and the remaining subjects were used as the test set. We report the Rank-1 accuracies and Verification rates at false acceptance rates (FAR) 1% as well as 0.1% for comparison.

ARL-VTF dataset: In [5], authors made available the DEVCOM Army Research Laboratory Visible-Thermal Face Dataset (ARL-VTF). The dataset contains heterogeneous data from 395 subjects with three visible spectra as well as one thermal (long-wave infrared- LWIR) camera, with over 500,000 images altogether. The dataset contains variability in terms of expressions, pose, and eyewear. We evaluate the models with the protocols originally provided with the dataset. The dataset also provides annotations for face landmarks. Several protocols evaluating the effects of the pose, expressions, and eyewear are also provided with the dataset. The $test$ set for each setting is fixed to enable direct comparisons with state-of-the-art methods. The naming of each protocol is as follows: 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.

In summary, the structure of protocol names are as follows:

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

A detailed description of these protocols can be found in [5].

CASIA NIR-VIS 2.0 dataset: The CASIA NIR-VIS 2.0 Face Database [44] provides images of subjects captured with both visible spectrum as well as near-infrared lighting, with a total of 725 identities. Each subject in the dataset has 1-22 visible images and 5-50 near-infrared (NIR) images. The experimental protocols provided uses a 10-fold cross-validation protocol with 360 identities used for training. The gallery and probe set for evaluation consist of 358 identities. The train and test sets are made with disjoint identities. We perform experiments in each fold and the mean and standard deviation of the performance metrics are reported.

SCFace dataset: The SCFace [45], dataset contains high quality mugshot for enrollment for FR. The probe samples correspond to surveillance scenarios coming from different cameras and are of low quality. Depending on the distance and quality of probe samples, four different protocols are present. They are close, medium, combined, and far. The "far" protocol is the most challenging one. The dataset contains 4,160 static images (in the visible and infrared spectrum) from 130 subjects.

MCXFace Dataset: We present a new *HFR* dataset named Multi-Channel Heterogeneous Face Recognition dataset (MCXFace). The dataset is derived from the HQ-WMCA dataset we have created earlier [46], [47]. The dataset contains images of 51 subjects collected in different channels under three different sessions and various illumination conditions. The channels available are color (RGB), thermal, near-infrared (850 nm), short-wave infrared (1300 nm), Depth, Stereo depth, and depth estimated from RGB images using 3DDFA [48] method. All the channels are registered spatially and temporally across all the modalities. The details about the sensors and data collection sessions can be found in our earlier work [46], [47]. In the MCXFace dataset, only bonafide samples are present. Further, the files are divided into $train$ and dev sets with a disjoint set of identities to make experiments in different homogeneous and heterogeneous settings possible. For each of the protocols, we have created five different folds, by randomly dividing the subjects in $train$ and dev partitions. Each of the protocol names is of the following form: $< SOURCE > - < TARGET > -split < split >$. In addition to the images, annotations for left and right eye centers for all the images are also provided. The dataset will be available publicly in the following link [5].

CUFSF dataset: The CUHK Face Sketch FERET Database (CUFSF) [23] contains 1194 faces from the FERET dataset [49] and each image in the FERET dataset has a corresponding sketch image drawn by an artist. The dataset is challenging since the sketches have more shape exaggerations compared to the source photos. We follow the same protocol as reported in [50], where 250 identities were used for training the model, and the rest of 944 identities are reserved as the testing set. The Rank-1 accuracies are reported for comparison.

B. Metrics

To evaluate the models we follow several different metrics corresponding to previous literature. We have used a subset of metrics from the following performance metrics, Area Under the Curve (AUC), Equal Error Rate

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

(EER), Rank-1 identification rate, Verification Rate with different false acceptance rates (0.01%, 0.1%, 1%, and 5%).

C. Experimental results

The experiments performed in the different datasets and the results are discussed in this section.

*1) **Experiments with Polathermal dataset**:* Here we perform experiments in thermal to visible recognition scenarios. The results in Table. I shows the average Rank-1 identification rate in the five protocols of the Polathermal 'thermal to visible protocols' (using the reproducible protocols in [13]). The reimplemented DSU-Iresnet100 baseline achieves better results compared to the original model from [13], indicating that the use of a better pre-trained model improves the results. It can be seen that the proposed approach achieves an average Rank-1 accuracy of 97.1% with a standard deviation of 1.3%, which is much greater than the results from other baselines reported in the literature.

TABLE I
POLA THERMAL - AVERAGE RANK ONE RECOGNITION RATE

Method	**Mean (Std. Dev.)**	**Info**
DPM in [42]	75.31 % (-)	Paper
CpNN in [42]	78.72 % (-)	Base-
PLS in [42]	53.05% (-)	lines
LBPs + DoG features in [18]	36.8% (3.5)	Repro-
ISV in [51]	23.5% (1.1)	ducible
GFK in [52]	34.1% (2.9)	baselines
DSU(Best Result) [13]	76.3% (2.1)	Reproducible
DSU-Iresnet100	88.2% (5.8)	Reproducible
PDT (Proposed)	**97.1% (1.3)**	Reproducible

*2) **Experiments with Tufts face datasets**:* Table. II shows the performance of the proposed approach against other state-of-the-art methods in the VIS-Thermal protocol of the Tufts face dataset. The Tufts face dataset is very challenging due to pose and other types of variations. The challenging pose variations in Tufts face dataset are depicted in Fig. 4. Performance of even visible spectrum face recognition systems degrade in such extreme yaw angles, so it is expected that the performance of *HFR* would also degrade. Nevertheless, it can be seen that the proposed approach achieves the best results in verification rate, and is only second to DVG-Face [26] in Rank-1 accuracy, showing the effectiveness of the proposed approach.

*3) **Experiments with CASIA-VIS-NIR 2.0 dataset**:* Though most of the focus is given to Thermal-VIS *HFR*, we perform experiments in the CASIA-VIS-NIR 2.0 dataset to showcase the effectiveness of the proposed approach in other heterogeneous scenarios, specifically

Fig. 4. Challenging pose variations in VIS-Thermal protocol of the TUFTS face dataset.

TABLE II
EXPERIMENTAL RESULTS ON VIS-THERMAL PROTOCOL OF THE TUFTS FACE DATASET.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
LightCNN [53]	29.4	23.0	5.3
DVG [54]	56.1	44.3	17.1
DVG-Face [26]	**75.7**	68.5	36.5
DSU-Iresnet100	49.7	49.8	28.3
PDT (Proposed)	65.71	**69.39**	**45.45**

NIR-VIS recognition. As it can be seen from the baselines, the domain gap appears to be less in this scenario and even some of the pre-trained FR model trained using VIS modality achieves reasonable performance. Due to this, the evaluations are done with even tighter thresholds and VR@FAR=0.1% and VR@FAR=0.01% are used for comparison. There are 10 sub-protocols in the dataset, and we report the mean and standard deviation of all ten folds in the comparison. The results are compared in Tab. III. From the results, it can be seen that the proposed approach achieves the best performance compared to other state-of-the-art methods. The superior performance indicates that our framework generalizes across several heterogeneous scenarios.

TABLE III
EXPERIMENTAL RESULTS ON CASIA NIR-VIS 2.0.

Method	Rank-1	VR@FAR=0.1%	VR@FAR=0.01%
IDNet [55]	87.1±0.9	74.5	-
HFR-CNN [56]	85.9±0.9	78.0	-
Hallucination [57]	89.6±0.9	-	-
TRIVET [58]	95.7±0.5	91.0±1.3	74.5±0.7
W-CNN [59]	98.7±0.3	98.4±0.4	94.3±0.4
PACH [60]	98.9±0.2	98.3±0.2	-
RCN [61]	99.3±0.2	98.7±0.2	-
MC-CNN [62]	99.4±0.1	99.3±0.1	-
DVR [63]	99.7±0.1	99.6±0.3	98.6±0.3
DVG [54]	99.8±0.1	99.8±0.1	98.8±0.2
DVG-Face [26]	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) **Experiments with SCFace dataset**:* We have performed a set of experiments on the SCFace dataset using the protocols provided for visible images in the dataset. In this dataset, the heterogeneity arises from the quality difference between gallery and probe images, i.e., gallery images are high-resolution mugshots and probe images and low-resolution images coming from a surveillance camera. The results are tabulated in Table. IV, the values reported corresponds to the "evaluation" set of the protocols. The baseline model is a pre-trained *Iresnet100* model as used in other experiments, and

the rows with PDT denotes the proposed model where the PDT model is trained using contrastive training. Results with reimplemented DSU-Iresnet100 model is also added for comparison. It can be seen that with the proposed approach the performance of the baseline model improves. The improvement is obvious in the "far" protocol where the quality of the probe images is very poor. It can be seen the PDT module, in this case, helps in learning quality and resolution invariant features improving the results.

TABLE IV
PERFORMANCE OF THE PROPOSED APPROACH IN THE SCFACE DATASET, THE BASELINE IS A PRETRAINED *Iresnet100* MODEL, AND THE PDT IS WITH 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

*5) **Experiments with MCXFace dataset**:* The experiments in the MCXFace dataset gives a unique opportunity to evaluate the performance of the models in several heterogeneous scenarios, including VIS-Thermal, VIS-Depth, VIS-SWIR, VIS-NIR, and so on. The results containing the average performance among five folds are shown in Table. V. For each modality, (eg. VIS-Thermal), the values reported are aggregated over the five folds in the dataset. In each protocol, we perform a baseline evaluation which is essentially using the pre-trained *Iresnet100* FR model on the new modality. The lower baseline performance indicates a large domain gap. For example, running a vanilla FR model itself achieves excellent performance for NIR and SWIR channels, whereas the performance is poor for depth and thermal channels. In the Table. V, the rows with PDT show the results with our proposed approach. In addition, the reimplemented DSU-Iresnet100 model is also added as a baseline. It can be seen that the performance improves greatly in the case of the thermal channel. However, for the depth channel, even though the performance improves, the final performance is not satisfactory, it indicates that the depth channel needs a different treatment, and using the depth data as range images may not be the optimal choice. Using representations like normals or point clouds could be well suited for the depth modality.

TABLE V
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

*6) **Experiments with ARL-VTF dataset**:* The ARL-VTF offers a large-scale VIS-Thermal dataset with different variations such as pose, expressions, and eyewear making it possible to perform experiments to evaluate the effect of different factors. We have used the standard protocols shipped with the datasets for our evaluations. The test sets of different protocols are fixed and the best model selected from the cross-validation was used to evaluate the performance. The performance of the proposed approach against the state-of-the-art methods is shown in Table. VI. It can be seen that the proposed approach achieves state-of-the-art performance in most cases. The performance improvement is very clear in challenging protocols with pose variations (P_TP-). In the $P\_TB0 - G\_VB0-$ protocol, where all the images are frontal, the proposed method achieves a verification rate of 98.57% for FAR1% (and a corresponding Rank-1 accuracy of 99.14%), whereas in the P_TP0-G_VB0- protocol where the samples contain pose variations the verification rate drops to 60.80% (and a corresponding Rank-1 accuracy of 60.23%). The Rank-1 accuracy for just frontal faces is 99.14% whereas with pose variations it is 60.23%. It can be seen that this drop is present in all other methods as well, indicating that *HFR* also suffers performance degradation with pose variations as with homogeneous face recognition.

*7) **Experiments with CUFSF dataset**:* Here we perform experiments with the challenging sketch to photo recognition task. The Rank-1 accuracies obtained with the baseline and the other methods are shown in Table. VII, using the protocols in [50]. It can be seen that the proposed approach obtains a Rank-1 accuracy of 71.08%. Nevertheless, the absolute accuracy in the sketch to photo recognition is low. Several methods in literature have shown (on a different protocol) that the sketch recognition performance could be improved with specifically designed features [68] and specially designed neural network models [69]. The sketch modality is very different compared to other heterogeneous imaging modalities we have considered so far like thermal, near-infrared, and so on. Though the CUFSF dataset contains viewed hand-drawn sketch images [70] which appears to

TABLE VI
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 [64]	71.12	33.80	6.95	21.28	75.22	30.42	8.28	27.63
	GANVFS [65]	97.94	8.14	75.00	88.93	98.58	6.94	79.09	91.04
	Di *et al.* [66]	99.28	3.97	87.95	96.66	99.49	3.38	90.52	97.81
	Fondje *et al.* [67]	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 [64]	68.77	38.02	6.69	20.28	52.11	48.88	2.22	4.66
	GANVFS [65]	99.36	3.77	84.88	97.66	87.34	18.66	7,00	29.66
	Di *et al.* [66]	99.63	2.66	91.55	98.88	89.24	19.49	16.33	41.22
	Fondje *et al.* [67]	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 [64]	69.10	35.98	7.01	16.44	73.97	31.87	7.93	19.60
	GANVFS [65]	96.81	10.51	70.41	84.00	97.73	8.90	74.20	86.80
	Di *et al.* [66]	98.46	6.44	81.11	92.49	98.89	5.60	84.23	93.94
	Fondje *et al.* [67]	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 [64]	68.78	36.24	7.75	18.06	51.05	49.11	2.26	4.95
	GANVFS [65]	98.66	5.93	73.17	92.82	83.68	22.41	6.77	22.13
	Di *et al.* [66]	99.30	3.84	82.55	97.44	86.12	21.68	9.88	31.62
	Fondje *et al.* [67]	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 [64]	54.86	47.22	3.13	9.78	56.50	46.03	4.01	10.84
	GANVFS [65]	63.70	41.66	16.55	23.73	65.58	40.19	17.95	25.68
	Di *et al.* [66]	65.06	40.24	17.33	24.56	67.13	38.67	18.91	26.46
	Fondje *et al.* [67]	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 [64]	54.31	47.04	2.93	8.44	50.08	49.67	0.60	4.33
	GANVFS [65]	65.79	40.35	17.84	25.48	59.51	44.04	4.29	15.47
	Di *et al.* [66]	67.27	39.00	18.16	26.02	60.10	43.57	5.77	15.97
	Fondje *et al.* [67]	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 [64]	59.68	41.72	3.33	3.33	67.08	36.44	2.68	11.11
	GANVFS [65]	87.61	20.16	20.55	44.66	96.82	8.66	46.77	83.00
	Di *et al.* [66]	91.11	17.43	22.33	55.66	97.96	7.21	60.11	88.70
	Fondje *et al.* [67]	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

be holistically similar for humans, there exists a domain gap in the context of an automatic face recognition system as evidenced from the baseline performance. The pretrained model achieves a Rank-1 accuracy of 56.57%, indicating the domain gap. The modalities like thermal, NIR, and SWIR, were all the "imaging" modalities and it shares the same high-level representation of the face, but a different aspect of the face. In sketch recognition, the sketch images contain exaggerations depending on the artist, and may not optimally preserve the discriminative information which a face recognition network might be looking for. This could explain the larger performance gap of sketch-photo recognition compared to the performance in other imaging modalities.

D. Analysis of the Framework

To further understand the effectiveness of the proposed approach, we have performed additional experiments on the ARL-VTF dataset due to the large number of subjects present. Specifically, we performed experiments to understand the effect of the amount of training data, the type of supervision used in training, and the performance

TABLE VII
CUFSF: RANK-1 RECOGNITION IN SKETCH TO PHOTO RECOGNITION

Method	Rank-1
Baseline	56.57
IACycleGAN [50]	64.94
DSU-Iresnet100	67.06
PDT (Proposed)	**71.08**

of different FR architectures. All these experiments were performed on the G_VB0-P_TB- protocol of the ARL-VTF dataset.

*1) **Performance with limited amount training data**:* The amount of paired training data available to train *HFR* model is often limited and expensive to acquire. In this regard, we perform a set of experiments to understand how the amount of training data available to train the model affects its performance. We used the ARL-VTF dataset for this set of experiments because it had a larger number of subjects. The test samples are kept the same for this set of experiments, with the only difference being the amount (or percentage) of training and validation samples. We start with using 100% of the training data for training the PDT module and gradually lower the number of samples in the intervals of 10%, and finally 1% intervals. For context, we also note the number of subjects in the training set for these scenarios. The results of this set of experiments are tabulated in Table. VIII. Remarkably, the proposed approach achieves a Rank-1 accuracy of 94.67% with just 2% percentage of the training data, for context, just with data from 4 subjects. This could be due to the parameter efficiency of our approach. The learnable component of the PDT block contains approximately just 1.4K parameters, and hence requires a very minimal amount of data to achieve good performance. This is an important observation, as *HFR* datasets are often small in size.

*2) **Experiments with unpaired images**:* So far in the experiments, the networks were trained using contrastive loss in a supervised manner. We assume that paired heterogeneous samples are available at the training time. Here we try to emulate an unpaired setting, meaning we do not have paired heterogeneous samples, and we do not have the identity information of the samples. In this scenario, we supervise our framework to match the feature distributions of source and target modalities using Maximum Mean Discrepancy (MMD) [71] loss. We have experimented with applying the MMD loss in three different ways, 1) the MMD loss was applied to both the *transformed* and source images, thereby attempting to match the FR network's inputs.(ip), 2) The MMD loss is applied to the embeddings obtained from the FR network (op). and 3) where both output embeddings and inputs were supervised by MMD loss (ip+op).

TABLE VIII
EXPERIMENTS WITH SUBSETS OF TRAINING DATA. THE TEST SET IS KEPT THE SAME IN 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

The results obtained from this set of experiments are shown in Table. IX. The first row shows the baseline with a pre-trained *Iresnet100* and the last row shows the results with supervised learning with the PDT approach. From the results, it can be seen that even in unpaired settings the proposed framework performs reasonably well. The best results are obtained when MMD is used on both the input and output. This indicates that the proposed framework can be employed even if paired samples aren't available. It was possible to achieve reasonable performance just by matching the distributions of the source and target modality features. However, as previously discussed, having a small number of labeled samples improves performance significantly when compared to using unpaired samples in training.

TABLE IX
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

*3) **Experiments with different Face Recognition models**:* We have used *Iresnet100* as the pre-trained FR model for all the experiments discussed in the previous section. In this section, we investigate whether the proposed approach generalizes to other FR architectures. We again perform these experiments in the ARL-VTF dataset using the G_VB0-P_TB- protocol. For each

TABLE X
COMPARISON WITH DIFFERENT ARCHITECTURES FOR THE HFACE 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

experiment, we just switch the pre-trained FR model in the PDT framework. In addition to *Iresnet100*, we have used several other pre-trained FR models which were available publicly [6].

We first run these pre-trained models without the PDT module to get a baseline performance. The results with the trained PDT module are also shown in Table. X. From the table, it can be seen that the proposed approach works well with all the architectures used. The complexity and performance of each of these models are different. From these results, it is clear that the proposed approach can be used with any FR model architecture, given the PDT module is trained together with it. Pre-trained FR models can thus be chosen for *HFR* based on a tradeoff between accuracy and computational complexity.

*4) **Generalizability of the PDT weights across architectures**:* In the previous sub-section, it has been shown that the proposed approach works with different FR models despite the differences in architecture and complexity. It is to be noted that, the architecture of the PDT remains the same even though we used different FR architectures. We further investigated whether the PDT module learned for one FR architecture would work for another. These experiments could provide insights into the learned transformations and could also indicate whether PDT modules could be used with black-box models. For example, in Table. XI, AttentionNet(C) (column)- EfficientNet(A) (row), means that the PDT was added to AttentionNet model during training, the weights of the PDT module trained with AttentionNet (checkpoint) was used in conjunction with EfficientNet architecture for evaluation. From the results in Table. XI, it can be seen that the proposed approach works well for most of the scenarios with the notable exception of GhostNet [72], which could be due to GhostNet's unique architecture. GhostNet modules try to reduce the computational complexity of a network by reducing the filters needed to compute redundant feature maps. This is achieved by creating additional feature maps using cheap linear operations such as depth-wise separable convolutions on the output feature maps. The original feature maps and the newly created (ghost) features are concatenated in the ghost module to maintain the feature map sizes. We suspect that this operation makes the PDT learned together with GhostNet specifically optimized to the filters in the pre-trained backbone. This results in the low performance of the PDT module trained with the GhostNet module when it is used with other architectures. In summary, while the PDT module works in most cases, the parameters of the PDT module are ultimately learned in conjunction with a specific architecture, and it may not always generalize when used with black-box models. However, using multiple FR models to learn the PDT weights in a distillation framework could improve the performance in the case of black-box FR models. This is left as future work and is beyond the scope of the current contribution.

*5) **Visualization of the intermediate features**:* The t-SNE plots of thermal and visible images at various stages in the pipeline are shown in Fig. 5. The thermal and visual modalities are readily distinguished in the t-SNE plots at the input stage. The heterogeneous samples appear to form different distributions even after the PDT stage. As we progress in the layers in the feature maps of the CNN, the distribution of thermal and visible features becomes more aligned. Towards the last layers and the final embedding output, the modalities align very well, i.e., the embeddings of the same identity in the thermal and visible spectrum grow closer in the embedding space, which is exactly what we need for the *HFR* task. One important observation is that even after the PDT block the distribution of features is disjoint.Despite the distribution disparity, the *transformed* image following the PDT block helps align the embeddings in the final layers.

Since the feature maps after the PDT module are also three-channel images with the same resolution, they can be used to inspect the PDT module's output. For example, Fig. 6, shows the thermal, visible, and *transformed-thermal* image of the same person. In a broad sense, the face embeddings extracted from the *transformed* images and visible spectrum images have a higher match score. Intuitively, one would expect the *transformed* image to resemble the VIS image in order for this to occur; however, the results show that this is not the case; instead, the *transformed* image only needs to preserve discriminative information for the *HFR* task.

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

TABLE XI
RANK-1 ACCURACIES FOR THE CROSS-TEST BETWEEN DIFFERENT ARCHITECTURES, ROWS ARE THE ARCHITECTURE USED FOR EVALUATION (A), AND COLUMNS ARE THE ARCHITECTURE USED FOR TRAINING THE PDT MODULE (C), RESULTS FOR THE SAME TRAIN-TEST ARCHITECTURE ARE HIGHLIGHTED 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

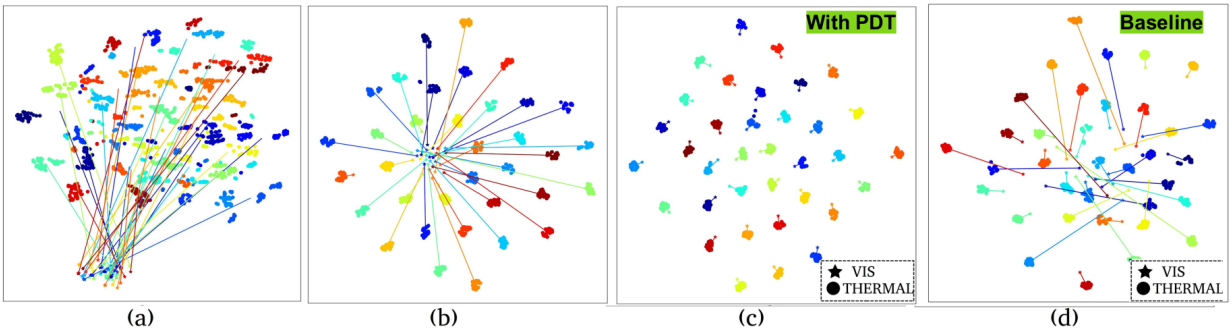

Fig. 5. t-SNE plots for visible and thermal images from various stages, different colors indicate different identities. For the reference modality, only one cluster center is shown for clarity. The lines connect the cluster center of visible and thermal images for each identity. a) shows visible images and transformed thermal images in pixel space, b) from an intermediate feature map of the CNN, c) final embedding space with PDT, and d) final embedding space without PDT (baseline). It can be seen that the identities match in the final embedding space with the PDT block added.

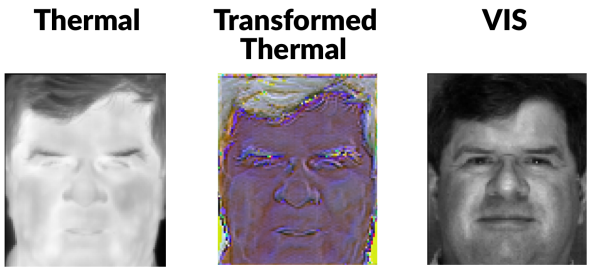

Fig. 6. A visualization of thermal to vis *HFR* scenario in Polathermal dataset, the *Transformed-Thermal* is the intermediate output from the PDT module, even though this image doesn't look visually similar to the VIS image, the embedding obtained from the *transformed* image produces a high match score with the embedding extracted from the VIS image.

V. DISCUSSIONS

As opposed to the computationally expensive synthesis based methods, the proposed approach introduces a simple, computationally simpler, plug and play module for heterogeneous face recognition that can be trained with a minimal number of samples. The proposed approach can be used even in scenarios where paired samples are not available. The PDT approach achieves this by learning a parameter efficient transformation in the pixel domain that reduces the domain gap while retaining the discriminative information.

The proposed approach achieves state-of-the-art performance in many challenging benchmarks, as demonstrated by experiments in different thermal to VIS as well as other heterogeneous protocols. The new PDT module added is parameter efficient and is generic enough for several heterogeneous scenarios. As evidenced from the experiments in Table. VIII, the amount of training data required for our framework is very small, this makes it suitable for the real-world data starved heterogeneous settings. The approach was also found to perform well even in the absence of paired samples, this is of practical importance in heterogeneous scenarios where paired data is not readily available. The approach itself is modular and can be added to any pre-trained FR model. To summarize, we show that prepending a learnable neural network module to a pre-trained FR model yields state-of-the-art performance in a variety of challenging *HFR* scenarios.

A. Limitations and Future directions

Currently, the architecture of the PDT block is designed to be general so as to serve a wide range of

heterogeneous scenarios. The current design of PDT provides flexibility in terms of the receptive field due to the multi-branch architecture and CBAM module. This simple approach achieves comparable, and in many cases outperforms many computationally expensive state-of-the-art models. The receptive field of the branches in PDT is a design parameter that can be optimized even further. It could be possible to optimize the architecture of PDT for a specific heterogeneous scenario, which might boost the performance even further. Since the Prepended Domain Transformer framework depends on the performance of the pre-trained network used in the framework, the overall performance would be impacted by the performance of the pre-trained network. For example, most of the standard pre-trained FR models struggle with extreme yaw angles and profile faces, meaning such a network would result in lower *HFR* performance when there is extreme yaw angles present (as we see from the results in Tufts face dataset). Nevertheless, the proposed approach is generic enough to be applied to newer and robust face recognition models. Further, the PDT approach is better suited for heterogeneous "imaging" modalities, as they share more structural similarities with visible face images. Heterogeneous modalities like drawn sketches do not satisfy this criterion and may be harder to adapt in the image domain. Methods like generative approaches might be better suited to these scenarios. Though the PDT blocks trained with one architecture do not necessarily work well with other architectures, this could be enhanced to work with black-box models by employing multiple models in a teacher-student fashion. The proposed approach can also be combined with GAN-based generation methods, and can also be improved by using triplet or quadruplet loss functions. The proposed approach can be further extended with more tuning of the PDT architecture with the likes of neural architecture search [35], training schedules, data augmentation, triplet training, and so on. However, as the title suggests, we hope that this approach will serve as a simple yet strong baseline motivating further research in heterogeneous face recognition.

VI. Conclusions

In this work, we have proposed a simple yet effective framework for heterogeneous face recognition. Essentially, to convert a pre-trained FR model to an *HFR* network we just prepend a novel neural network module for the target modality. The new module, called Prepended Domain Transformer (PDT) is parameter efficient and does not require a lot of samples to train. Since the method performs well across a variety of FR architectures, it can be used to convert any face recognition model to a heterogeneous one. The design of the framework, training schedule, loss functions, and parameter selection are intentionally left simple to demonstrate the efficacy of the proposed approach. These can be further tuned to improve the results further. The proposed approach was found to outperform state-of-the-art methods in many challenging heterogeneous datasets. Further, we also introduce MCXFace heterogeneous face recognition dataset which contains multiple modalities which can be used for *HFR* evaluations. Lastly, the source code, protocols, and datasets used are publicly available to make further extensions of the work possible.

Acknowledgment

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

References

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

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

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

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

[5] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman, X. Di, C. N. Fondje, V. M. Patel, N. J. Short, B. S. Riggan, N. M. Nasrabadi *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*, 2021, pp. 1559–1568.

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

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

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

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

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

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

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

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

[14] U. Cheema, M. Ahmad, D. Han, and S. Moon, "Heterogeneous visible-thermal and visible-infrared face recognition using cross-modality discriminator network and unit-class loss," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[17] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "Tv-gan: Generative adversarial network based thermal to visible face recognition," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 174–181.

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

[19] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning multi-scale block local binary patterns for face recognition," in *International Conference on Biometrics*. Springer, 2007, pp. 828–837.

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

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

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

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

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

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

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

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

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

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

[31] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

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

[33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

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

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

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

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

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

[40] J. Wang, Y. Liu, Y. Hu, H. Shi, and T. Mei, "Facex-zoo: A pytorch toolbox for face recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3779–3782.

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

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

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

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

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

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

[47] Z. Mostaani, A. George, G. Heusch, D. Geissenbuhler, and S. Marcel, "The high-quality wide multi-channel attack (hq-wmca) database," Idiap, Idiap-RR Idiap-RR-22-2020, 9 2020.

[48] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "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*. Springer, 2020, pp. 152–168.

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

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

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

[52] A. F. Sequeira, L. Chen, J. Ferryman, P. Wild, F. Alonso-Fernandez, J. Bigun, K. B. Raja, R. Raghavendra, C. Busch, T. de Freitas Pereira *et al.*, "Cross-eyed 2017: Cross-spectral

iris/periocular recognition competition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 725–732.

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

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

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

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

[57] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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

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

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

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

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

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

[64] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

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

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

[67] C. N. Fondje, S. Hu, N. J. Short, and B. S. Riggan, "Cross-domain identification for thermal-to-visible face recognition," *arXiv preprint arXiv:2008.08473*, 2020.

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

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

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

[71] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[72] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1580–1589.

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

Amir Mohammadi obtained his PhD from EPFL, Switzerland in 2020 where he worked on face presentation attack detection and developed novel domain adaptation methods. He was a post-doctoral researcher at Idiap research institute where he worked on heterogeneous face recognition. Currently, he is working as senior data scientist at Eyeware working on head and gaze tracking.

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