xEdgeFace: Efficient Cross-Spectral Face Recognition for Edge Devices

Anjith George and Sébastien Marcel
Idiap Research Institute
Rue Marconi 19, CH - 1920, Martigny, Switzerland
{anjith.george, sebastien.marcel}@idiap.ch

Abstract

Heterogeneous Face Recognition (HFR) addresses the challenge of matching face images across different sensing modalities, such as thermal to visible or near-infrared to visible, expanding the applicability of face recognition systems in real-world, unconstrained environments. While recent HFR methods have shown promising results, many rely on computation-intensive architectures, limiting their practicality for deployment on resource-constrained edge devices. In this work, we present a lightweight yet effective HFR framework by adapting a hybrid CNN-Transformer architecture originally designed for face recognition. Our approach enables efficient end-to-end training with minimal paired heterogeneous data while preserving strong performance on standard RGB face recognition tasks. This makes it a compelling solution for both homogeneous and heterogeneous scenarios. Extensive experiments across multiple challenging HFR and face recognition benchmarks demonstrate that our method consistently outperforms state-of-the-art approaches while maintaining a low computational overhead.

1. Introduction

Facial recognition (FR) has become a key component in modern biometric systems, especially for access control, thanks to its efficiency and non-intrusive nature. With the rise of deep learning, particularly convolutional neural networks (CNNs), face recognition has reached near-human performance under unconstrained conditions [43]. However, most of these systems are built for homogeneous settings, where both gallery and probe images are captured using visible-spectrum cameras.

In many real-world scenarios, such as surveillance, mobile authentication, or defense applications, relying solely on visible-light imagery is limiting. Images captured beyond the visible spectrum, such as near-infrared (NIR) [30] or thermal, offer clear advantages. For instance, NIR is more robust to changes in lighting and is harder to spoof [47, 21]. Despite these benefits, training effective models on such modalities remains a challenge due to the scarcity of large-scale, annotated heterogeneous datasets. Heterogeneous Face Recognition (HFR) aims to bridge the gap between different sensing modalities, such as matching a thermal or NIR image to a visible-light reference [40, 2]. A key subtask here is Cross-spectral Face Recognition (CFR), which deals with extreme appearance differences caused by spectral shifts between domains. CFR is especially crucial in low-light or long-range environments where visible imaging is impractical.

While recent advances in deep neural networks (DNNs) have significantly improved Heterogeneous Face Recognition (HFR), the task remains challenging due to the inherent modality gap between source and target domains, which often causes RGB-trained models to perform poorly on non-RGB data [32]. Moreover, collecting large-scale, paired cross-modal datasets is both costly and logistically demanding, highlighting the need for models that generalize well under limited supervision. At the same time, the heavy architectures used in many state-of-the-art HFR methods hinder deployment on edge devices. This has led to a growing interest in lightweight models that strike a balance between efficiency and accuracy. Recent developments in Vision Transformers (ViTs), known for capturing long-range dependencies [38], present a compelling alternative to CNNs. Combining the strengths of both paradigms opens new possibilities for building compact yet robust HFR systems suitable for real-world, resource-constrained environments.

In this work, we introduce a novel HFR framework based on a hybrid CNN-Transformer architecture EdgeFace [19], originally developed for RGB-based face recognition. Our approach starts from a backbone pre-trained on large-scale RGB datasets and adapts it to cross-modal settings with minimal heterogeneous supervision. Unlike existing methods that require extensive paired data or large mod-

els, our solution enables efficient end-to-end training in a lightweight design suited for real-time applications. Our model delivers strong performance across both homogeneous and heterogeneous settings.
The main contributions of this work are:

- We propose a framework to train a lightweight hybrid CNN-Transformer architecture for HFR, enabling robust cross-modal matching with a minimal amount of paired data.
- Our model is designed for efficiency, making it practical for deployment on edge devices.
- Extensive experiments across multiple challenging HFR and RGB benchmarks show that our approach achieves competitive or superior performance compared to state-of-the-art methods while being extremely lightweight. Code available at [1].

2. Related works

Heterogeneous Face Recognition: HFR aims to match facial images captured under different sensing modalities, such as visible (VIS), near-infrared (NIR), thermal, or sketch domains. A major challenge in HFR lies in the modality gap, i.e., the large distribution shift between modalities, which significantly degrades recognition performance when conventional face recognition models trained on RGB data are applied directly to new modalities. To mitigate this, recent research has proposed a wide range of techniques broadly categorized into three main paradigms: invariant feature learning, common-space projection, and synthesis-based approaches.

Invariant feature-based methods aim to extract robust facial descriptors that remain consistent across modalities. Early works leveraged handcrafted features such as Difference of Gaussian (DoG) filters, multi-scale LBP [48], SIFT, and MLBP [39] to model local texture patterns. Later approaches incorporated deep CNNs to learn modality-invariant embeddings [31, 32], while others introduced novel handcrafted descriptors like the Local Maximum Quotient (LMQ) [64] or composite feature integration at the score level [51].

Common-space projection methods attempt to reduce domain discrepancies by mapping multi-modal features into a shared latent space. Classical methods include Canonical Correlation Analysis (CCA) [79], Partial Least Squares (PLS) [70], and coupled regression models [44]. These techniques align modalities through linear or nonlinear transformations that preserve discriminative information. More recent methods adopt deep learning-based solutions, such as domain-specific units [9], domain-invariant units [24], coupled attribute-guided loss functions [50], and semi-supervised collaborative representations [52], which improve cross-domain alignment with minimal manual supervision or paired data. Recent works [23, 22] have shown that conditional modulation of the intermediate feature maps could address the domain gap, which was later extended to be modality agnostic [25].

Synthesis-based methods take a different route by generating modality-translated images, often in the visible domain, so that standard face recognition pipelines can be directly applied. Early techniques relied on patch-level synthesis using Markov Random Fields [72] or manifold learning methods like LLE [53]. The introduction of GANs and CycleGAN [85] has significantly advanced this line of research, enabling unpaired image translation and photo-realistic face synthesis [80, 17]. Recent innovations include latent disentanglement models [49], memory-modulated transformers for unsupervised reference-based generation [55], and plug-and-play modules like Prepended Domain Transformers (PDT) [27], which align cross-domain features without explicit synthesis. However, the synthesize-based methods increase the computation required heavily as we need to use both image translation and another face recognition network for matching.

Lightweight Face Recognition: With the widespread adoption of handheld mobile devices and edge computing, the focus in face recognition (FR) research has shifted toward developing lightweight models that maintain high accuracy while operating under strict computational constraints. This has led to a surge of interest in efficient network designs tailored for FR. MobileFaceNets [8], based on the MobileNet architecture [34, 65], were among the first to demonstrate high accuracy with fewer than 1M parameters. MixFaceNets [6] adopted the MixConv concept [71] to further improve efficiency, and ShiftFaceNet [73] leveraged ShiftNet's operations to reach competitive performance with just 0.78M parameters. ShuffleFaceNet [58], inspired by ShuffleNetV2 [56], provided flexible model sizes from 0.5M to 4.5M parameters while maintaining high accuracy. Architecture search has also played a key role: PocketNet [7] was derived using DARTS on CASIA-WebFace [78], employing multi-step knowledge distillation (KD), while VarGFaceNet [77], the winner of the ICCV 2019 LFR challenge [11], used variable group convolutions to balance computational load. More recently, SynthDistill [69] showed that synthetic data [26] and online KD can effectively train TinyFaR networks [29], allowing student models to mimic powerful teacher networks. GhostFaceNets [1] exploited redundancy in convolutional operations to build compact models with as few as 61M FLOPs using depthwise convolutions. Finally, EdgeFace [20] fused convolutional and transformer modules inspired by EdgeNeXt [57], incorporating low-rank linear modules

[1] https://www.idiap.ch/paper/xedgeface

to reduce both parameter count and computational cost, achieving near state-of-the-art performance at a fraction of the complexity, ranking first among the compact models in the EFaR 2023 challenge [42].

Lightweight Heterogeneous Face Recognition: Lightweight face recognition (FR) models are well-suited for edge deployment, yet their extension to the more challenging Heterogeneous Face Recognition (HFR) task remains underexplored. Existing HFR methods often rely on computationally heavy architectures or synthesis-based pipelines that introduce prohibitive computational overhead, limiting real-world applicability on resource-constrained devices. To bridge this gap, we propose a compact and efficient HFR framework designed for edge devices, achieving strong cross-modal performance with minimal computational and data requirements.

3. Proposed Approach

Heterogeneous face recognition (HFR) presents a unique challenge due to the scarcity of paired training data. A commonly adopted strategy to mitigate this issue involves leveraging large-scale pretrained models on visible-spectrum (RGB) data and adapting them to the heterogeneous domain. However, fine-tuning such models directly on the small HFR datasets often leads to overfitting and catastrophic forgetting, where the model's original RGB face recognition capability is significantly degraded. To address this, several prior works [9, 27] have proposed architectural modifications that introduce modality-specific branches or asymmetric processing paths. Although such designs help preserve RGB performance, they come at the cost of increased model complexity and parameter redundancy, an issue particularly detrimental when targeting lightweight models. Moreover, these approaches often assume a fixed representation for the RGB modality, requiring the heterogeneous domain (e.g., NIR, sketch, or thermal) to map into the same latent space. This rigid alignment can become a performance bottleneck, especially when modality-specific variations are large and nonlinear. In this work, we aim to design a unified model capable of handling both homogeneous and heterogeneous face recognition tasks without incurring significant compute overhead or compromising performance in either domain. Specifically, we propose a lightweight yet effective adaptation strategy for pretrained face recognition models that avoids catastrophic forgetting while enabling robust cross-modal generalization. We base our approach on EdgeFace [19], a state-of-the-art lightweight architecture that combines convolutional backbones with transformer-based components. A key insight in our method is the pivotal role of Layer Normalization (LayerNorm [18]) in modality adaptation. Rather than altering the backbone or duplicating pathways, we utilize LayerNorm as a modulation point to adapt modality-specific statistics, allowing the network to learn discriminative features for both domains within a shared architecture. To achieve this, we utilize a contrastive self-distillation training framework. The training objective consists of two components: 1) Contrastive Modality Alignment, which enforces that embeddings from paired samples (e.g., RGB-NIR) are close in the feature space, promoting modality-invariant representations. 2) Self-Distillation Regularization, which maintains the original RGB recognition capabilities by distilling knowledge from the pretrained model into the adapted one, thus mitigating catastrophic forgetting.

LayerNorm Adaptation: Previous studies have demonstrated that the statistical properties of feature maps in deep neural networks (DNNs) can effectively capture the stylistic elements of images, as first illustrated by Gatys et al. [18]. Building on this understanding of internal representations, normalization techniques have become essential for stabilizing and improving the training of deep models. Layer Normalization (LayerNorm), introduced by Ba et al. [5], addresses the limitations of batch normalization by computing normalization statistics across the features of individual samples rather than across the batch dimension. This design ensures consistent behavior during both training and inference, making it especially suitable for scenarios involving variable input sizes or non-i.i.d. data. Recent work by Zhao et al. [82] has shown that fine-tuning LayerNorm parameters within each attention block can significantly enhance model efficiency while reducing computational overhead, outperforming other parameter-efficient approaches such as Low-Rank Adaptation (LoRA) [35]. LayerNorm serves as a natural modulation point, as its scale and shift parameters can reshape the entire feature distribution at each layer without altering the underlying network weights. As noted in Xu et al. [76] and [82], with the increase in network depth, the expected LayerNorm gradient converges to zero and its variance remains small, which are properties associated with better generalization. These low-magnitude, low-variance gradients allow the model to steer feature representations toward new domains by updating only the scale and shift parameters, preserving the stability of the pretrained network. This is particularly advantageous in our setting, where adapting to new modalities without significantly modifying the rest of the network is critical (in preventing catastrophic forgetting).

Problem Formulation: We begin with a pretrained face recognition network F, parameterized by weights Θ_{FR}, trained on a large-scale dataset in the visible (RGB) spectrum. Let (X_{s_i}, X_{t_i}, y_i) denote a triplet consisting of a pair of images X_{s_i} and X_{t_i} from the source (e.g., RGB) and target (e.g., NIR or thermal) modalities respectively, and a binary identity label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates that both images correspond to the same identity and $y_i = 0$ otherwise.

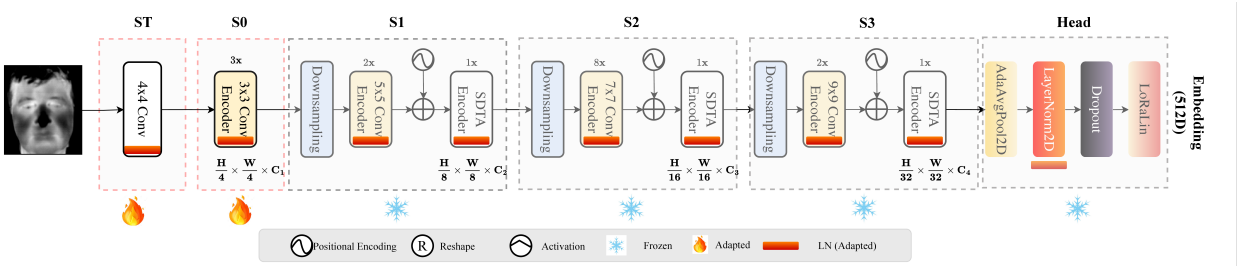

Figure 1. Model architecture of xEdgeFace models: The highlighted modules (LN-LayerNorm, ST-Conv. Stem, Stages-S0, S1, S2) are adapted while other network components remain frozen. The two loss components ensure modality alignment, keeping the source FR performance. Computational complexity remains unchanged in new models.

The goal is to adapt F into a heterogeneous face recognition network $\hat{F}$, parameterized by Θ_{HFR}, such that the resulting embeddings $e_{s_i} = \hat{F}(X_{s_i})$ and $e_{t_i} = \hat{F}(X_{t_i})$ are well-aligned in the shared embedding space if they belong to the same identity, while also preserving the discriminative ability of the original model F on the source modality.

We initialize $\hat{F}$ with the pretrained parameters Θ_{FR}, and decompose Θ_{HFR} into three disjoint subsets:

$$\Theta_{\text{HFR}} = \left\{ \Theta_{\text{LN}}^{(1:K)}, \Theta_{\text{Adapted}}, \Theta_{\text{Frozen}} \right\}, \tag{1}$$

where $\Theta_{\text{LN}}^{(1:K)}$ denotes the set of all LayerNorm parameters (from K layers), Θ_{Adapted} includes all trainable parameters except LayerNorms, and Θ_{Frozen} refers to the set of parameters that remain fixed during training.

To enforce alignment between embeddings from different modalities, we use a cosine-based contrastive loss defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{C}}(e_{s_i}, e_{t_i}, y_i) = y_i \cdot (1 - \cos(e_{s_i}, e_{t_i})) \\ + (1 - y_i) \cdot \max\left(0, \cos(e_{s_i}, e_{t_i}) - m\right),\end{aligned} \tag{2}$$

where $\cos(e_{s_i}, e_{t_i}) = \frac{e_{s_i} \cdot e_{t_i}}{\|e_{s_i}\|_2 \|e_{t_i}\|_2}$ is the cosine similarity between the embeddings and $m \in [0, 1]$ is a contrastive margin.

To prevent the model from forgetting the original face recognition capability on the source modality, we introduce a self-distillation loss that encourages the adapted model $\hat{F}$ to preserve the source modality embeddings of the pretrained model F. This is defined as:

$$\mathcal{L}_{\text{SDL}}(e_{F_{s_i}}, e_{\hat{F}_{s_i}}) = 1 - \cos(e_{F_{s_i}}, e_{\hat{F}_{s_i}}), \tag{3}$$

where $e_{F_{s_i}} = F(X_{s_i})$ is the frozen embedding from the original model and $e_{\hat{F}_{s_i}} = \hat{F}(X_{s_i})$ is the adapted embedding for the same image.

The final loss function for optimizing the adapted network $\hat{F}$ is a weighted combination of the contrastive loss for modality alignment and the self-distillation loss for performance preservation:

$$\begin{aligned}\mathcal{L}_{\text{total}} = (1 - \lambda) \cdot \mathcal{L}_{\text{C}}(e_{s_i}, e_{t_i}, y_i) \\ + \lambda \cdot \mathcal{L}_{\text{SDL}}(e_{F_{s_i}}, e_{\hat{F}_{s_i}}),\end{aligned} \tag{4}$$

where $\lambda \in [0, 1]$ is a balancing hyperparameter controlling the trade-off between cross-modal alignment and self-regularization.

In all our experiments, we set $\lambda = 0.75$ and margin $m = 0$ unless stated otherwise. This configuration empirically provided the best balance between adapting to the heterogeneous domain and retaining source modality performance.

Face Recognition Backbone We adopt the pre-trained EdgeFace [20] model as the face recognition (FR) backbone. EdgeFace is a hybrid convolutional-transformer architecture that incorporates LayerNorm modules in place of the more commonly used BatchNorm. The model is trained on the WebFace12M dataset [86], comprising over 12 million images from more than 600,000 unique identities. Input images are RGB with a spatial resolution of 112×112 pixels. Before inference, all face images are aligned using a similarity transformation to normalize eye positions to fixed coordinates. For thermal (single-channel) inputs, the image is triplicated across channels to produce a three-channel tensor suitable for processing by the RGB-trained backbone.

Implementation Details. Our HFR framework leverages a frozen copy of the pre-trained EdgeFace model as a regularizer network to guide the learning of a shallow, trainable surrogate network (Figure 1) through self-distillation. Specifically, the surrogate is initialized with the pretrained model weights and fine-tuned on the target domain. During training, only select lower-level modules, including LayerNorm layers, are unfrozen for adaptation, while the rest of the network remains fixed. To enhance cross-modal alignment, we integrate a contrastive loss applied to the feature embeddings produced by the surrogate model across both modalities (RGB and thermal).

The framework is implemented in PyTorch and builds on the Bob library [4, 3][2]. We use the Adam optimizer with a

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

learning rate of 1×10^{-4}, training for 20 epochs with a batch size of 256. The margin parameter m in the contrastive loss is set to 0, and λ is fixed at 0.75 across all experiments. Although both the pretrained and surrogate networks are used during training, only the adapted surrogate needs to be retained for inference.

4. Experiments

This section presents the results from a comprehensive set of experiments conducted using the proposed framework. We assessed the heterogeneous face recognition (HFR) performance on standard benchmarks and compared it to state-of-the-art methods in the literature. Additionally, we tested the framework on standard face recognition benchmarks to ensure that adaptation did not result in catastrophic forgetting. For all experiments, the standard cosine distance metric was employed for performance comparison.

4.1. Datasets and Protocols

For our evaluations, we utilized the following datasets:

Tufts face Dataset: The Tufts Face Database [61] comprises a diverse collection of face images across various modalities, designed for HFR tasks. In our evaluation, we specifically utilized the thermal images available in the dataset following the VIS-Thermal protocol. This dataset includes 113 identities, with a demographic breakdown of 39 males and 74 females from various regions. Each subject is represented across different modalities. Following the protocol established in [17], we randomly selected 50 identities for the training set and allocated the remaining subjects to the test set.

MCXFace Dataset: The MCXFace Dataset [27, 60] comprises images of 51 individuals captured under varying illumination conditions across three distinct sessions and utilizing multiple channels. These channels include RGB color, thermal, near-infrared (850 nm), short-wave infrared (1300 nm), depth, and depth estimated from RGB images. The dataset features five different folds, each created by randomly dividing the subjects into training and development partitions. Our evaluations focused on the challenging "VIS-Thermal" protocols of this dataset.

Polathermal Dataset: The Polathermal dataset [36], collected by the U.S. Army Research Laboratory (ARL), is a heterogeneous face recognition (HFR) dataset that includes both polarimetric LWIR imagery and color images for 60 subjects. Additionally, the dataset provides both conventional thermal images and polarimetric images for each subject. In our experiments, we utilized the conventional thermal images, following the five-fold partitioning method introduced in [9]. We used 25 identities for the training set and the remaining 35 identities for the test set.

SCFace Dataset: The SCFace dataset [28] features high-quality enrollment images for face recognition alongside low-quality probe samples taken in diverse surveillance settings using various cameras. This dataset is organized into four protocols that vary based on the quality and distance of the probe samples: close, medium, combined, and far, with the "far" protocol posing the greatest challenge. Overall, the dataset comprises 4,160 static images from 130 subjects, captured across both visible and infrared spectra.

CUFSF Dataset: The CUHK Face Sketch FERET Database (CUFSF) [81] is composed of 1,194 face images from the FERET dataset [62], each paired with a corresponding sketch created by an artist. The artistic exaggerations in the sketches make this dataset particularly challenging for heterogeneous face recognition (HFR) tasks. Following the protocols in [15], we trained our model using 250 identities and tested it on the remaining 944 identities.

CASIA NIR-VIS 2.0 Dataset: The CASIA NIR-VIS 2.0 Face Database [46] includes images of 725 individuals captured under both visible spectrum and near-infrared (NIR) lighting conditions. Each individual in the dataset is represented by 1-22 visible spectrum photos and 5-50 NIR photos. The experimental setup employs a 10-fold cross-validation approach, with 360 identities designated for training. The gallery and probe sets for evaluation contain 358 unique individuals, ensuring that the identities in the training and testing sets are completely distinct.

Metrics: Our evaluation of the models employs a range of performance metrics that are well-established in prior literature. These include the 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%). When multiple folds are present, we report the mean and the standard deviation along the folds.

4.1.1 Ablation Studies

Given the large set of design choices and hyperparameters involved, we first conduct a comprehensive set of ablation studies to analyze the contribution of individual components to the overall performance of the model. All ablation experiments are performed on the Tufts Face Dataset using the VIS-Thermal protocol, which presents the most challenging heterogeneous face recognition (HFR) scenario due to its substantial domain gap. Adapted variants of the base models are denoted as **xEdgeFace** in the following experiments.

Model Complexity: A key objective of this work is to focus on lightweight heterogeneous face recognition (HFR) models. Consequently, it is essential to compare the computational complexity of the proposed models against those commonly employed in existing literature. We evaluate the computational efficiency of our approach by reporting two key metrics: the number of floating-point operations (GFLOPs) and the total number of parameters (mea-

sured in millions, denoted as MPARAMs). As shown in Table 1, the proposed xEdgeFace variants operate with significantly reduced computational overhead and parameter count, highlighting their suitability for deployment in resource-constrained environments.

Table 1. Comparison of computational complexity between the proposed method and state-of-the-art HFR approaches, reported in terms of floating point operations (GFLOPs) and number of parameters (MPARAMs).

	GFLOPS	MPARAMS
CAIM(1-3) [22]	26.3	65.6
DIU [24]	24.2	65.2
SSMB [25]	24.2	65.5
PDT [27]	24.2	65.2
xEdgeFace - Base	1.39	18.23
xEdgeFace - S ($\gamma = 0.5$)	0.31	3.65
xEdgeFace - XS ($\gamma = 0.6$)	0.15	1.77
xEdgeFace - XXS	0.09	1.24

Adapting Different Sets of Layers: To determine the optimal set of layers for adaptation, we conduct a series of controlled ablation experiments. Specifically, we progressively adapt the LayerNorm (LN) layers, the initial convolutional stem (ST), and successive network stages: Stage 0 (S0), Stage 1 (S1), and Stage 2 (S2). The results, summarized in Table 2, reveal that adapting only the LayerNorm layers yields a substantial performance gain over the baseline pretrained model. Further adaptation of the stem and early network stages continues to improve performance. However, the extent to which layers can be adapted is closely tied to the amount of available training data. Notably, configurations such as (LN, ST) and (LN, ST, S0) achieve a favorable trade-off, delivering strong performance while limiting the number of parameters adapted.

Table 2. Ablation study on the Tufts Face Dataset using different configurations of adapted layers.

Adapted Layers	AUC	EER	Rank-1	VR@FAR=1%
Baseline	87.44	20.37	42.73	42.86
LN	96.68	9.09	73.97	77.18
ST	91.21	16.49	51.89	52.50
LN, ST	97.52	7.98	75.76	79.59
LN, ST, S0	97.91	6.68	**82.59**	**86.83**
LN, ST, S0, S1	**98.36**	6.71	81.33	83.30
LN, ST, S0, S1, S2	98.29	**6.64**	79.53	84.23

Effect of Varying λ: The hyperparameter λ controls the trade-off between supervision from the pretrained model and modality alignment objective, both of which are critical for heterogeneous face recognition (HFR). The results of this ablation study are presented in Table 3. Setting $\lambda = 0$ emphasizes only the modality alignment objective, completely omitting supervision from the pretrained model. While this encourages domain adaptation, it leads to rapid overfitting due to the limited training data. Conversely, $\lambda = 1$ relies solely on the pretrained supervision, resulting in poor cross-modal performance due to insufficient domain alignment. Although both λ = 0.50 and λ = 0.75 perform very well, we find that $\lambda = 0.75$ offers the best balance, assigning greater importance to pretrained supervision given the relatively small size of the fine-tuning dataset. This configuration not only improves HFR performance but also mitigates catastrophic forgetting, preserving recognition performance in the original RGB domain (Table 5).

Table 3. Ablation study with varying values of hyperparameter λ.

λ	AUC	EER	Rank-1	VR@FAR=1%
0.00	94.52	12.23	52.96	46.57
0.25	97.87	7.24	79.53	81.82
0.50	**98.46**	**5.95**	**83.3**	84.79
0.75	97.91	6.68	82.59	**86.83**
1.00	87.47	20.22	42.19	42.86

Experiments with Other EdgeFace Variants: While the main experiments were conducted using the EdgeFace-Base model, we also evaluated the effectiveness of our proposed adaptation on lighter variants of the architecture. Table 4 presents a comparison between the original pretrained models and their adapted counterparts, denoted as xEdgeFace. The results indicate that the absolute heterogeneous face recognition (HFR) performance correlates with the performance of the original pretrained weights. However, our adaptation consistently yields significant improvements in all model sizes, achieving relative performance gains of 103%, 194%, 257%, and 361% from the largest to the most compact variant. These results highlight the scalability and generalizability of our method, demonstrating its ability to enhance HFR performance even in extremely lightweight architectures.

Table 4. Comparison with different variants of EdgeFace

Model	AUC	EER	Rank-1	VR@FAR=1%
EdgeFace - Base	87.44	20.37	42.73	42.86
xEdgeFace - Base	97.91	6.68	82.59	**86.83** (↑ **103**%)
EdgeFace - S ($\gamma = 0.5$)	80.93	27.30	24.78	25.05
xEdgeFace - S ($\gamma = 0.5$)	96.93	8.89	71.10	**73.65** (↑ **194**%)
EdgeFace - XS ($\gamma = 0.6$)	77.76	30.24	18.67	19.11
xEdgeFace - XS ($\gamma = 0.6$)	96.28	10.02	68.22	**68.27** (↑ **257**%)
EdgeFace - XXS	75.72	31.73	17.41	12.80
xEdgeFace - XXS	95.14	11.35	60.50	**59.00** (↑ **361**%)

Face Recognition Performance of Adapted HFR Models: To evaluate the face recognition (FR) performance of the adapted models beyond the HFR setting, we evaluate the xEdgeFace model on standard FR benchmarks. We evaluate the models on LFW [37], CA-LFW [84], CP-LFW [83], CFP-FP [67], and AgeDB-30 [59] and report the accuracies. We compare the xEdgeFace model adapted for both the VIS-NIR and VIS-Thermal HFR settings, with the latter presenting the most significant domain gap among all evaluated scenarios. As shown in Table 5, the adapted

model achieves near-parity in face recognition performance with the original EdgeFace model, even under the challenging VIS-Thermal scenario. At the same time, xEdgeFace shows substantial gains in HFR performance, clearly demonstrating its capability to perform robustly across both homogeneous and heterogeneous FR tasks. This highlights the effectiveness of the self-distillation loss, which serves as a regularization mechanism that mitigates catastrophic forgetting and preserves the model's discriminative ability in the original RGB domain. Essentially, the proposed training scheme extends the model's capabilities to heterogeneous face recognition without compromising its original performance, thereby enabling its effective use in both homogeneous and heterogeneous recognition tasks.

Table 5. Face recognition performance of the pretrained and adapted model.

Model	LFW [37]	CALFW [84]	CPLFW [83]	CFP-FP [67]	AGEDB-30 [59]
EdgeFace - Base	99.83 ± 0.24	96.07 ± 1.03	93.75 ± 1.16	97.01 ± 0.94	97.60 ± 0.70
xEdgeFace - Base (VIS-Thermal)	99.78 ± 0.27	95.85 ± 1.16	93.62 ± 1.31	96.83 ± 0.99	97.50 ± 0.88
xEdgeFace - Base (VIS-NIR)	99.82 ± 0.26	96.07 ± 0.99	93.78 ± 1.24	96.94 ± 0.97	97.28 ± 0.83

4.2. Comparison with State-of-the-art

In this section, we present a comparative evaluation of the proposed xEdgeFace model against state-of-the-art heterogeneous face recognition (HFR) methods reported in the literature. Notably, xEdgeFace is significantly more lightweight than the models it is compared against, highlighting the efficiency of our approach. For all experiments, we employ the xEdgeFace-Base variant with two adaptation configurations: (LN, ST) and (LN, ST, S0). The adaptation loss weight λ is fixed at 0.75 across all evaluations to ensure consistency. However, this value can be further fine-tuned for individual datasets based on their size and face recognition performance requirements.

Experiments with Tufts face dataset: Table 6 presents the performance of xEdgeFace and 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, particularly extreme yaw angles, which degrade the performance of both visible-spectrum and heterogeneous face recognition systems. Despite these challenges, the xEdgeFace model with the (LN, ST, S0) configuration achieves the highest verification rate (69.02%) and Rank-1 accuracy (82.59%), despite being extremely lightweight.

Experiments with MCXFace Dataset: Table 7 reports the average performance over five folds for the VIS-Thermal protocol on the MCXFace dataset. The results represent the mean and standard deviation values across all five folds. The baseline corresponds to the performance of the pretrained *IresNet100* face recognition model evaluated directly on thermal images. As shown, the proposed xEdgeFace approach outperforms all compared methods, achieving the highest average Rank-1 accuracy of 91.68%.

Table 6. Experimental results on VIS-Thermal protocol of the Tufts Face dataset.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
LightCNN [74]	29.4	23.0	5.3
DVG [16]	56.1	44.3	17.1
DVG-Face [17]	75.7	68.5	36.5
DSU-Iresnet100 [27]	49.7	49.8	28.3
PDT [27]	65.71	69.39	45.45
CAIM [22]	73.07	76.81	46.94
SSMB [25] (N=1)	75.04	78.29	53.99
SSMB [25] (N=2)	78.46	80.33	54.55
xEdgeFace-Base (LN, ST)	75.76	79.59	62.89
xEdgeFace-Base (LN, ST, S0)	**82.59**	**86.83**	**69.02**

Table 7. Performance of the proposed approach in the VIS-Thermal protocol of MCXFace dataset, the Baseline is a pretrained *Iresnet100* model.

Method	AUC	EER	Rank-1
Baseline	84.45 ± 3.70	22.07 ± 2.81	47.23 ± 3.93
DSU-Iresnet100 [27]	98.12 ± 0.75	6.58 ± 1.35	83.43 ± 5.47
PDT [27]	98.43 ± 0.78	6.52 ± 1.45	84.52 ± 5.36
CAIM [22]	98.97 ± 0.24	5.05 ± 0.91	87.24±2.75
xEdgeFace-Base (LN, ST)	99.12±0.14	4.71±0.31	89.98±2.47
xEdgeFace-Base (LN, ST, S0)	99.50±0.21	3.42±0.78	**91.68±2.67**

Experiments with Polathermal Dataset: We performed experiments on the thermal-to-visible face recognition protocols of the Polathermal dataset, and the results are summarized in Table 8. The table reports the average Rank-1 identification accuracy across the five protocols defined in [9]. Our proposed xEdgeFace approach achieves the highest performance, with an average Rank-1 accuracy of 97.31%.

Table 8. Pola Thermal - Average Rank-1 recognition rate.

Method	Mean (Std. Dev.)
DPM in [36]	75.31 % (-)
CpNN in [36]	78.72 % (-)
PLS in [36]	53.05% (-)
LBPs + DoG in [48]	36.8% (3.5)
ISV in [10]	23.5% (1.1)
GFK in [68]	34.1% (2.9)
DSU(Best Result) [9]	76.3% (2.1)
DSU-Iresnet100 [27]	88.2% (5.8)
PDT [27]	97.1% (1.3)
CAIM [22]	95.00% (1.63)
xEdgeFace-Base (LN, ST)	95.98% (1.90)
xEdgeFace-Base (LN, ST, S0)	**97.31% (1.96)**

Experiments with SCFace Dataset: The SCFace dataset presents a heterogeneity challenge due to the quality disparity between the gallery (high-resolution mugshots) and probe (low-resolution surveillance camera) images. Table 9 presents the results, focusing on the "far" protocol, which is the most challenging among the evaluation settings. It can be seen that our approach achieved the best Rank-1 performance of 96.36%. This indicates that aspects such as image quality degradation can also be addressed in our framework.

Table 9. Performance of the proposed approach in the SCFace dataset, performance reported on the *far* protocol.

Method	AUC	EER	Rank-1
DSU-Iresnet100 [27]	97.18	8.37	79.53
PDT [27]	98.31	6.98	84.19
CAIM [22]	98.81	5.09	86.05
SSMB [25] (N=1)	98.77	5.91	87.73
SSMB [25] (N=2)	98.67	6.36	86.82
xEdgeFace-Base (LN, ST)	**99.86**	**1.82**	**96.36**
xEdgeFace-Base (LN, ST, S0)	99.71	2.27	93.18

Experiments with CUFSF Dataset: In this section, we evaluate our approach on the challenging task of sketch-to-photo face recognition. Table 10 presents the Rank-1 accuracies achieved by the competing methods, following the protocols defined in [15]. Our method achieves a Rank-1 accuracy of 80.30%, ranking second only to SSMB [25]. However, overall performance in this modality remains lower compared to others, such as thermal or near-infrared, reflecting the inherent difficulty of the task. The CUFSF dataset consists of viewed hand-drawn sketches [41], which, while appearing visually similar to the original subjects to humans, often lack the discriminative features critical for face recognition models. Unlike other imaging modalities, sketches are influenced by artistic interpretation and exaggeration, further widening the domain gap. Nonetheless, our method demonstrates promising performance, highlighting its robustness even under such extreme modality shifts.

Table 10. CUFSF: Rank-1 recognition rate in sketch to photo recognition.

Method	Rank-1
IACycleGAN [15]	64.94
DSU-Iresnet100 [27]	67.06
PDT [27]	71.08
CAIM [22]	76.38
SSMB [25] (N=1)	81.14
SSMB [25] (N=2)	**81.67**
xEdgeFace-Base (LN, ST)	80.30
xEdgeFace-Base (LN, ST, S0)	78.81

Experiments with CASIA-VIS-NIR 2.0 Dataset: We evaluate the proposed method on the CASIA NIR-VIS 2.0 dataset to assess its VIS-NIR matching performance. Owing to the relatively small domain gap, VIS-pretrained models already achieve competitive performance. To enable rigorous comparison, we adopt stricter evaluation metrics: VR@FAR=0.1% and 0.01%. Following the standard protocol, we report average performance and standard deviation across 10 folds. As shown in Table 11, our method consistently outperforms existing state-of-the-art approaches, demonstrating strong generalization capabilities.

Discussions: The extensive experimental results in six different heterogeneous face recognition (HFR) benchmarks demonstrate the effectiveness, generalizability, and efficiency of the proposed xEdgeFace framework. Despite its lightweight nature, xEdgeFace consistently outperforms or closely matches state-of-the-art methods across modalities, including thermal, NIR, sketch, and low-resolution surveillance images. In particular, our approach achieves top Rank-1 accuracy in challenging datasets while maintaining minimal degradation in standard face recognition benchmarks. Ablation studies show insights into which components are most effective for adaptation and the effectiveness of self-distillation in balancing domain alignment and the retention of pretrained knowledge. The framework also scales effectively to extremely compact model variants, achieving significant relative gains, which highlights its suitability for edge deployment. Furthermore, strong performance under large domain gaps (e.g., sketch-photo) confirms the robustness of our adaptation strategy across challenging heterogeneous settings.

Table 11. Experimental results on CASIA NIR-VIS 2.0.

Method	Rank-1	VR@FAR=0.1%	VR@FAR=0.01%
IDNet [63]	87.1±0.9	74.5	-
HFR-CNN [66]	85.9±0.9	78.0	-
Hallucination [45]	89.6±0.9	-	-
TRIVET [54]	95.7±0.5	91.0±1.3	74.5±0.7
W-CNN [33]	98.7±0.3	98.4±0.4	94.3±0.4
PACH [14]	98.9±0.2	98.3±0.2	-
RCN [13]	99.3±0.2	98.7±0.2	-
MC-CNN [12]	99.4±0.1	99.3±0.1	-
DVR [75]	99.7±0.1	99.6±0.3	98.6±0.3
DVG [16]	99.8±0.1	99.8±0.1	98.8±0.2
DVG-Face [17]	99.9±0.1	99.9±0.0	99.2±0.1
PDT [27]	99.95±0.04	99.94±0.03	99.77±0.09
CAIM [22]	99.96±0.02	**99.95±0.02**	99.79±0.11
xEdgeFace-Base (LN, ST)	99.96±0.02	99.91±0.02	99.83±0.04
xEdgeFace-Base (LN, ST, S0)	**99.99±0.01**	99.93±0.02	**99.86±0.04**

5. Conclusions

In this work, we proposed xEdgeFace, an efficient framework for lightweight HFR that extends existing FR models for cross-modal scenarios. By selectively adapting early convolutional and layer normalization (LN) layers within a contrastive self-distillation framework, our method enables strong cross-modal generalization while preserving the model's original capabilities in the visible spectrum. This design ensures that the adapted model performs robustly on diverse and challenging HFR tasks such as VIS-Thermal, VIS-NIR, and sketch-photo recognition, but also maintains competitive performance on standard FR benchmarks, effectively mitigating catastrophic forgetting. Extensive experiments show that xEdgeFace consistently outperforms or matches state-of-the-art methods, even with highly compact model variants, making it well-suited for edge deployment.

6. Acknowledgements

This research was funded by the European Union project CarMen (Grant Agreement No. 101168325).

References

[1] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 2023.

[2] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva. Beyond the visible: A survey on cross-spectral face recognition. *Neurocomputing*, 611:128626, 2025.

[3] 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.

[4] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*, Oct. 2012.

[5] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[6] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper. Mixfacenets: Extremely efficient face recognition networks. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[7] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multi-step knowledge distillation. *IEEE Access*, 10:46823–46833, 2022.

[8] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer, 2018.

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

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

[11] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

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

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

[14] 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.

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

[16] 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.

[17] 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.

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

[19] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel. EdgeFace: Efficient face recognition model for edge devices. *arXiv preprint arXiv:2307.01838*, 2023.

[20] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel. Edgeface: Efficient face recognition model for edge devices. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(2):158–168, 2024.

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

[22] A. George and S. Marcel. Bridging the gap: Heterogeneous face recognition with conditional adaptive instance modulation. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[23] A. George and S. Marcel. From modalities to styles: Rethinking the domain gap in heterogeneous face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(4):475–485, 2024.

[24] A. George and S. Marcel. Heterogeneous face recognition using domain invariant units. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4780–4784. IEEE, 2024.

[25] A. George and S. Marcel. Modality agnostic heterogeneous face recognition with switch style modulators. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024.

[26] A. George and S. Marcel. Digi2real: Bridging the realism gap in synthetic data face recognition via foundation models. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1469–1478, 2025.

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

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

[29] K. Han, Y. Wang, Q. Zhang, W. Zhang, C. Xu, and T. Zhang. Model rubik's cube: Twisting resolution, depth and width for tinynets. *Advances in Neural Information Processing Systems*, 33:19353–19364, 2020.

[30] S. Happy, A. Dasgupta, A. George, and A. Routray. A video database of human faces under near infra-red illumination for

human computer interaction applications. In *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, pages 1–4. IEEE, 2012.

[31] 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.

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

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

[34] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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

[37] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[38] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

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

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

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

[42] J. N. Kolf, F. Boutros, J. Elliesen, M. Theuerkauf, N. Damer, M. Alansari, O. A. Hay, S. Alansari, S. Javed, N. Werghi, et al. Efar 2023: Efficient face recognition competition. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–12. IEEE, 2023.

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

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

[45] 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.

[46] S. Li, D. Yi, Z. Lei, and S. Liao. The CASIA Nir-Vis 2.0 face database. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 348–353, 2013.

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

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

[49] D. Liu, X. Gao, C. Peng, N. Wang, and J. Li. Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis. *IEEE transactions on neural networks and learning systems*, 33(10):5611–5625, 2021.

[50] D. Liu, X. Gao, N. Wang, J. Li, and C. Peng. Coupled attribute learning for heterogeneous face recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4699–4712, 2020.

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

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

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

[54] X. Liu, L. Song, X. Wu, and T. Tan. Transferring deep representation for Nir-Vis heterogeneous face recognition. In *International Conference on Biometrics*, 2016.

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

[56] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[57] M. Maaz, A. Shaker, H. Cholakkal, S. Khan, S. W. Zamir, R. M. Anwer, and F. Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 3–20. Springer, 2023.

[58] Y. Martindez-Diaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza. Shufflefacenet: A lightweight face architecture for efficient

and highly-accurate face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[59] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

[60] Z. Mostaani, A. George, G. Heusch, D. Geissbuhler, and S. Marcel. The high-quality wide multi-channel attack (hq-wmca) database. *arXiv preprint arXiv:2009.09703*, 2020.

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

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

[63] 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.

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

[65] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[66] S. Saxena and J. Verbeek. Heterogeneous face recognition with CNNs. In *European Conference on Computer Vision*, 2016.

[67] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

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

[69] H. O. Shahreza, A. George, and S. Marcel. Knowledge distillation for face recognition using synthetic data with dynamic latent sampling. *IEEE Access*, 2024.

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

[71] M. Tan and Q. V. Le. Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, 2019.

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

[73] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer. Shift: A zero flop, zero parameter alternative to spatial convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9127–9135, 2018.

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

[75] 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.

[76] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

[77] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[78] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

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

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

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

[82] B. Zhao, H. Tu, C. Wei, J. Mei, and C. Xie. Tuning layernorm in attention: Towards efficient multi-modal LLM finetuning. In *The Twelfth International Conference on Learning Representations*, 2024.

[83] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.

[84] T. Zheng, W. Deng, and J. Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[85] 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.

[86] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.