Rethinking the Domain Gap in Near-infrared Face Recognition

Michail Tarasiou Jiankang Deng Stefanos Zafeiriou
Imperial College London
{michail.tarasiou10,j.deng16,s.zafeiriou}@imperial.ac.uk

Abstract

Heterogeneous face recognition (HFR) involves the intricate task of matching face images across the visual domains of visible (VIS) and near-infrared (NIR). While much of the existing literature on HFR identifies the domain gap as a primary challenge and directs efforts towards bridging it at either the input or feature level, our work deviates from this trend. We observe that large neural networks, unlike their smaller counterparts, when pre-trained on large scale homogeneous VIS data, demonstrate exceptional zero-shot performance in HFR, suggesting that the domain gap might be less pronounced than previously believed. By approaching the HFR problem as one of low-data fine-tuning, we introduce a straightforward framework: comprehensive pre-training, succeeded by a regularized fine-tuning strategy, that matches or surpasses the current state-of-the-art on four publicly available benchmarks. Given its simplicity and demonstrably strong performance, our method could be used as a practical solution for adjusting face recognition models to HFR as well as a new baseline for future HFR research. Corresponding training and evaluation codes can be found at `https://github.com/michaeltrs/RethinkNIRVIS`.

1. Introduction

Face recognition (FR) is one of the most important and well-studied fields in computer vision [1, 62]. It was for many years one of the main driving forces for the development of new lines of research in machine learning and was one of the first wins of Deep Neural Networks (DNNs) versus human perception [41]. Nowadays, FR technologies are widely adopted from cell-phones and laptops Face ID sensors to border control and immigration to name just a few. The most adopted and used systems currently operate with Near-Infrared (NIR) images due to their high robustness to illumination changes, as well as because it can be easily combined with liveness detection systems [5, 53]. This is in contrast with most research that is conducted and published in academia which uses images captured by conventional Visible spectrum (VIS) camera, mainly because the publicly available databases are a product of web harvesting [64]. Thus, we argue that research on face recognition using non-visible light sources, and in particular NIR light, is of significant importance for building enhanced liveness detection systes and robust anti-spoofing frameworks.

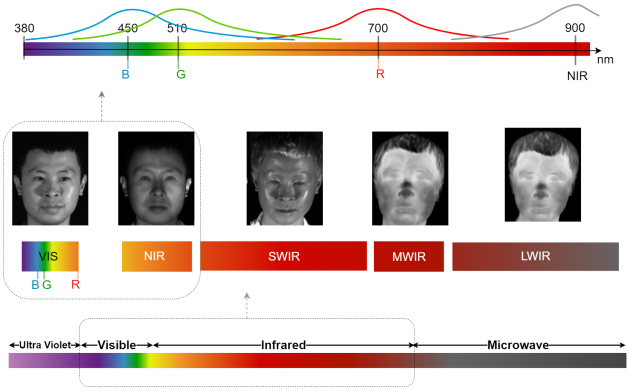

Figure 1. Face photo captured under visible and infrared light [22]. The infrared spectrum can be divided into four sub-bands: NIR (0.75–1.4μm), SWIR (1.4–3μm), MWIR (3–8 μm), and LWIR (8–15μm) [42]. The spectral sensitivity of NIR imagery is much closer to that of the VIS spectrum opposed to images captured at the far end of the IR spectrum.

Moreover, while NIR sensors are increasingly used for capturing face images during deployment (probes), these probes will need to be compared to images collected in a face database (gallery). In contrast to probes, face enrollment to gallery typically takes place in controlled environments which reduces the need for robustness to illumination conditions offered by NIR sensors. Additionally, because gallery images typically include images initially captured for use in official documents, e.g. passports, most galleries contain images captured in the VIS domain. As a result, FR systems will need to adress the problem of adequately matching faces between the two modalities. Heterogeneous NIR-VIS face recognition (HFR) [15, 16, 21, 44, 56] deals with the problem of face matching between the NIR and

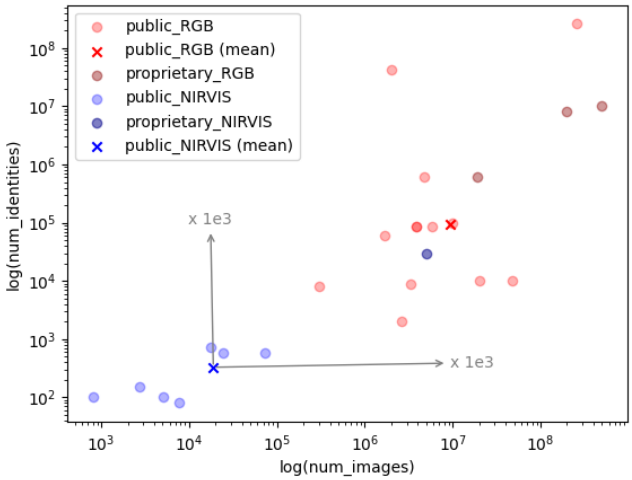

Figure 2. Size of FR datasets (#images, #identities). The average size of NIR-VIS datasets is three orders of magnitude smaller than RGB datasets.

VIS modalities and is becoming essential in modern FR systems. Most published HFR works suggest the presence of a domain gap as one of the main challenges in HFR [15, 16] and propose techniques to bridge that gap.

We follow a fundamentally different approach. Motivated by the perceptual similarities between VIS and NIR imagery (Fig.1) and the richness of VIS FR datasets (Fig.2) we employ transfer learning for solving the HFR problem. Our main observations and contributions are the following:

1. **Domain gap**: we have determined that large CNNs, when pre-trained on extensive VIS data, show remarkable zero-shot performance in NIR-VIS HFR, even outperforming current benchmarks. This observation contrasts the prevailing HFR narrative of a large domain gap and has been missed by the HFR literature which has focused exclusively on training smaller models that do not exhibit this behaviour. Instead, it indicates that large scale VIS data contain enough information to address the HFR problem, however, typically employed architectures fail to learn these cross modality features.
2. **VIS pre-training**: based on the above finding, we shift our focus towards harnessing large-scale VIS data for HFR and introduce pre-training strategies which lead to demonstrably improved zero-shot performance.
3. **NIR-VIS fine-tuning**: standard fine-tuning is found to disrupt the embedding space developed during pre-training. Two simple methods are presented that do not only rectify previous issues but also set new performance benchmarks on four public NIR-VIS HFR datasets. Furthermore, we show that further harnessing large-scale VIS data during the fine-tuning stage leads to further improvements in sensor generalization performance, making HFR systems generalize to imagery captured using novel NIR sensors.

2. Related Work

2.1. Primer on face recognition.

Recent years have witnessed a number of successful deep face recognition techniques, such as DeepFace [48], DeepID [45–47], FaceNet [40], SphereFace [32], CosFace [51] and ArcFace [9]. The majority of the advancements are based on the evolution of training loss functions. Most of the early works rely on metric-learning based loss, including contrastive loss [6] and triplet loss [40]. However, metric-learning based methods are usually inefficient on large-scale training datasets, suffering from the combinatorial explosion in the number of face pairs or triplets. Therefore, the research community has moved attention to the classification-based loss function. Wen et al. [55] develops a center loss to enhance the intra-class compactness. L_2-softmax [37] and NormFace [49] apply L_2 normalization constraint on both features and weights to improve face recognition under low-quality. Since then, several margin-based softmax losses [9, 32, 50, 51] progressively improve the performance on various celebrity benchmarks. Based on margin-based softmax loss, recent works further improve the performance by exploring adaptive parameters [30, 31, 59, 60], inter-class regularization [13, 17, 61], and sample mining [25, 52].

2.2. Heterogeneous face recognition.

Thanks to their insensitivity to the visual spectral range, images captured at the NIR spectrum are naturally robust to variations in ambient lighting conditions and, aided by IR illumination, allow for detail capturing in low light conditions at a close range. This property of NIR images suggests that the performance of trained face recognition systems will be robust on unconstrained illumination variations and in low light environments. As a result, such devices are preferred and extensively used in security and monitoring systems for which performance on low light conditions is highly desirable. Early studies on the use of NIR images for face recognition have verified the advantages of using NIR [26, 28], *Thermal infrared* (TIR) [43] and *Short wave infrared* (SWIR) [34] images compared to same dataset size of VIS images. However, VIS face recognition datasets include significant ambient light variations, thus, facilitating the development of models that are invariant to all but extreme illumination conditions, outperforming their NIR counterparts by a large margin. In this work we show that under appropriate training large scale VIS datasets can lead to strong perfromance in the heterogeneous case.

However, homogeneous face recognition in the NIR domain can be problematic for some applications such as security systems which requires face matching across NIR

and VIS modalities leading to the development of HFR. HFR methods have identified a domain gap between the two modalities and focus their attention on bridging that gap in order to transform the HFR problem into a homogeneous problem. Currently, there are two dominant approaches in the deep HFR literature:

1. **Image synthesis** methods propose to solve the HFR problem by bridging the domain gap at the level of model inputs, by learning to translate faces across domains [21, 38, 58]. A powerful VIS face recognition model is subsequently used for the face matching task.
2. **Domain-invariant feature learning** methods [18, 36, 39] aim at extracting facial identity features which are invariant to the source image domain, thus, bridging the domain gap at the level of extracted features.

Several studies use a combination of the above paradigms. Among these, [15, 16] rather than treating the image generation problem as conditional VIS generation from a NIR input, they choose an unconditional generative model trained to generate paired NIR-VIS images from random noise and generate a large amount of training samples which are used to train a network to learn a domain invariant feature space. To the best of our knowledge, the current state-of-the-art in HFR is achieved by [33], who reconstruct 3D face shape and reflectance from a large 2D facial dataset and transform the VIS reflectance to NIR reflectance in order to generate large-scale photorealistic data in the NIR and VIS spectra for further fine-tuning.

Our method offers a simplified solution to the HFR problem. We rely heavily on good initialization of model parameters from pre-trained checkpoints and proposed regularized finetuning techniques for extending the pre-trained foundational face recognition models in the new modality while retaining their performance in the original domain.

2.3. Transfer learning

Transfer learning aims at improving a learner's performance on a target task and data domain pair by "transferring" the knowledge already learned through training in different but somehow related source task and domain pair [35].

Depending on the discrepancy between the source and target domains, transfer learning can be categorized into two distinct groups [65]. In homogeneous transfer learning [54], inputs and ground truths, in both the source and target domains share the same feature space. Otherwise, the approach is referred to as heterogeneous transfer learning [8]. In the case where tasks are to remain the same and only the domains were to change, the problem is referred to as domain adaptation [14]. Furthermore, depending on how knowledge is transferred, methods can be split into two major categories [35]. Parameter-based techniques transfer knowledge in the form of inheriting parameters of a model pretrained on the source task. Similarly, instance-based methods aim at utilising a subset of the source data together with the target data. Our method is a heterogeneous method and utilises both types of knowledge transfer. We initialise model weights from models pretrained on large scale VIS data and further utilise the pretraining set during finetuning as a form of regularization.

Closely related to our work, transfer learning through reusing classifier weights has been extensively used as a means for knowledge distillation [2] including works on FR [10]. However, transfer learning for FR typically involves transferring to a different set of identities which discards the possibility of reusing classifier weights. To avoid this issue, [63] pre-compute the classifier as the mean per-class embedding of the pre-trained backbone and freeze these values to fine-tune the backbone for homogeneous FR. Additionally, they do not allow model parameters to deviate significantly from pre-trained values through an L2 regularization term. We find that building a classifier through averaging identity embeddings leads to very strong perfromance when combined with our finetuning pipeline but discard L2 regularization in favour of our proposed regulariziation scheme which is tailored to heterogeneous data.

3. Method

Our method is primarily motivated by the following observation: in contrast to VIS images, the use of NIR cameras is not ubiquitous, discarding the possibility of gathering large-scale NIR imagery data from the public domain. This showcases the important role of large-scale VIS data as a source of pre-training data. Furthermore, naive fine-tuning is found to reduce performance in most cases tested in section 4.3, suggesting the need for research on appropriate transfer learning techniques. More specifically, our method can be categorised as a domain invariant method, as we aim towards finding a common embedding space where face images of the same identity are mapped to points close to one another irrespective of the input domain. However, we do not design specific architectural or loss components to bridge the domain gap but find that large scale networks exhibit impressive zero-shot performance and few shot learning techniques are more effective compared to the current state-of-the-art while being significantly simpler to implement. Our proposed method could be particularly useful for building practical HFR systems and should replace naive finetuning as a baseline technique for future HFR research.

In the following a source FR dataset used for pre-training in the *VIS* domain $\{x_i^{preVIS}, y_i\}$, consists of x_i^{preVIS} face images and y_i associated identity labels. Similarly, a target HFR dataset used for *NIR-VIS* finetuning consists of face images and associated labels for two modalities $\{x_j^{VIS}, y_j\}$ and $\{x_k^{NIR}, y_k\}$. A schematic overview of the proposed framework is presented in Fig. 3.

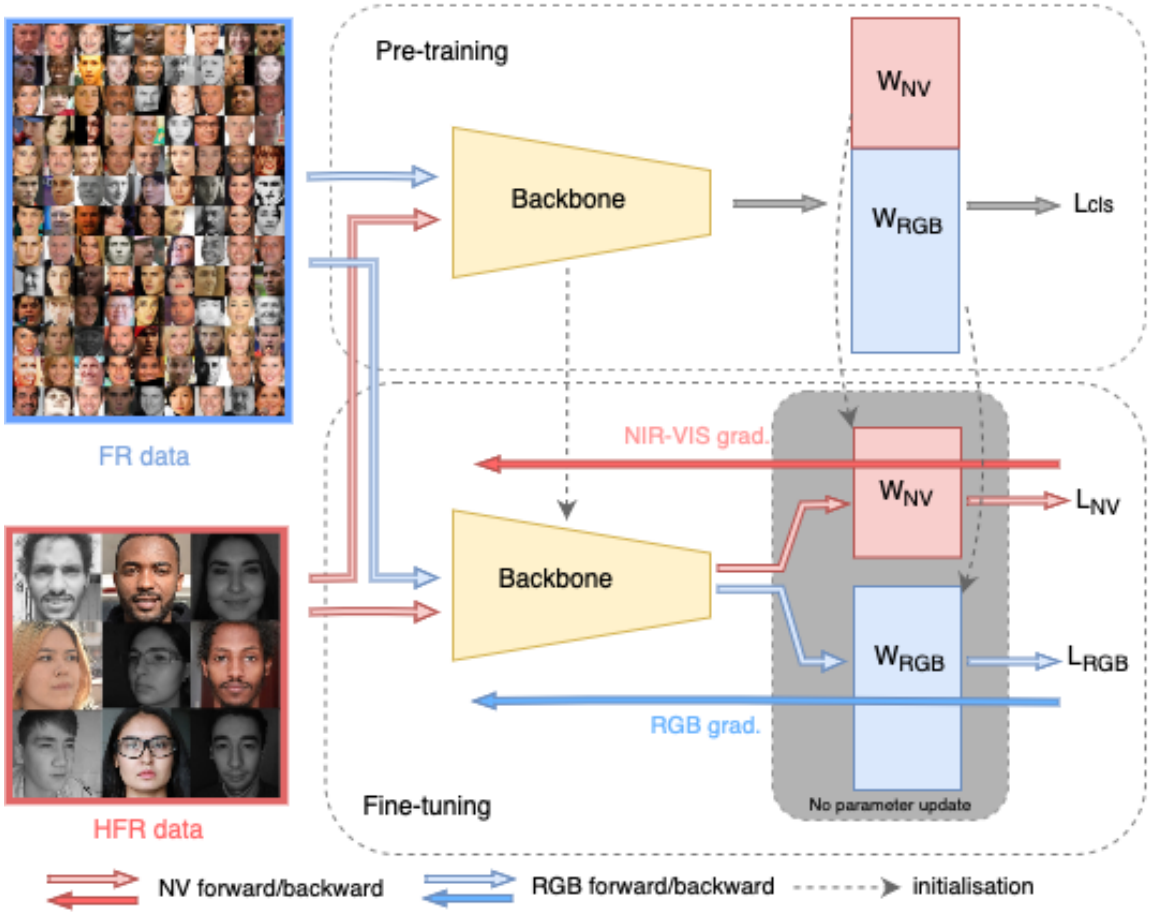

Figure 3. Proposed pre-training and fine-tuning with a subspace classifier for HFR. (top) we utilize both *source* and *target* data, use augmentation from Eq.(**??**) and train with a joint set of identities, (bottom) we initialize all modules from pre-trained counterparts, feed both *source* and *target* data to our backbone, freeze both linear classifier weights, and train with the combined loss presented in Eq.(3.2).

3.1. Pre-training with Large Scale VIS Data

To achieve strong HFR performance a model needs to be able to achieve feature invariance for both VIS and NIR modalities. Current FR models trained on large-scale VIS datasets have arguably achieved very strong performances [9, 51]. We assume that pre-training on large VIS data is enough to learn a robust embedding space for the VIS modality. Thus, we focus our attention on improving downstream transfer ability with regard to NIR images. Each face image can be decomposed into three color channels $x = \{x^R, x^G, x^B\}$ each of which is an intensity map of captured light at each respective spectral range. However, not all (R, G, B) channels share the same similarities with the NIR channel, the spectral sensitivity of the R channel has significantly higher overlap with the NIR spectral range than the B, G channels as shown in Fig.1. Motivated by this observation we are using the *red* channel as a means of shifting VIS images closer in appearance to the NIR spectrum. During pre-training we use the following data transformation with equal probability to augment our data towards the NIR spectrum:

Furthermore, assuming that we have access to some NIR FR data, we can combine the *source* (preVIS) and *target* (NIR-VIS) data for pre-training. In doing so it is possible to inject some real *target* data knowledge during pre-training that would force our model to learn some discriminative NIR features during pre-training while retaining good performance in the large scale in-the-wild VIS dataset. Additionally, we can extract the learnt classifiers for the set of target identities and transfer these directly during finetuning, which we find is very beneficial against downstream overfitting in the small scale NIR-VIS data.

3.2. Fine-Tuning on Target NIR-VIS Data

Fine-tuning DNNs directly for downstream tasks has been shown to potentially reduce performance in low data regimes [27], an observation which is also verified in sec-

tion 4.3 in this paper. While a pre-trained backbone transfers significant prior knowledge, FR classifier weights are typically initialized randomly and trained together with the backbone despite potentially having a larger capacity. This is because state-of-the-art face recognition systems are trained as image classifiers where each facial identity is treated as a single class. For example, training on MS1Mv3 [11, 19] which includes $100k$ identities with a class embedding $d = 512$ leads to classifiers with $50M$ parameters which can be more parameters than typically used deep neural networks, e.g. a ResNet50 contains around 43M parameters. For smaller and more parameter efficient backbones this discrepancy can be very significant as shown in Table 2. We propose two techniques for transferring knowledge for FR classifiers:

1. First, given the strong zero-shot performance of preVIS pre-trained models, it is reasonable to assume that the encoded representations of NIR-VIS data will also form compact clusters, the centers of which are expected to be strong identity predictors. We thus calculate classifier values for the identities found in the HFR data by enumerating all training data and taking the mean per-identity embedding as the class center. These embedding centers are then used to classify faces as one of the identites in the HFR data. We refer to this method as the *Mean Embedding Classifier* (MEC).
2. Second, assuming both *source* and *target* data are available, we combine the *source* and *target* datasets for pre-training. SUbsequently, we only keep the subspace of the learnt classifier that corresponds to *target* identities for further finetuning. In doing so our *target* class centers fit well with respective identities and by explicitly comparing them with large scale *source* centers during pre-training we end up with a more robust *target* embedding space. We refer to this method as the *Sub-Space Classifier* (SSC).

Importantly, in both cases we opt for freezing the classifier weights during finetuning and only update the parameters of the backbone model.

Finally, since there is a gap in model performance between *source* and *target* data, especially so for the smaller architectures which are of interest for deployment on mobile devices, we opt for a regularization scheme that does not penalize deviation from pre-trained parameter values as the one proposed in [63]. Instead, in the spirit of instance based transfer learning, we reuse the *source* data during fine-tuning and learn a simultaneously good solution for both HFR and homogeneous FR while placing no explicit constraint on model parameters through the following combined loss:

$$L_{finetune} = L_{cls}^{NIR-VIS} + \lambda L_{cls}^{preVIS} \quad (1)$$

where $L_{cls}^{NIR-VIS}$ and L_{cls}^{preVIS} are respectively FR classification losses for the *target* and *source* data and λ is a weight applied to the *source* data loss used for regularization purposes. We find that a particular value of λ is not critical for achieving strong HFR performance and simplify our scheme by using $\lambda = 1$ in all subsequent experiments.

Table 1. FR and HFR datasets used in experiments.

Database	Domain	N_{images}	$N_{subjects}$ *(eval)*	$Year$
Oulu-CASIA [3]	NIR-VIS	7,680	80 (40)	2009
BUAA [7]	NIR-VIS	2.7k	150 (40)	2012
CASIA 2.0 [29]	NIR-VIS	17.5k	725 (358)	2013
LAMP-HQ [57]	NIR-VIS	73.6k	573 (273)	2019
MS1Mv3 [11, 19]	VIS	5.1M	93k	2020

Table 2. Backbone Architectures used in experiments.

Model	input size	params (M)	FLOPS (G)
MFN [4]	112×112	10.48	0.23
LC29 [56]	128×128	10.48	3.70
IR18 [20]	112×112	24.03	2.62
IR50 [20]	112×112	43.59	6.32
IR100 [20]	112×112	65.15	12.12

4. Experiments

4.1. Implementation details

Datasets. We utilise the **MS1Mv3** dataset [11] for pre-training in the *VIS* domain. For finetuning, all models are initialized from MS1Mv3 pre-trained parameters. The following publicly available HFR datasets are used for finetuning our models:

- **BUAA-NIR-VIS** HFR dataset consists of NIR and VIS paired images of 150 subjects. For each individual and modality, a single image is captured, for a set of nine distinct poses and emotions, i.e., neutral-frontal, left-rotation, right-rotation, tilt-up, tilt-down, happiness, anger, sorrow, and surprise.
- **CASIA NIR-VIS 2.0** [29] contains images from 725 subjects from different age groups captured by VIS and NIR cameras. For each individual, CASIA NIR-VIS 2.0 includes 1-22 VIS and 5-50 NIR images with large pose, expression, and illumination variations, with or without accessories. We note that the different modalities of CASIA NIR-VIS 2.0 are captured independently, leading to unpaired NIR-VIS images for the same expression.
- **Oulu-CASIA-NIR-VIS** [3] contains a total of 80 individuals, each of which was photographed under three distinct illumination environments (strong, weak, and dark) at six different emotional states (anger, disgust, fear, happiness, sadness, and surprise).

Table 3. Zero-shot NIR-VIS performance after pre-training (TAR@FAR=10^{-4}). † fold-1, * with target train data.

Model	Lamp-HQ †			CASIA 2.0 †			Oulu-CASIA			BUAA		
	base	+ *red aug.*	+*target**	base	+ *red aug.*	+*target**	base	+ *red aug.*	+*target**	base	+ *red aug.*	+*target**
MFN	87.91	**88.68**	96.90	95.05	**95.75**	98.26	84.60	**88.36**	92.75	96.70	**96.73**	98.44
LC29	84.93	**86.37**	98.17	95.84	**95.97**	99.49	89.09	**89.41**	93.51	96.19	**96.24**	99.03
IR18	93.04	**93.23**	98.92	97.88	**98.76**	99.51	92.72	**94.80**	95.74	98.05	**98.45**	99.37
IR50	99.03	**99.16**	99.84	99.89	**99.90**	99.97	**99.52**	98.76	99.61	**99.84**	99.61	100.0
IR100	99.60	**99.65**	99.89	99.93	**99.97**	99.98	99.82	**99.87**	99.75	**99.92**	99.81	100.0

Table 4. Performance on NIR-VIS public datasets after naive (pre-train/fine-tune). TAR@FAR=10^{-4}.

	Model	Lamp-HQ†			CASIA 2.0†			Oulu-CASIA			BUAA		
		/✓	✓/	✓/✓	/✓	✓/	✓/✓	/✓	✓/	✓/✓	/✓	✓/	✓/✓
Naive	MFN	0.14	87.91	**92.75**	1.61	**95.05**	81.54	3.28	**84.60**	60.46	0.15	96.70	**97.58**
	IR18	3.98	93.04	**94.77**	0.73	**97.88**	86.10	8.20	**92.72**	54.94	0.19	**98.05**	93.80
	IR50	68.47	**99.03**	97.72	50.50	**99.89**	94.30	12.59	**99.52**	95.1	90.81	**99.84**	99.61
	IR100	71.52	**99.60**	96.85	52.35	**99.93**	93.86	4.08	**99.82**	92.76	89.80	**99.92**	99.77

- **LAMP-HQ** [57] contains 56,788 NIR and 16,828 high resolution VIS images of 573 people from three distinct races (Asian, Black, White) with large variations in pose ($0^o, \pm 45^o$ yaw angles, side, and bottom view), illumination (five different scenes including indoor natural light, indoor strong light, indoor dim light, outdoor natural light, and outdoor backlight) and accessories (different types of glasses, headdresses, and earings).

Information on all employed datasets is summarized in Table 1. Following common practice, we obtain normalized face crops by aligning all faces to a pre-defined template, as commonly followed in the FR literature [9, 32, 51], using five facial landmarks extracted by RetinaFace [12]. All images are subsequently scaled to (112×112) image size.

Architectures. Information for all employed architectures is summarized in Table 2. LC29 [56] has been explicitly proposed for HFR while remaining models have been proposed for general discriminative feature learning.

Training. We employ ArcFace [9] as the margin based FR loss. We pre-train for 24 epochs, batch size 512, and learning rate 0.1 which we decay by 0.1 at epochs 10, 18 and 22. We fine-tune for 20 epochs of target data, starting with learn rate 10^{-4} which we decay by 0.1 at epochs 10 and 15 and loss margin equal to 0.6. We use a batch size 64 for the finetuning data and keep the *source* data batch size to 512. All training takes place on $\times 8$ Nvidia V100 GPUs.

4.2. Zero-shot performance from VIS pre-training

We begin by assessing the zero-shot performance of pre-trained FR models in a HFR setting without further fine-tuning. Results are presented in Table 3.

First, it is observed that larger architectures (IR50, IR100) behave qualitatively differently from smaller ones, having very strong performance despite the domain shift in stark opposition to very clear performance degradation for smaller models (MFN, LC29). This finding suggests a new perspective for solving the HFR problem. It indicates that there exists adequate information in large-scale FR datasets to bridge the domain gap to NIR, however, this is not typically achieved due to the inability of smaller models to learn these features. This has critically not been observed in previous studies due to the small model capacity of employed architectures in the HFR literature.

To bridge the performance gap for smaller architectures, our proposed method for enhanced pre-training through red channel augmentation offers clear performance gains, which however are not as significant for the larger models.

Finally, we observe that including target data in the pre-train set is enough to bridge a significant portion of the performance gap between the zero-shot and fine-tuned models.

4.3. HFR Fine-Tuning Performance

In Table 4 we present experimental results on **naively fine-tuning** to target HFR data through randomly initializing classifier weights and training the model end-to-end. More specifically, we evaluate model performance with or without pre-training or fine-tuning. We observe that without pre-training all models perform substantially worse than pre-trained counterparts, in particular, the smaller architectures fail to learn any discriminative features. Thus, pre-training appears to be crucial for learning useful representations from small HFR datasets. Additionally, naive target set fine-tuning appears to destroy the embedding space learned during pre-training and lead to performance degradation. This is always the case for IR50, IR100, and almost always for IR18 and MFN.

In Table 5 we present experimental results for **regularized fine-tuning** methods. In contrast to naive finetuning (Table 4), here, we observe clear performance gains as most models reach performances close to 100% for most datasets and no performance degradation compared to no

Table 5. Performance on NIR-VIS public datasets (TAR@FAR=10^{-4}) after regularized fine-tuning with either MEC or SSC, with ($\lambda = 1$) or without ($\lambda = 0$) making use of source data. No target data have been used during pre-training.

	Model	Lamp-HQ†			CASIA 2.0†			Oulu-CASIA			BUAA		
		RCT [63]	λ=0	λ=1	RCT [63]	λ=0	λ=1	RCT [63]	λ=0	λ=1	RCT [63]	λ=0	λ=1
MEC	MFN	99.12	**99.50**	99.34	99.52	**99.61**	99.58	94.05	93.65	**96.61**	98.64	99.23	**99.52**
	IR18	99.67	**99.77**	99.70	99.83	99.89	**99.90**	96.59	95.82	**96.96**	99.61	**100.0**	99.84
	IR50	99.91	**99.93**	99.91	**99.98**	**99.98**	**99.98**	**99.88**	99.85	**99.88**	**100.0**	**100.0**	**100.0**
	IR100	**99.93**	**99.93**	**99.93**	**99.98**	**99.98**	**99.98**	**99.97**	**99.97**	**99.97**	**100.0**	**100.0**	**100.0**
SSC	MFN	99.34	**99.69**	99.54	99.60	**99.68**	**99.68**	95.12	95.73	**97.16**	**99.52**	99.34	99.41
	IR18	99.73	**99.78**	99.76	99.86	99.90	**99.92**	96.58	96.78	**99.45**	99.70	99.92	**99.95**
	IR50	99.91	**99.95**	99.93	**99.98**	**99.98**	**99.98**	99.88	99.90	**99.96**	**100.0**	**100.0**	**100.0**
	IR100	99.93	99.93	**99.94**	**99.98**	**99.98**	**99.98**	**99.97**	**99.97**	**99.97**	**100.0**	**100.0**	**100.0**

Table 6. Comparison with state-of-the-art. LC29 architecture with MEC and $\lambda = 0$. Folds 1-10.

Method	CASIA 2.0 †			Lamp-HQ †			Oulu-CASIA		BUAA	
	FAR=10^{-4}	10^{-3}	Rank-1	FAR=10^{-4}	FAR=10^{-3}	Rank-1	FAR=10^{-3}	Rank-1	FAR=10^{-3}	Rank-1
LAMP-HQ [57]	-	98.2 ± 0.2	99.2 ± 0.0	-	78.2 ± 3.0	97.3 ± 0.2	89.0	100.0	93.4	98.8
DFAL [24]	-	98.7 ± 0.2	99.1 ± 0.2	-	-	-	93.8	100.0	99.2	100.0
OMDRA [23]	-	99.4 ± 0.2	99.6 ± 0.1	-	-	-	92.2	100.0	99.7	100.0
DVG-Face [16]	99.2 ± 0.1	99.9 ± 0.0	99.9 ± 0.1	-	-	-	97.3	100.0	99.1	99.9
LC-29 [33]	**99.90 ± 0.06**	**100.0** ± 0.0	99.9 ± 0.1	98.6 ± 0.4	99.4 ± 0.3	99.1 ± 0.3	99.1	100.0	99.8	100.0
LC-29 (ours)	**99.9 ± 0.1**	99.95 ± 0.02	**100.0**	**99.35 ± 0.2**	**99.87 ± 0.05**	**100.0**	**99.62**	**100.0**	**99.90**	**100**

Table 7. Cross dataset evaluation (TAR@FAR=1^{-4}). Pre-trained MFN is fine-tuned with MEC (λ=0 /λ=1). † Fold 1.

		Evaluation			
		Lamp-HQ †	CASIA 2.0 †	Oulu-Casia	BUUA
Training	Lamp-HQ †	**99.50** / 99.34	99.17 / **99.35**	85.57 / **92.81**	92.91 / **97.66**
	CASIA 2.0 †	88.30 / **91.63**	**99.61** / 99.58	82.35 / **92.88**	91.27 / **98.28**
	Oulu-Casia	74.79 / **87.37**	84.49 / **96.88**	93.65 / **96.61**	87.79 / **96.50**
	BUUA	86.18 / **88.31**	96.86 / **98.16**	84.13 / **91.31**	99.23 / **99.52**
	no fine-tune	88.68	95.75	88.36	96.73

fine-tuning. Furthermore, we note that when using MEC and no regularization, the proposed finetuning scheme is conceptually simple, more memory efficient and faster to train than naive finetuning. We propose that this finetuning paradigm to be used as the baseline for judging the performance of future HFR research in place of naive finetuning. We additionally find that regularization w.r.t. parameter values of pre-trained network (RCT) [63] does not help and is almost always suboptimal compared to either no regularization ($\lambda = 0$) or our proposed regularization ($\lambda = 1$). This can be explained by the NIR-VIS discrepancy as RCT was proposed for homogeneous data. Compared to no regularization, our proposed regularization ($\lambda = 1$) is found to be somewhat less performant for the more diverse datasets (Lamp-HQ and CASIA) but offers important gains for the less diverse ones (Oulu-Casia and BUUA). In most cases tested our SSC outperforms MEC, albeit at the added cost of *target*-specific pre-training. As before, gains are significantly more pronounced for the smaller backbone architectures employed.

Further benefits of our regularization scheme can be observed in Table 7. There, we perform **cross-dataset evaluation** among the four HFR datasets with or without utilizing the *source* data during finetuning. Similarly, we note that apart from the more diverse datasets, Lamp-HQ and CASIA, $\lambda = 1$ outperforms $\lambda = 0$ in every case. Importantly, we observe very large performance gains in nondiagonal elements of Table 7 that have been trained and evaluated in different datasets. Thus, our proposed regularization can boost model performance in two settings. First, for small scale, low diversity datasets with limited data to finetune on our scheme reduces overfitting. Second, comparing Tables 7 and 3, we observe that when there are no *target* HFR data to finetune on, our regularization scheme can help utilize a different HFR dataset and improve performance compared to no finetuning.

Finally, in Table 6 we present a **comparison with state-of-the-art methods** for HFR. A LC29 model is pre-trained with red channel augmentation, no target data, and fine-tuned with MEC and $\lambda = 0$ for a fair comparison with literature. We observe similar performance for CASIA 2.0 and significant gains for all other datasets compared to state-of-the-art methods from the literature. Importantly, our framework is conceptually much simpler than competing methods which rely on expensive processes for generating synthetic data or employ complex architectures for domain invariant learning.

5. Conclusion

In this paper, we presented a simple method consisting of strong pre-training, followed by regularized fine-tuning, that demonstrated robust performance in HFR. Our experiments further revealed that large-scale models, in particular, showcase significant zero-shot performances compared to their smaller counterparts. This suggests that VIS data alone carry ample information to effectively address the HFR problem. While knowledge distillation (KD) might seem like a natural research avenue given these findings, our initial experiments with this technique did not yield the anticipated results, which could be attributed to various factors, including the intricacies of the HFR problem. Future work might focus on refining KD techniques applicable to HFR.

References

[1] Rama Chellappa, Charles L Wilson, and Saad Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–741, 1995. 1

[2] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11933–11942, 2022. 3

[3] Jie Chen, Dong Yi, Jimei Yang, Guoying Zhao, Stan Z. Li, and Matti Pietikainen. Learning mappings for face synthesis from near infrared to visual light images. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–163, 2009. 5

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. MobileFacenets: Efficient CNNs for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, 2018. 5

[5] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato. Non-local intrinsic decomposition with near-infrared priors. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2521–2530, 2019. 1

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 2

[7] Huang D and Wang Y Sun J. The buaa-visnir face database instructions. 2012. 5

[8] Oscar Day and Taghi M. Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 2017. 3

[9] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4, 6

[10] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 3

[11] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 5

[12] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multilevel face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[13] Yueqi Duan, Jiwen Lu, and Jie Zhou. UniformFace: Learning deep equidistributed representation for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[14] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. In *Advances in Data Science and Information Engineering*, pages 877–894, Cham, 2021. Springer International Publishing. 3

[15] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dual variational generation for low shot heterogeneous face recognition. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 1, 2, 3

[16] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dvg-face: Dual variational generation for heterogeneous face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1, 2, 3, 7

[17] Baris Gecer, Vassileios Balntas, and Tae-Kyun Kim. Learning deep convolutional embeddings for face representation using joint sample-and set-based supervision. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1665–1672, 2017. 2

[18] Dihong Gong, Zhifeng Li, Weilin Huang, Xuelong Li, and Dacheng Tao. Heterogeneous face recognition: A common encoding feature discriminant approach. *IEEE Transactions on Image Processing*, 26(5):2079–2089, 2017. 3

[19] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 2016. 5

[20] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6307–6315, 2017. 5

[21] Ran He, Jie Cao, Lingxiao Song, Zhenan Sun, and Tieniu Tan. Adversarial cross-spectral face completion for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1025–1037, 2020. 1, 3

[22] Shuowen Hu, Nathaniel Short, Benjamin Riggan, Matthew Chasse, and M. Sarfraz. Heterogeneous face recognition: Recent advances in infrared-to-visible matching. pages 883–890, 2017. 1

[23] Weipeng Hu and Haifeng Hu. Orthogonal modality disentanglement and representation alignment network for nir-vis face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3630–3643, 2022. 7

[24] Weipeng Hu, Wenjun Yan, and Haifeng Hu. Dual face alignment learning network for nir-vis face recognition. *IEEE*

Transactions on Circuits and Systems for Video Technology, 32(4):2411–2424, 2022. 7

[25] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. CurricularFace: adaptive curriculum learning loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[26] Brendan Klare and Anil K. Jain. Heterogeneous face recognition: Matching nir to visible light images. In *2010 20th International Conference on Pattern Recognition*, pages 1513–1516, 2010. 2

[27] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. 4

[28] Stan Z. Li, Rufeng Chu, Shengcai Liao, and Lun Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, 2007. 2

[29] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013. 5

[30] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *IEEE International Conference on Computer Vision*, 2019. 2

[31] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[32] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 6

[33] Yunqi Miao, Alexandros Lattas, Jiankang Deng, Jungong Han, and Stefanos Zafeiriou. Physically-based face rendering for nir-vis face recognition. In *Advances in Neural Information Processing Systems*, pages 22752–22764. Curran Associates, Inc., 2022. 3, 7

[34] Francesco Nicolo and Natalia A. Schmid. Long range cross-spectral face recognition: Matching swir against visible light images. *IEEE Transactions on Information Forensics and Security*, 7(6):1717–1726, 2012. 2

[35] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 3

[36] Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. Dlface: Deep local descriptor for cross-modality face recognition. *Pattern Recognition*, 90:161–171, 2019. 3

[37] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv:1703.09507*, 2017. 2

[38] Benjamin S. Riggan, Nathanial J. Short, Shuowen Hu, and Heesung Kwon. Estimation of visible spectrum faces from polarimetric thermal faces. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2016. 3

[39] M. Saquib Sarfraz and Rainer AU Stiefelhagen. Deep perceptual mapping for cross-modal face recognition. *International Journal of Computer Vision 122, 426–438 (2017)*. 3

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[42] Reza Shoja Ghiass, Ognjen Arandjelović, Abdelhakim Bendada, and Xavier Maldague. Infrared face recognition: A comprehensive review of methodologies and databases. *Pattern Recognition*, 47(9):2807–2824, 2014. 1

[43] D.A. Socolinsky and A. Selinger. A comparative analysis of face recognition performance with visible and thermal infrared imagery. In *2002 International Conference on Pattern Recognition*, pages 217–222 vol.4, 2002. 2

[44] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Hybrid deep learning for face verification. In *IEEE International Conference on Computer Vision*, 2013. 1

[45] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Neural Information Processing Systems*, 2014. 2

[46] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[47] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. DeepID3: Face recognition with very deep neural networks. *arXiv:1502.00873*, 2015. 2

[48] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[49] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *ACM International Conference on Multimedia*, 2017. 2

[50] Feng Wang, Weiyang Liu, Haijun Liu, and Jian Cheng. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 2018. 2

[51] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4, 6

[52] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *AAAI*, 2020. 2

[53] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin'ichi Satoh. Beyond intra-modality:

A survey of heterogeneous person re-identification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2021. 1

[54] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 2016. 3

[55] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 2016. 2

[56] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11): 2884–2896, 2018. 1, 5, 6

[57] Aijing Yu, Haoxue Wu, Huaibo Huang, Zhen Lei, and Ran He. Lamp-hq: A large-scale multi-pose high-quality database and benchmark for nir-vis face recognition. *International Journal of Computer Vision*, 2021. 5, 6, 7

[58] He Zhang, Benjamin S. Riggan, Shuowen Hu, Nathaniel J. Short, and Vishal M. Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision*, 127(6-7):845–862, 2019. 3

[59] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[60] Xiao Zhang, Rui Zhao, Junjie Yan, Mengya Gao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. P2sgrad: Refined gradients for optimizing deep face models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[61] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. RegularFace: Deep face recognition via exclusive regularization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[62] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003. 1

[63] Wenbin Zhu, Chien-Yi Wang, Kuan-Lun Tseng, Shang-Hong Lai, and Baoyuan Wang. Local-adaptive face recognition via graph-based meta-clustering and regularized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20301–20310, 2022. 3, 5, 7

[64] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[65] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the Institute of Radio Engineers*, 109(1):43–76, 2021. 3