Noname manuscript No.
(will be inserted by the editor)

Benchmarking Lightweight Face Architectures on Specific Face Recognition Scenarios

Yoanna Martínez-Díaz · Miguel Nicolás-Díaz · Heydi Méndez-Vázquez · Luis S. Luevano · Leonardo Chang · Miguel Gonzalez-Mendoza · Luis Enrique Sucar

Received: date / Accepted: date

Abstract This paper studies the impact of lightweight face models on real applications. Lightweight architectures proposed for face recognition are analyzed and evaluated on different scenarios. In particular, we evaluate the performance of five recent lightweight architectures on five face recognition scenarios: image and video based face recognition, cross-factor and heterogeneous face recognition, as well as active authentication on mobile devices. In addition, we show the lacks of using common lightweight models unchanged for specific face recognition tasks, by assessing the performance of the original lightweight versions of the lightweight face models considered in our study. We also show that the inference time on different devices and the computational requirements of the lightweight architectures allows their use on real-time applications or computationally limited platforms. In summary, this paper can serve as a baseline in order to select lightweight face architectures depending on the practical application at hand. Besides, it provides some insights about the remaining challenges and possible future research topics.

Keywords lightweight architectures · face recognition · efficient face models

1 Introduction

Face recognition (FR) is an active research topic in computer vision and image understanding. It is one of the most used and extended biometric techniques, mainly due to the fact that the face is the most common characteristic used by humans for recognition [48]. Figure 1 shows the whole pipeline of an automatic face recognition system consisting on two main modules: preprocessing and recognition. In

Y. Martínez-Díaz · M. Nicolás-Díaz · H. Méndez Vázquez
Advanced Technologies Application Center (CENATAV)
7A No. 21406, Siboney, Playa, P.C. 12200, Havana, Cuba
E-mail: ymartinez@cenatav.co.cu

L.S. Luevano · L. Chang · M. Gonzalez-Mendoza
Tecnologico de Monterrey, School of Engineering and Science, México

L. Enrique Sucar
INAOE, Puebla, Mexico

the preprocessing step the face in an image is detected and then aligned. Afterwards, a feature extractor is used to obtain a face representation and the system compares the extracted features with the gallery faces to do face matching. There are two different tasks in face matching: face verification and face identification. Face verification computes one-to-one similarity between the gallery and probe face descriptors, to determine whether they belong or not to the same subject. On the other hand, face identification is a one-to-many matching to determine the specific identity of a probe face descriptor. Many methods have been proposed to address each module including face preprocessing [7,16,62,111] and recognition [2,6,27]. In the remainder of this paper we only focus on the recognition module including face verification and identification tasks.

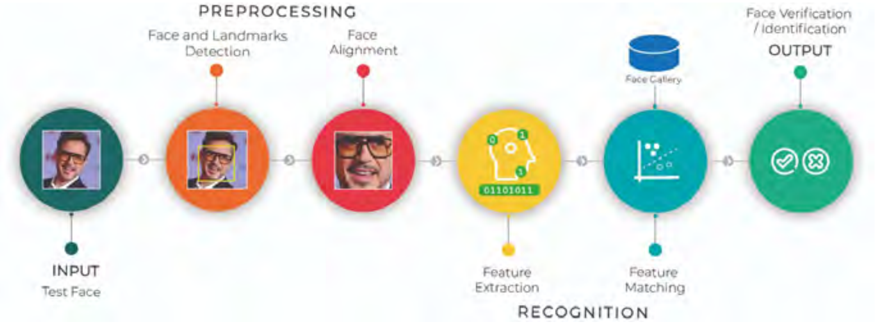


Fig. 1: Pipeline of an automatic face recognition system including preprocessing and recognition modules.

The demands of FR are also growing quickly in recent years due to its extensive applications in video surveillance, law enforcement, access control, marketing and so forth. In practice, however, unconstrained face recognition is affected by many factors such as pose variation, illumination changes, low resolution, and motion blur, resulting in low recognition accuracies. Traditional algorithms, such as the Eigenfaces [88], Fisherfaces [4], Local Binary Patterns [2], Fisher Vector based [65, 72], etc., still suffer from limitations on robustness against these complex nonlinear facial appearance variations.

In the last decade, deep neural networks through multiple layers and massive training data, have reshaped the research landscape of face recognition [27,90,92], due to their increased effectiveness, ability and generalization to learn the essential features of data by constructing powerful representations from the low-level pixels. Previous work such as DeepFace [82], FaceNet [77] and VGG-Face [73] first explored the advantages of stacking up convolutional layers in depth and width. Residual neural networks [6,32,95] became very popular due to their ability to construct extremely deep networks and to prevent data degradation by adding the original residual unit input to the computed weights in the subsequent layers of the unit, resulting in a better recognition accuracy. At the same time, the computational complexity and resource consumption (e.g. large memory and powerful GPUs) of these networks continue to increase, which make them unfeasible to deploy in real-time applications or on resource-limited devices such as mobile and embedded systems.

Therefore, designing efficient deep networks without significantly lowering the accuracy has become a hot topic within the deep learning community. To deal with this issue, many efforts aiming at mitigate the computational complexity challenge have been made for general visual recognition tasks, including the introduction of MobileNets [35,36,76], SqueezeNet [45], ShuffleNets [61,113], CondenseNet [39], EfficientNet [84] and VarGNet [112] which dramatically reduce memory requirements (layers and number of parameters) and computational complexity by factorizing standard convolutions, introducing bottleneck convolutions, and adjusting parameters such as model width, depth, and input resolution. Other approaches have tried to optimize existing networks by using binarization [12,13,74] or pruning techniques [30,33].

Recently, several attempts to build small and efficient neural networks specifically tailored for face recognition have emerged, reaching high levels of accuracy [9, 63,99,103]. Some authors have proposed to learn a compact embedding on large-scale face data while maintaining millions of parameters [99]. Others [9,63,103] have been inspired on existing lightweight mobile architectures but introducing significant modifications in order to improve their discriminative and generalization ability for face recognition. These latest face models aim at extreme model compactness and computation efficiency, overcoming the weaknesses of common mobile networks by improving their accuracy on this specific task. However, existing lightweight face networks have been trained and tested with different experimental settings and datasets, which might result in unfair comparisons, making it difficult to understand how different these models are. Moreover, very limited effort has been made towards benchmarking this kind of models on specific face recognition scenarios that are still challenging for deep learning such as large-scale image FR and video FR, and heterogeneous FR.

This paper aims at benchmarking lightweight face architectures that have been recently introduced to improve their discriminative ability in face recognition by conducting a systematic evaluation on different scenarios, which include many specific issues and challenges that must be addressed in facial recognition. Specifically, we evaluate the performance of the selected lightweight architectures on images and videos, cross-factor and heterogeneous face recognition, as well as on active authentication on mobile devices. In addition, we compare them with respect to their original lightweight models in order to analyze their breakthrough and impact for face recognition applications. On the other hand, although different lightweight face architectures have been proposed in the literature, they have not been compared each other regarding their inference time on different devices and computational requirements. We conduct here this analysis and we also compare them with some state-of-the-art face recognition models, in order to demonstrate their feasibility for real-time applications and computationally limited platforms. To the best of our knowledge, this is the first study that provides an extensive performance comparison of lightweight deep face networks on specific FR scenarios. There is a large number of papers in the literature that focus on the behavior of very deep CNN models on face recognition [3,27,62,79,92,117], where existing lightweight face architectures are not widely addressed.

In summary, in this paper we show that:

- Lightweight deep face networks can be applied to specific face recognition scenarios, that have not been fully covered in the literature, and this deserves further attention.
- It is possible to enhance the discriminative ability of existing lightweight architectures, achieving a competitive high-accuracy while preserving their small computational cost.
- Even though the generalization ability of lightweight face models on different face recognition scenarios is good, there are still some challenges such as large-scale face recognition, low-resolution and facial sketch recognition, that need to be addressed.
- The efficiency of lightweight face models makes it possible to implement face recognition on limited computational platforms such as mobile and embedded systems.

This work serves as a baseline for future research on different face recognition tasks by using lightweight deep networks. The rest of this paper is structured as follows. In Section 2 we review main efficient deep neural networks in the literature, focusing on those proposed for the specific case of face recognition. Section 3 briefly introduces the lightweight face architectures used for our study. Experimental results on different face recognition scenarios are presented in Section 4. In Section 5 we present our final remarks and discuss some remaining challenges in face recognition using lightweight deep networks. Finally, Section 6 concludes this work.

2 Efficient deep neural networks architectures

Designing deep neural network architectures for an optimal trade-off between accuracy and efficiency has been an active research area in recent years [11, 30] given the great demand for the deployment of deep CNN in embedded and mobile devices. Researchers have varied ways to speed up the network architectures which can be categorized into: 1) compressed networks, 2) efficient building block design, and 3) efficient deep networks tailored to face recognition.

2.1 Compressed networks

There are several new advances in compressing and accelerating deep CNNs in order to optimize the execution and storage for a given network [11]. For example, pruning methods [33, 52, 58] have been used to remove unimportant or unnecessary parameters in deep networks and obtain their lightweight form with comparable accuracy. In this way, the parameters require less disk storage and the computational complexity of the networks is reduced.

Low-rank decomposition has been proposed to approximate the weight matrix in neural networks with a low-rank matrix but in a more computationally efficient way. Techniques like singular value decomposition [18, 114], block-term decomposition [91] and CP-decomposition [50] have been employed to impose low-rank constraints according to in how many components the filters are decomposed. Although these methods achieve high compression levels, notable speed-ups are not obtained since they work well on fully-connected layers, and computationally-intensive operations in CNN mainly come from convolutional layers.

On the other hand, quantizing real-valued weights and activations into binary/ternary weights have been used to dramatically optimize the network's storage footprint [13, 38, 51]. Using binary quantization, the network can be compressed by a factor of about 32 compared with 32-bit floating-point networks. Besides, by binarizing both weights and activations, the network computations can be conducted using only fixed-point operations and the storage is also reduced [12, 74]. However, these low-bit representations usually come with a loss in accuracy.

Knowledge distillation technique has been adopted in many works [75, 34, 107] to compress a complex model into a simpler model that is much easier to deploy. This approach is different from the network compression or acceleration methods since it trains a student network using a teacher network, and the student network can be designed with a different network architecture. Its basic idea is to transfer the dark knowledge from a larger teacher network to a small student network by learning the class distributions provided by the teacher via softened softmax. Thus, the key component of knowledge distillation is the trade-off between speed and performance that can achieve a higher accuracy than training merely through the class labels.

2.2 Efficient building block design

Another line of inquiry to speed up a network is to design a more efficient but low-cost architecture. By integrating small modules or compact blocks we can reduce the number of weights, the memory usage and the computational cost for the inference stage. In [32, 45, 54, 81] the amount of parameters and computational cost is reduced by using 1×1 convolutions and decreasing filter sizes, which enhance the network capacity while keeping the overall computational complexity small.

Depth-wise separable convolutions were introduced in MobileNetV1 [36] as an efficient replacement for traditional convolution layers. In addition, MobileNetV2 [76] introduced inverted residuals and linear bottleneck structure in order to make even more efficient layer structures by leveraging the low-rank nature of the problem. Built upon the MobileNetV2 structure, MnasNet [83] introduced lightweight attention modules based on squeeze and excitation into the bottleneck structure. Lastly, MobileNetV3 [35] uses a combination of these layers as building blocks in order to build most effective models.

On the other hand, ShuffleNets [61, 113] utilize group convolution and channel shuffle operations to increase the information change within multiple groups. Moreover, SeesawNet [109] provides a novel search space for designing basic blocks by using uneven point-wise group convolution without any channel shuffle operation, which obtains a better trade-off between representation capability and computational cost. VarGNet [112] proposes a novel network design mechanism by fixing the number of channels in a group convolution for efficient embedded computing.

In the past few years, network architecture search (NAS) has shown itself to be a very powerful tool for discovering and optimizing network architectures. However, NAS algorithms require a lot of training time and computing capability. Other methods such as DARTS [55], SNAS [100] and ProxylessNAS [5] have also been developed to accelerate the search process by reducing search space and changing search strategies, at cost of dropping the accuracy.

2.3 Efficient deep networks tailored to face recognition

Developing very efficient and lightweight networks for the specific case of face recognition is a topic that has gained attention from the research community in recent years. FaceNet [77] is one of the first attempts to directly learn an Euclidean embedding to simplify face verification, face recognition, and face clustering tasks. The proposal uses a deep convolutional network trained to optimize the embedding itself by minimizing intra-class and maximizing inter-class squared L2 distances, rather than an intermediate bottleneck layer as in previous deep learning approaches. The benefit of the proposal is a much greater representational efficiency resulting on 128-bytes per face; however the computational complexity of this approach makes it impossible to use on low-compute devices. A Light CNN framework for face representation was proposed in [99], which introduces a Max-Feature-Map (MFM) activation function in order to reduce the number of weights and select the less noisy ones from the input. As result, the learned single model with a 256-D representation achieves state-of-the-art results on five face recognition benchmarks. The proposed Light CNN framework leads to better performance in terms of speed and storage space compared with state-of-the-art face models; however, it is not so efficient compared with mobile networks (it has 12.6 million parameters and about 3.9G FLOPs).

More recent face models have been inspired on existing lightweight mobile networks but introducing some modifications in order to improve their performance on face recognition. MobileFaceNet [9] and ShuffleFaceNet [63], which are based on MobileNetV2 [76] and ShuffleNetV2 [61], respectively, replace the Global Average Pooling layer for a Global Depth-wise Convolution layer, and use the Parametric Rectified Linear Unit (PReLU) activation function instead of the Rectified Linear Unit (ReLU) function. Similarly, MobiFace [22] network was proposed aiming at maximizing the information embedded in final feature vector while maintaining the low computational cost by adopting the Residual Bottleneck block with expansion layers as the building block. It adopts a fast downsampling strategy by quickly reducing the spatial dimensions of layers/blocks with the input size larger than 14×14, uses PReLU non-linear activation function over ReLU function and uses the Fully Connected (FC) layer in the last stage of embedding process. VarGFaceNet [103] improves the discriminative ability of VarGNet [112] by using an efficient variable group convolutional network for lightweight face recognition. Moreover, to improve the interpretation ability of this lightweight network, an equivalence of angular distillation loss is employed as objective function and a recursive knowledge distillation strategy is introduced. Inspired by different network design strategies such as using seesaw block, squeeze-and-excitation optimization and Swish non-linear activation function, SeesawFaceNets [110] achieve a competitive performance in terms of accuracy and computation cost against both lightweight and large-scale deep face recognition networks.

Note that techniques such as pruning [33,52,58], low-bit quantization [74, 38,51], and knowledge distillation are able to improve the efficiency of these lightweight networks additionally, but these are not included in the scope of this paper. Given the variety of existing lightweight face architectures, training datasets and test benchmarks including specific face recognition problems, it is quite difficult to select appropriate face models to conduct face recognition tasks. In this study, we aim at benchmarking the performance of different lightweight face architectures in specific face recognition problems, as well as assessing to their computational complexity to deploy in real-time applications or resource-limited devices.

3 Lightweight Face Architectures for Face Recognition

In this section, we describe in detail the face CNN models considered in our study. Specifically, we select three recent lightweight architectures (VarGFaceNet [103], MobileFaceNet [9] and ShuffleFaceNet [63]) tailored for face recognition which are based on common mobile networks. In addition, we modify the MobileNetV1 [36] and the ProxylessNAS [5] taking into account some of the strategies used in these lightweight face models to improve their discriminative ability on face recognition and additionally, enlarge our benchmark study.

3.1 VarGFaceNet architecture

VarGFaceNet [103] consists of an efficient variable group convolutional network based on VarGNet [112] for lightweight face recognition. Different from the blocks in VarGNet, it adds squeeze and excitation (SE) block [37] and employes PReLU activation function instead of ReLU to increase the discriminative ability of their blocks. Moreover, VarGFaceNet removes the downsample process at the start of network to preserve more information and applies variable group convolution after last convolution to shrink the feature tensor to $1 \times 1 \times 512$ before FC layer. Table 1 shows the overall architecture of the lightweight network VarGFaceNet. As we can see, 3×3 Convolution with stride 1 is used at the start of network instead of 3×3 Convolution with stride 2 as in VarGNet, which reserves the discriminative ability in lightweight networks.

Table 1: VarGFaceNet architecture.

Name	Kernel/Stride	Output Size
Image	-	112 × 112 × 3
Conv	3 × 3/1	112 × 112 × 40
Head Block	- /2	56 × 56 × 40
Stage2	- /2	28 × 28 × 80
	- /1	28 × 28 × 80
Stage3	- /2	14 × 14 × 160
	- /1	14 × 14 × 160
Stage4	- /2	7 × 7 × 320
	- /1	7 × 7 × 320
Conv5	1 × 1/1	7 × 7 × 1024
Group Conv	7 × 7/1	1 × 1 × 1024
Pointwise Conv	1 × 1/1	1 × 1 × 512
FC	-	512

3.2 MobileFaceNet architecture

MobileFaceNet [9] was specifically tailored for high-accuracy real-time face verification on mobile and embedded devices. It uses the residual bottlenecks proposed in MobileNetV2 [76] as their main building blocks. The detailed structure of MobileFaceNet architecture is shown in Table 2. Each line describes a sequence of operators, repeated n times. All layers in the same sequence have the same output channels. The first layer of each sequence has a specific stride and all the others use stride equal to 1. All spatial convolutions in the bottlenecks use 3×3 kernels and the expansion factor t is always applied to the input size. Particularly, expansion factors for bottlenecks in MobileFaceNet architecture are much smaller than those in MobileNetV2. Also, a fast downsampling strategy at the beginning of the network and an early dimensional reduction strategy at the last several convolutional layers are employed. Finally, a Global Depth-wise Convolution (GDC) layer followed by a linear 1×1 convolution layer is used instead of a Global Average Pooling layer used in MobileNetV2 as the feature output layer. Moreover, PReLU function is applied as the non-linearity instead of ReLU, which has shown to be better for face recognition.

Table 2: MobileFaceNet architecture.

Name	Kernel/Stride	t	n	Output Size
Image	-	-	-	112 × 112 × 3
Conv	3 × 3/2	-	1	56 × 56 × 64
DWConv	3 × 3/1	-	1	56 × 56 × 64
Bottleneck	-	2	5	28 × 28 × 64
Bottleneck	-	4	1	14 × 14 × 128
Bottleneck	-	2	6	14 × 14 × 128
Bottleneck	-	4	1	7 × 7 × 128
Bottleneck	-	2	2	7 × 7 × 128
Conv	1 × 1/1	-	1	7 × 7 × 512
GDConv	7 × 7/1	-	1	1 × 1 × 512
LinearConv	1 × 1/1	-	1	1 × 1 × 128

3.3 ShuffleFaceNet architecture

ShuffleFaceNet [63] extends the extremely efficient network ShuffleNetV2 [61] to the domain of face recognition. In Table 3 is presented the detailed structure of ShuffleFaceNet architecture, where the building blocks in Stages 2-4 consist of DenseNet blocks [40] and the number of channels in each block is scaled to generate four networks of different complexities, denoted as 0.5×, 1×, 1.5× and 2×. Different from ShuffleNetV2, a fast downsampling strategy is applied at the beginning of the ShuffleFaceNet architecture and a linear 1×1 convolution layer following a GDC layer is used to output the feature vector. Moreover, in order to deal with non-linearities, PReLU function is employed instead of ReLU function, which increases the discriminative ability of the network.

Table 3: ShuffleFaceNet architecture for four different levels of complexity.

Name	Kernel/Stride	Output Size	Output Channels			
			0.5×	1×	1.5×	2×
Image	-	112 × 112	3	3	3	3
Conv1	3 × 3/2	56 × 56	24	24	24	24
Stage2	-	28 × 28	48	116	176	244
Stage3	-	14 × 14	96	232	352	488
Stage4	-	7 × 7	192	464	704	976
Conv5	1 × 1/1	7 × 7	1024	1024	1024	2048
GDConv	7 × 7/1	1 × 1	1024	1024	1024	2048
LinearConv	1 × 1/1	1 × 1	128	128	128	128

Although it is known that the higher the complexity, the higher the accuracy, in [63] it was demonstrated that ShuffleFaceNet 1.5× presents the best speed-accuracy trade-off [63]. Therefore, we will use this model configuration for our study and we will refer to it as ShuffleFaceNet in the remaining of our work.

3.4 MobileFaceNetV1 architecture

Inspired on some of the strategies that have been introduced in the previous lightweight face architectures, we have decided to modify the MobileNetV1 [36] in order to enhance its discriminative ability for the specific case of face recognition and enlarge our study about this kind of networks. As result, the proposed architecture, namely MobileFaceNetV1, is built on top of the MobileNetV1 which utilizes 3 × 3 depth-wise separable convolutions to greatly reduce computational requirements and all layers are followed by a batch-norm [47]. As particularity, we adopt a fast downsampling strategy at the beginning of the network and use PReLU as non-linearity activation function instead of ReLU. Moreover, the feature output is given by a linear 1 × 1 convolution layer following a GDC layer. In Table 4 we present the overall architecture of proposed MobileFaceNetV1.

3.5 ProxylessFaceNAS architecture

We propose a modified version of the ProxylessNAS network [5]. The proposed ProxylessFaceNAS is based on the efficient CPU model found by ProxylessNAS which uses MobileNetV2 [76] as the backbone to build the architecture space. For the specific case of face recognition, we use PReLU to take place of ReLU as the activation function and replace the last global average pooling layer with a GDC layer followed by a linear 1 × 1 convolution layer. Table 5 presents the detailed architecture of ProxylesFaceNAS. MBConv3 and MBConv6 denote mobile inverted bottleneck convolution layer with an expansion ratio of 3 and 6, respectively.

Table 4: MobileFaceNetV1 architecture.

Name	Kernel/Stride	Output Size
Image	-	112 × 112 × 3
Conv	3 × 3/2	112 × 112 × 32
Conv_dw	3 × 3/1	112 × 112 × 32
Conv	1 × 1/1	112 × 112 × 64
Conv_dw	3 × 3/2	56 × 56 × 64
Conv	1 × 1/1	56 × 56 × 128
Conv_dw	3 × 3/1	56 × 56 × 128
Conv	1 × 1/1	56 × 56 × 128
Conv_dw	3 × 3/2	28 × 28 × 128
Conv	1 × 1/1	28 × 28 × 256
Conv_dw	3 × 3/1	28 × 28 × 256
Conv	1 × 1/1	28 × 28 × 256
Conv_dw	3 × 3/2	14 × 14 × 256
Conv	1 × 1/1	14 × 14 × 512
Conv_dw	3 × 3/1	14 × 14 × 512
Conv	1 × 1/1	14 × 14 × 512
Conv_dw	3 × 3/2	7 × 7 × 512
Conv	1 × 1/1	7 × 7 × 1024
Conv_dw	3 × 3/2	7 × 7 × 1024
Conv	1 × 1/1	7 × 7 × 1024
GDConv	7 × 7/1	1 × 1 × 1024
LinearConv	1 × 1/1	1 × 1 × 128

Table 5: ProxylessFaceNAS architecture.

Name	Kernel/Stride	Output Size
Image	-	112 × 112 × 3
Conv	3 × 3/1	112 × 112 × 40
MBConv1	3 × 3/1	112 × 112 × 24
MBConv6	3 × 3/1	56 × 56 × 32
MBConv3	3 × 3/1	56 × 56 × 32
MBConv3	3 × 3/1	56 × 56 × 32
MBConv3	3 × 3/1	56 × 56 × 32
MBConv6	3 × 3/1	28 × 28 × 48
MBConv3	3 × 3/1	28 × 28 × 48
MBConv3	3 × 3/1	28 × 28 × 48
MBConv3	5 × 5/1	28 × 28 × 48
MBConv6	3 × 3/1	14 × 14 × 88
MBConv3	3 × 3/1	14 × 14 × 88
MBConv6	5 × 5/2	14 × 14 × 104
MBConv3	3 × 3/1	14 × 14 × 104
MBConv3	3 × 3/1	14 × 14 × 104
MBConv3	3 × 3/1	14 × 14 × 104
MBConv6	5 × 5/1	7 × 7 × 216
MBConv3	5 × 5/1	7 × 7 × 216
MBConv3	5 × 5/1	7 × 7 × 216
MBConv3	3 × 3/1	7 × 7 × 216
MBConv6	5 × 5/1	7 × 7 × 360
GDConv	7 × 7/1	1 × 1 × 1024
LinearConv	1 × 1/1	1 × 1 × 128

4 Experiments

In this section, we evaluate the selected lightweight face architectures on several benchmarks, which cover specific scenarios of face recognition (FR) to get approximate real applications. According to their characteristics, we divide these scenarios into five categories: image FR, video FR, cross-factor FR, heterogeneous FR, and active authentication on mobile devices. First, we describe the implementation details. Then, in order to demonstrate the effectiveness of the lightweight face models, we compare them with state-of-the-art methods for the five FR scenarios. After that, we analyze the advantages of extending common lightweight architectures to the case of face recognition by comparing the lightweight face models with their original versions. Finally, we show the great computational efficiency of these lightweight models compared with state-of-the-art deep face models.

4.1 Implementation details

4.1.1 Datasets

We employ the MS1M-RetinaFace dataset [15] which is a semi-automatic refined version of the MS1M dataset [28], containing 5.1M images collected from 93K identities. All face images in this dataset are detected and aligned by using five facial landmarks predicted from RetinaFace [16] and then, resized to 112×112. The template is normalized into [1, 1] by subtracting the mean pixel value, i.e. 127.5, and then divided by 128. In addition, we use LFW [41], CFP-FP [78] and AgeDB-30 [69] as validation datasets to check the improvement from different settings.

As given in Table 6, we used different benchmarks to test the effectiveness of lightweight face models on the defined FR scenarios, that show their main caracteristics. Besides, efficient face verification image datasets (e.g. LFW [41] and CFP-FF [78]), we also report the performance of the lightweight networks on large-scale image datasets such as MegaFace [49], IJB-B [96], IJB-C [67] and Trillion-Pairs [1]. Moreover, we extensively test on video datasets (e.g. YTF [97], COX Face [43] and IQIYI-Video [15]) and cross-pose (e.g. CFP-FP [78] and CPLFW [115]) and cross-age (e.g. AgeDB-30 [69] and CALFW [116]) datasets. In addition, we explore low-resolution (e.g. SCface [26]) and photo-sketch (e.g. UoM-SGFS [24]) datasets to assess the performance on heterogeneous domains. Finally, we use UMDAA-01 dataset [23] for face-based continuous authentication.

4.1.2 Experimental settings

For training the lightweight models, we adopt Stochastic Gradient Descent (SGD) optimizer with the batch size of 128/256/512 due to limited GPU memory. Specifically, we use two Nvidia GeForce GTX 1080Ti (11GB) GPUs. The learning rate is initialized to 0.1 and decreased by a factor of 10 periodically at 100K, 140K, 160K iterations. The total iteration step is set as 200K. The momentum parameter is set to 0.9 and weight decay at 5e-4. The parameter initialization for convolution is Xavier with random sampling from a Gaussian normal distribution. All of the

Table 6: Benchmarks used for testing on different face recognition scenarios.

Database	# imgs/videos	# subjects	Task	Metrics	Key Features
Image FR					
LFW [41]	13,233/-	5,749	1:1	Acc	unconstrained images
CFP-FF [78]	5,000/-	500	1:1	Acc, EER, AUC	unconstrained images
MegaFace [49]	1M/-	690,572	1:1 1:N	VR@FAR Rank-1	large-scale, 1M of distractors
IJB-B [96]	21.8K/7,011	1,845	1:1 1:N	TAR@FAR Rank-1, Rank-5, Id@FPIR	large-scale, full pose variation
IJB-C [67]	31.3K/11,779	3,531	1:1 1:N	TAR@FAR Rank-1, Rank-5, Id@FPIR	large-scale, full pose variation
DeepGlint-Image [15]	1.8M/-	5.7K	1:1	TPR@FPR	large-scale, trillion-level pairs
Video FR					
YTF [97]	/-3,425	1,595	1:1 1:N	Acc, EER, AUC, TAR@FAR DIR(Rank-1)@FAR	video vs. video, video vs. image
COX Face [43]	1,000/3,000	1,000	1:N	Rank-1	video vs. video, video vs. image
IQIYI-Video [15]	6.3M/200K	4,934	1:1	TPR@FPR	large-scale, billion-level pairs
Cross-Factor FR					
CFP-FP [78]	2,000/-	500	1:1	Acc, EER, AUC	cross-pose
AgeDB [69]	16,488/-	568	1:1	Acc	cross-age
CPLFW [115]	11,652/-	3,968	1:1	Acc	cross-pose
CALFW [116]	12,174/-	4,025	1:1	Acc	cross-age
Heterogeneous FR					
SCface [26]	2,250/-	150	1:N	Rank-1	high vs. low resolution
UoM-SGFS [24]	1,800/-	600	1:N	Rank-N	sketch vs. photo
Active Authentication on mobile devices					
UMDAA-01 [23]	-/750	50	1:N	Rank-1	lighting variations

lightweight face models in this work are trained under this same general setting from scratch and implemented on the MxNet framework [10].

In the case of the loss function, we use the best available option for each model. In [63] the results of ShuffleFaceNet for various loss functions are reported and ArcFace exhibit the best results. Similarly, VargFaceNet and MobileFaceNet, were originally designed with ArcFace as the best option in [103] and [9] respectively. In the three cases ArcFace [14] loss function was used with an angular margin $m = 0.5$. For the proposed MobileFaceNetV1 and ProxylessFaceNAS, we explore different loss functions (SoftMax, CosFace [89] and ArcFace [14]), and evaluate their performance on the validation datasets, achieving CosFace the highest verification accuracy. Thus, both MobileFaceNetV1 and ProxylessFaceNAS are trained by CosFace loss function.

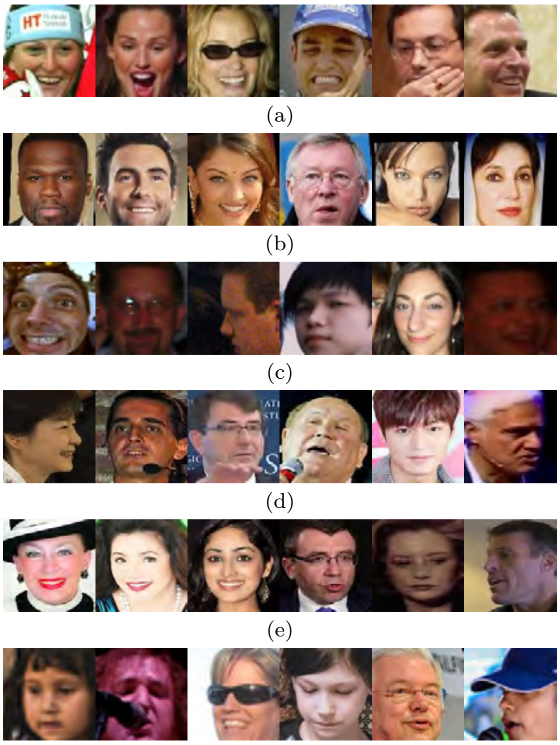
During testing, in order to generate the normalised face crops (112×112), we follow the same preprocessing adopted with the training data by using RetinaFace detector [16]. Moreover, we only keep the feature embedding of each network without the fully connected layer and extract the corresponding features for each normalised face.

4.2 Comparison with state-of-the-art methods

In order to asses the effectiveness of the lightweight face models, we conducted an extensive experimental evaluation on several face datasets representing the different FR scenarios selected for our study. It is important to note that we do not re-train or fine-tune the models for any specific problem. Thus, we directly use the feature vector given for each network and do the comparison of these features by using the cosine similarity. For each dataset, we compare the lightweight face architectures with the state-of-the-art results reported in the literature.

4.2.1 Image Face Recognition

First, we explore efficient face verification datasets such as Labeled Faces in the Wild (LFW) [41] and Celebrities in Frontal-Profile (CFP) [78]. Then, we also report the face recognition performance on recent large-scale image datasets including MegaFace [49], IJB-B [96], IJB-C [67] and Trillion-Pairs [1]. Figure 2 illustrates example images from these databases.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2: Example face images from (a) LFW, (b) CFP-FF, (c) MegaFace, (d) IJB-B, (e) IJB-C and (f) Trillion-Pairs databases.

Results on LFW. LFW contains 13,233 web-collected images from 5,749 different identities, with large variations in pose, expression and illuminations. There are 6,000 face pairs, which are divided into ten subsets, each having 300 positive pairs and 300 negative pairs. Following the standard protocol of unrestricted with labeled outside data, we give the verification accuracy on 6,000 face pairs as it is shown in Table 7. We also show in the table some details of the networks and amount of training data used. It can be seen that the five lightweight face models achieve similar results, which are very close to the best performing ones.

Table 7: Verification accuracy (%) on LFW database.

Method	#Nets	Training Data	Accuracy
DeepFace [82]	3	4M	97.4
FV-DCNN + pool5 [8]	1	0.5M	98.1
VGG-Face [73]	1	2.6M	98.9
CenterLoss [95]	1	0.7M	99.3
LightCNN-29 [99]	1	4M	99.3
Li-ArcFace [53]	1	5.1M	99.3
SphereFace [56]	1	0.5M	99.4
DeepID3 [80]	50	0.2M	99.5
Marginal Loss [17]	1	5M	99.5
FaceNet [77]	1	200M	99.6
Seesaw-shuffleFaceNet [110]	1	5.8M	99.6
CosFace [89]	1	5M	99.7
MobiFace [22]	1	3.8M	99.7
UniformFace [20]	1	3.8M	**99.8**
ResNet100-ArcFace [14]	1	5.8M	**99.8**
MobileFaceNet	1	5.1M	99.7
ShuffleFaceNet	1	5.1M	99.7
VarGFaceNet	1	5.1M	99.7
MobileFaceNetV1	1	5.1M	99.4
ProxylessFaceNAS	1	5.1M	99.2

Results on CFP-FF. CFP dataset [78] contains 10 frontal and 4 profile images of 500 subjects. Two separate experiments of Frontal-Frontal (FF) and Frontal-Profile (FP) face verification are defined, each having 10 splits with 350 same-person pairs and 350 different-person pairs. Table 8 presents the mean and standard deviation of Accuracy, Equal Error Rate (EER) and Area Under Curve (AUC) over the 10 splits for the CFP-FF experiment. As we can see, all the lightweight face models outperform the reported face recognition methods, reaching MobileFaceNet the state-of-the-art.

Results on MegaFace. The Megaface dataset [49] is one of the largest publicly available testing benchmarks for evaluating face recognition performance at the million scale of distractors. It includes a gallery set and a probe set. The gallery set consists of a subset of Flickr photos from Yahoo [85], containing more than one million images from 690K different individuals. The probe sets are two existing databases: FaceScrub and FGNet. In this work, we use FaceScrub [70] as the probe set that contains 100K photos of 530 unique individuals. In addition, we report the performance on the refined version of MegaFace with cleaned labels from [14].

Table 8: Verification results (%) on CFP-FF database.

Method	Frontal-Frontal		
	Accuracy	EER	AUC
HoG + Sub-SML [78]	88.3 ± 1.3	11.4	94.8
FV + Sub-SML [78]	91.3 ± 0.8	8.8	96.8
Deep features [78]	96.4 ± 0.7	3.5	99.4
Human [78]	96.2 ± 0.7	5.3	98.2
FV-DCNN [8]	98.4 ± 0.5	1.5	99.9
DR-GAN [86]	98.4 ± 0.7	-	-
MobileFaceNet	**99.6 ± 0.2**	**0.4**	**100**
ShuffleFaceNet	99.6 ± 0.3	0.5	100
VarGFaceNet	99.5 ± 0.2	0.5	100
MobileFaceNetV1	99.5 ± 0.2	0.6	99.9
ProxylessFaceNAS	98.8 ± 0.5	1.1	99.9

Table 9 shows the results obtained by the tested lightweight face models and state-of-the art methods reported for both the identification and verification tasks on the original and the refined MegaFace datasets. True Acceptance Rate (TAR) under False Acceptance Rate (FAR) of 10^{-6} is used to report the verification results, while the Rank-1 face accuracy is employed to the case of identification. Since the training set used has more than 0.5 million images it is regarded as large.

From Table 9 we can see that the lightweight face models achieve comparable results with respect to very deep state-of-the-art face models. Among them, MobileFaceNet obtains the best performance, surpassing strong baselines such as Marginal Loss [17] and ResNet50-ArcFace [14] and outperforming other lightweight models such as Light CNN-4, -9 and -29 [99]. On the refined MegaFace, both MobileFaceNet and ShuffleFaceNet surpass ResNet50-ArcFace [14] by a clear margin on verification and identification tasks. In the case of MobileFaceNetV1, it obtains comparable results, while ProxylessFaceNAS considerably drops its performance.

Results on IJB-B and IJB-C. The IARPA Janus Benchmark-B (IJB-B) dataset [96] consists of 1,845 subjects with 21,798 still images and 55,026 frames from 7,011 videos; whereas the IARPA Janus BenchmarkC (IJB-C) [67] is a further extension of IJB-B, by increasing dataset size and variability. In total, IJB-C contains 3,531 subjects with 31,334 still images and 117,542 frames from 11,779 videos. Both datasets include test protocols that closely model operational face recognition use cases. Specifically, we follow the evaluation protocols 1:1 verification and 1:N (mixed media) identification including both closed and open-set protocols. As performance metrics, true accepted rates (TAR) under varying false accepted rates (FAR) are reported for the 1:1 verification protocol, while Cumulative Match Characteristic (CMC) and Identification Error Trade-off (IET) curves are reported for the 1:N closed-set and 1:N open-set identification protocols, respectively.

For the verification task, IJB-B provides 12,115 templates with 10,270 genuine matches and 8M impostor matches, while IJB-C contains 23,124 templates with 19,557 genuine matches and 15,639 impostor matches. In Table 10, we compared the TAR at different FAR values (1e-5, 1e-4 and 1e-3) of considered lightweight face models with state-of-the-art models on both IJB-B and IJB-C datasets. As we can observe, the very deep ResNet100-ArcFace model [14] achieves the best verification results followed by MobileFaceNet, the second best performance reported. For the

Table 9: Face identification and verification evaluation on the original and the refined MegaFace using FaceScrub as test set.

Method	Identification (Rank-1)	Verification (TAR@FAR=1e-6)
Original version of MegaFace [49]		
Light CNN-4 [99]	60.2	62.3
Vocord-DeepVo1	75.1	67.3
Light CNN-9 [99]	67.1	77.5
Light CNN-29 [99]	73.5	84.7
FaceNet [77]	70.5	86.5
Deepsense-large	74.8	87.8
ResNet50-ArcFace [14]	77.5	92.3
UniformFace [20]	80.0	95.4
Marginal Loss [17]	80.3	92.6
AirFace [53]	80.8	96.5
CosFace [89]	**82.7**	96.7
ResNet100-ArcFace [14]	81.0	**97.0**
MobileFaceNet	79.3	95.2
VarGFaceNet	78.2	93.9
ShuffleFaceNet	77.4	93.0
MobileFaceNetV1	76.0	91.3
ProxylessFaceNAS	69.7	82.8
Refined version of MegaFace [14]		
MobiFace [22]	-	91.3
ResNet50-ArcFace [14]	91.8	93.7
AirFace [53]	98.0	97.9
ResNet100-ArcFace [14]	**98.4**	**98.5**
MobileFaceNet	95.8	96.8
VarGFaceNet	94.9	95.6
ShuffleFaceNet	94.1	94.6
MobileFaceNetV1	91.7	93.0
ProxylessFaceNAS	82.1	84.8

remaining lightweight face models, we can see that VarGFaceNet, ShuffleFaceNet and MobileFaceNetV1 are able to achieve very close results, outperforming the other state-of-the-art models. In the case of ProxylessFaceNAS, although it obtains a lower performance, it is able to improve well-established face models such as VGG-Face2.

In the case of the 1:N identification task, Table 11 and Table 12 present the Rank-1 and Rank-5 accuracy for closed-set protocol and the identification performance at FPIR=0.01 and FPIR=0.1 for open-set protocol on the IJB-B and IJB-C datasets, respectively. Note that, for both datasets, all lightweight face models, except ProxylessFaceNAS, considerably boost the IET performance of all the reported state-of-the-art methods. As in the 1:1 verification task, also in this case MobileFaceNet provides the best IET results, specially for low FPIR values like 0.01, followed by VarGFaceNet, ShuffleFaceNet and MobileFaceNetV1, while ProxylessFaceNAS is only able to improve the VGG-Face2 and MN methods. Regarding to the closed-set performance, we can observe that most of compared methods perform very similar and just slightly improvements are shown between the best performing ones. For example, in the IJB-B, DDL [42] obtains 95.4 and

Table 10: Verification TAR at different FARs on the IJB-B and IJB-C databases.

	IJB-B			IJB-C		
Method	1e-5	1e-4	1e-3	1e-5	1e-4	1e-3
VGG-Face2 [6]	67.1	80.0	88.7	74.7	84.1	90.9
MN [102]	70.8	83.1	90.9	77.1	86.2	92.7
DCN [101]	-	84.9	93.7	-	88.5	94.7
RKD [71]	78.4	89.6	94.7	85.5	92.1	96.1
SP [87]	79.4	89.8	94.9	85.9	92.3	96.2
ResNet50-ArcFace [14]	80.5	89.9	94.5	86.1	92.1	96.0
DDL [42]	83.4	90.7	95.2	88.4	93.1	96.3
ResNet100-ArcFace [14]	-	**94.2**	-	**93.2**	**95.6**	-
MobileFaceNet	**87.9**	92.8	**95.6**	92.2	94.7	**96.6**
VarGFaceNet	87.7	92.9	95.6	91.6	94.7	96.7
ShuffleFaceNet	86.5	92.3	95.2	91.3	94.3	96.3
MobileFaceNetV1	85.7	92.0	94.7	90.4	93.9	95.9
ProxylessFaceNAS	75.8	87.1	92.8	83.0	89.7	94.4

97.2 for Rank-1 and Rank-5, respectively, while MobileFaceNet achieves 95.3 and 96.9.

Table 11: 1:N (mixed media) Identification on the IJB-B database.

Method	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5
VGG-Face2 [6]	70.6	83.9	90.1	94.5
RKD [71]	70.6	87.6	93.4	96.5
SP [87]	72.4	88.0	93.8	96.6
ResNet50-ArcFace [14]	73.1	88.2	93.6	96.5
DDL [42]	76.3	89.5	93.9	**96.6**
MobileFaceNet	**81.3**	**92.2**	**94.0**	96.5
VarGFaceNet	81.0	92.1	94.0	96.5
ShuffleFaceNet	80.5	91.4	93.6	96.2
MobileFaceNetV1	78.0	90.7	93.2	95.8
ProxylessFaceNAS	67.1	84.7	90.7	94.4

Results on DeepGlint-Image The DeepGlint-Image dataset [15] is used as the large-scale image test set, which contains about 274K face images from 5.7K identities from celebrities in the LFW name list and 1.58M face images from Flickr as distractors. Different from MegaFace [49], IJB-B [96], IJB-C [67] datasets, it defines an extensive comparison metric for an unbiased evaluation of face recognition models by comparing all possible positive and negative pairs. Thus, every pair between gallery and probe set is used for evaluation (0.4 trillion pairs in total). We follow the evaluation protocol proposed in the DeepGlint-Light challenge of Lightweight Face Recognition Challenge (LFR) [15] and report the verification accuracy in terms of True Positive Rate (TPR) at different False Positive Rates (FPRs).

In Table 13 we compare the TPR corresponding to 1e-8 and 1e-9 FPR values of lightweight face models with state-of-the-art methods. We include in the compar-

Table 12: 1:N (mixed media) Identification on the IJB-C database.

Method	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5
VGG-Face2 [6]	74.6	84.2	91.2	94.9
RKD [71]	79.3	89.1	94.6	96.9
ResNet50-ArcFace [14]	79.6	89.5	94.8	96.9
SP [87]	79.9	89.5	94.7	97.0
DDL [42]	85.4	91.1	**95.4**	**97.2**
MobileFaceNet	**90.3**	**93.5**	95.3	96.9
VarGFaceNet	89.7	93.2	95.3	96.9
ShuffleFaceNet	88.9	93.0	95.0	96.7
MobileFaceNetV1	88.5	92.3	94.7	96.4
ProxylessFaceNAS	77.9	86.5	91.8	94.8

ison the top-3 ranked algorithms of the LFR2019 [46], being “YMJ”, “count” and “NothingLC”, the first, second and third ranked methods, respectively. We can observe that the top-3 solutions remarkably outperform the selected lightweight face models. However, these results are expected since some of these solutions are based on the lightweight architectures considered in our study but they explore additional techniques in order to improve their performance and enhance their interpretation ability, such as new loss designs and knowledge distillation strategy. For example, “YMJ”method is based on VarGFaceNet network but including an equivalence of angular distillation loss to guide the lightweight network and a recursive knowledge distillation to relieve the discrepancy between the teacher model and the student model. “count” improves the performance of MobileFaceNet by increasing the network depth and width and adding attention module, as well as designing a novel loss function named Li-ArcFace [53] based on ArcFace. “NothingLC” employed a teacher-student framework, where DenseNet is used as the teacher model and a modified version of the ProxylessNAS mobile network as student model. All these strategies suggest that for large-scale datasets including high-level pairs of comparisons, lightweight face architectures need additional modifications aim at increase their discriminative ability.

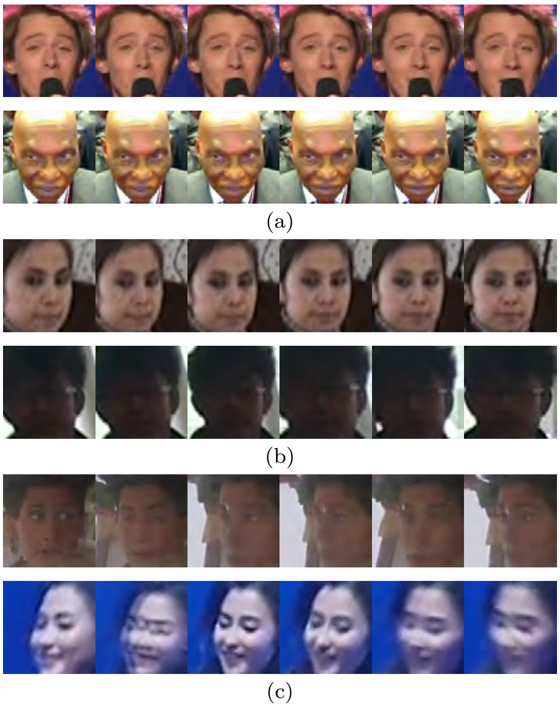
Table 13: Verification accuracy (%) on the TrillionPairs dataset.

Method	FPR=1e-9	FPR=1e-8
AM-Softmax [89]	61.61	-
SV-AM-Softmax [94]	72.71	-
ResNet100-ArcFace [14]	78.60	-
YMJ [15]	**81.89**	**88.78**
count [15]	81.20	88.42
NothingLC [15]	81.79	88.14
MobileFaceNet	70.57	80.39
VarGFaceNet	67.02	77.13
ShuffleFaceNet	64.83	75.31
MobileFaceNetV1	60.75	70.70
ProxylessFaceNAS	41.19	52.80

4.2.2 Video Face Recognition

Ideally, deep models are trained with massive images per person and are tested with one image per person, but the situation will be different in reality. Sometimes, both probe and gallery sets are represented using images or videos, but in other cases we need to match a query video against still face images or vice versa.

In order to evaluate the lightweight face networks in this kind of scenario, we select the YouTube Faces (YTF) [97], the COX Face [43] and the iQIYI-VID [57] databases. In Figure 3 we show example frames from some videos of each one of these databases. In the case of YTF and COX Face, for each video, we choose the 50 most frontal frames and a video is represented by the average of the corresponding 50 face descriptors.



(a)

(b)

(c)

Fig. 3: Examples face frames of videos from (a) YTF, (b) COX Face and (c) IQIYI-Video databases.

Results on YTF. The YTF database [97] is a large video dataset for unconstrained face recognition in videos. It contains 3,425 videos of 1,595 subjects with significant variations on expression, illumination, pose, resolution and background. The standard protocol of the YTF database provides a pair-matching benchmark corresponding to 5,000 video pairs. Specifically, these pairs are divided into ten splits, each one containing 250 positive pairs and 250 negative ones. Under this protocol, three metrics are usually considered to report the verification results: the mean accuracy, the area under curve (AUC) and the equal error rate (EER).

Recently, new relevant evaluation protocols for the YTF database (REP-YTF) were designed in order to better fit operational face recognition systems [66]. The proposal includes large-scale face verification at low FAR values and open/closed-set identification protocols for both video-to-video and video-to-image comparisons. Under these protocols, the YTF database is divided into ten random trials of training and test sets, containing on average, more than 800 face videos available for training and more than 2,500 face videos for testing for each trial. For the verification protocol, the test set of each trial is used to compute the matching scores, resulting on 2,277 genuine and 3,314,989 impostor not-duplicated comparisons on average. Thus, over 3,317,266 video pairs comparison scores are computed for each trial. Mean True Acceptance Rate (TAR) at 0.1% and 1% FAR values, and mean EER over the ten random trials are used as verification performance metrics. Both Linear Discriminant Analysis (LDA) metric learning and Cosine (Cos) similarity are used to perform the comparison.

In Tables 14 and 15 we present the verification results obtained by the lightweight face architectures and state-of-the-art methods on the YTF database, using the standard protocol [97] and the REP-YTF verification protocol [66], respectively. As we can see from Table 14, by using the standard protocol, the five lightweight face models obtain very similar verification results, which are comparable to the best state-of-the-art methods. In the case of REP-YTF protocol, it can be seen from Table 15, that these lightweight models achieve the state-of-the-art by using cosine similarity. In general, among the lightweight face architectures, MobileFaceNet and VarGFaceNet obtain the best results on both protocols, followed by ShuffleFaceNet, MobileFaceNetV1 and ProxylessNAS.

Table 14: Verification performance on YTF database using the standard protocol.

Method	Accuracy	AUC	EER
LBinVF^2 [65]	83.3	93.2	14.6
VF^2 [72]	84.7	93.0	14.9
ShiftFaceNet [98]	90.1	96.1	-
DeepFace-single [82]	91.4	96.3	8.6
CenterLoss [95]	94.9	-	-
TBE-CNN [19]	95.0	-	-
SphereFace [56]	95.0	-	-
FaceNet [77]	95.1	-	-
Light CNN-29 [99]	95.5	-	-
NAN [105]	95.7	**98.8**	-
Marginal Loss [17]	96.0	-	-
VGG-Face [73]	97.3	-	-
CosFace [89]	97.6	-	-
UniformFace [20]	97.7	-	-
ResNet100-ArcFace [14]	**98.0**	-	-
MobileFaceNet	96.2	98.4	**4.5**
VarGFaceNet	96.0	98.3	5.1
ShuffleFaceNet	95.7	98.2	5.3
MobileFaceNetV1	95.2	98.0	5.6
ProxylessFaceNAS	94.4	98.0	6.1

Table 15: Verification results (%) on YTF database using the REP-YTF protocol.

Method	TAR@FAR=0.1%	TAR@FAR=1%	EER
LBinVF^2 + LDA [66]	21.27 ± 0.5	39.59 ± 0.8	18.12 ± 0.7
VF^2 + LDA [66]	20.84 ± 0.4	40.68 ± 0.8	16.37 ± 0.5
VGG-Face + JB [66]	43.04 ± 1.9	66.91 ± 1.4	9.93 ± 0.8
Dlib + LDA [66]	50.70 ± 1.2	75.98 ± 0.9	7.59 ± 0.4
ProxylessFaceNAS + LDA	69.64 ± 1.2	85.96 ± 0.8	6.17 ± 0.4
MobileFaceNetV1 + LDA	76.72 ± 1.1	88.45 ± 0.6	6.06 ± 0.3
MobileFaceNet + LDA	80.25 ± 0.9	89.94 ± 0.5	5.95 ± 0.3
ShuffleFaceNet + LDA	81.69 ± 0.9	90.85 ± 0.5	5.56 ± 0.4
VarGFaceNet + LDA	84.30 ± 0.7	91.69 ± 0.5	5.26 ± 0.3
ProxylessFaceNAS + Cos	82.86 ± 0.6	90.96 ± 0.5	5.16 ± 0.3
MobileFaceNetV1 + Cos	86.90 ± 0.6	91.90 ± 0.6	5.02 ± 0.3
ShuffleFaceNet + Cos	89.61 ± 0.5	93.16 ± 0.5	4.52 ± 0.4
MobileFaceNet + Cos	90.16 ± 0.6	93.53 ± 0.5	4.55 ± 0.3
VarGFaceNet + Cos	**90.28 ± 0.6**	**93.56 ± 0.5**	**4.39 ± 0.4**

For both open and closed-set REP-YTF identification protocols [66], three different configurations of the test set are obtained for each trial by using the openness values: 0.2, 0.5 and 0.9, resulting on different gallery sizes. Table 16 and 17 show the mean Detection and Identification Rate (DIR) at rank-1 and False Acceptance Rate (FAR) of 1%, achieved by the lightweight face models and the best performing methods in [66], over the ten random trials defined in the open and closed-set REP-YTF identification protocols, respectively, both including video-to-video and video-to-image comparisons.

Table 16: Mean DIR (%) at rank-1 and FAR = 1% for REP-YTF open-set identification protocol in video-to-video and video-to-image settings.

Method	video-to-video			video-to-image		
	(0.2)	(0.5)	(0.9)	(0.2)	(0.5)	(0.9)
LBinVF^2 + LDA [66]	10.05	8.57	8.18	6.58	4.78	4.53
VF^2 + LDA [66]	10.67	8.47	8.84	5.95	4.92	4.82
VGG-Face + JB [66]	22.83	18.16	16.28	17.33	14.20	13.14
Dlib + LDA [66]	25.97	20.12	17.99	16.62	14.26	11.41
ProxylessFaceNAS + LDA	45.77	38.73	38.60	42.65	36.04	32.76
MobileFaceNetV1 + LDA	55.23	49.66	47.42	57.81	52.74	50.23
MobileFaceNet + LDA	59.69	53.67	51.86	60.04	56.25	53.79
ShuffleFaceNet + LDA	59.79	54.21	51.44	64.33	59.64	57.57
VarGFaceNet + LDA	65.71	60.02	57.34	66.83	61.46	59.69
ProxylessFaceNAS + Cos	68.23	63.16	60.78	60.37	55.82	53.02
MobileFaceNetV1 + Cos	76.04	73.35	71.79	69.12	64.42	61.49
ShuffleFaceNet + Cos	82.30	79.14	77.40	76.08	73.05	**71.87**
VarGFaceNet + Cos	82.44	79.74	76.65	76.96	**73.22**	70.85
MobileFaceNet + Cos	**83.48**	**79.84**	**77.91**	**78.00**	73.21	71.34

As we can see in Table 16, the open-set identification results of lightweight face architectures are significantly superior with respect to the existing methods. The best performance is obtained by MobileFaceNet using cosine similarity, which

Table 17: Mean identification rates for REP-YTF closed-set identification protocol in video-to-video and video-to-image settings.

Method	video-to-video			video-to-image		
	(0.2)	(0.5)	(0.9)	(0.2)	(0.5)	(0.9)
LBinVF^2 + LDA [66]	37.70	30.93	29.85	31.55	24.46	24.01
VF^2 + LDA [66]	38.22	32.65	30.80	35.23	28.92	27.46
VGG-Face + LDA [66]	60.60	54.59	52.33	53.87	47.49	45.02
Dlib + LDA [66]	71.91	66.53	64.18	56.78	50.90	48.33
ProxylessFaceNAS + LDA	80.07	75.62	73.64	74.16	69.17	67.14
MobileFaceNetV1 + LDA	83.94	81.98	80.84	80.40	78.71	76.70
MobileFaceNet + LDA	86.34	84.25	83.66	82.74	80.18	79.06
ShuffleFaceNet + LDA	86.83	85.52	84.61	84.40	81.89	80.36
VarGFaceNet + LDA	88.22	86.81	85.93	86.12	83.49	82.69
ProxylessFaceNAS + Cos	86.25	84.60	83.86	83.69	80.29	78.85
MobileFaceNetV1 + Cos	88.79	87.38	86.74	84.94	82.96	81.50
ShuffleFaceNet + Cos	90.71	89.70	89.24	88.45	86.54	85.61
MobileFaceNet + Cos	91.14	89.99	89.63	**88.90**	**87.10**	86.32
VarGFaceNet + Cos	**91.17**	**90.32**	**89.95**	88.49	87.06	**86.40**

outperforms the Dlib method in more than 50%. A similar behavior is presented in Table 17 for the closed-set protocol, where the recognition rates at rank-1 are at least 20% better.

Results on COX Face. The COX Face database [43] was designed to simulate video surveillance applications, including both still images and videos of 1,000 Chinese subjects. For each subject, the database contains one high-quality still image and three video clips each captured from a different camera (Cam1, Cam2 and Cam3). The videos have natural variations in pose, expression, lighting, blur and face resolution. COX Face database presents evaluation protocols for three different video-based face recognition modalities: 1) Video-to-Still (V2S); 2) Still-to-Video (S2V); and 3) Video-to-Video (V2V), respectively, taking video or still image as query or target. For all the evaluations, ten random 300/700 partitions are provided and the mean and the standard deviation of the face recognition results on the 10 runs are reported.

Table 18 and 19 present the rank-1 recognition rates of the tested methods for the V2V and V2S/S2V face recognition modalities, respectively. It can be seen that ShuffleFaceNet, VarGFaceNet and MobileFaceNet perform very similar on the three modalities of this database, slightly better than MobileFaceNetV1 and ProxylessFaceNAS. We can also observed that, in the case of V2V, the performance of lightweight models are comparable to those of state-of-the-art methods, but when we compare a video against an image, the lightweight models perform much better.

Results on iQIYI-VID. The IQIYI-Video dataset [15] is a large-scale benchmark for multi-modal person identification including face, cloth, voice, gait and subtitles, for character identification. It contains 200K videos of 10K identities from IQIYI variety shows, films and television dramas. For evaluating, we follow the Protocol-3 proposed in the IQIYI-Light challenge of LFR2019 [15] and report the verification accuracy in terms of True Positive Rate (TPR) at different False Positive Rates (FPRs). To get the embedding features for videos, the feature centre of all frames

Table 18: Recognition rates at Rank-1 for V2V face recognition on the COX Face database.

Method	V2V face recognition rate					
	V2-V1	V3-V1	V3-V2	V1-V2	V1-V3	V2-V3
PSCL-SS [43]	57.70	73.17	67.70	62.77	78.26	68.91
LERM-SS [44]	65.94	78.24	70.67	64.44	80.53	72.96
GGDA [31]	70.80	76.23	71.99	69.17	76.77	77.43
CDL [93]	78.43	85.31	79.71	75.56	85.84	81.87
BinVF^2 [64]	93.37	96.81	95.93	93.17	96.56	96.39
VGG-Face [73]	94.51	95.34	96.39	93.39	96.10	96.60
TBE-CNN [19]	98.07	98.16	97.93	97.20	99.30	**99.33**
LBinVF^2 [65]	97.83	98.96	98.63	97.99	98.93	98.87
VF^2 [72]	98.27	**99.34**	**99.06**	**98.33**	**99.47**	99.23
ShuffleFaceNet	**98.36**	98.81	98.94	97.81	98.99	99.03
VarGFaceNet	97.81	98.74	98.99	98.04	99.06	99.10
MobileFaceNet	97.13	98.57	98.64	96.79	98.84	98.70
MobileFaceNetV1	96.96	98.51	98.69	96.87	98.54	98.76
ProxylessFaceNAS	96.77	98.36	98.94	96.90	98.91	99.11

Table 19: Recognition rates at Rank-1 for V2S/S2V face recognition on the COX Face database.

Method	V2S face recognition rate			S2V face recognition rate		
	V1-S	V2-S	V3-S	S-V1	S-V2	S-V3
PSCL-ES [43]	38.60	33.20	53.26	36.39	30.87	50.96
LERM-EA [44]	45.03	42.53	59.76	43.17	41.51	60.26
LERM-ES [44]	45.71	42.80	58.37	49.07	44.16	63.83
VGG-Face [73]	88.36	80.46	90.93	69.61	68.11	76.01
TBE-CNN [19]	93.57	93.69	98.96	88.24	87.86	95.74
VarGFaceNet	**95.11**	94.97	**99.80**	**96.44**	**95.90**	99.79
MobileFaceNet	94.80	**95.33**	99.63	96.06	95.57	**99.86**
ShuffleFaceNet	94.40	95.03	99.69	95.74	95.36	99.56
ProxylessFaceNAS	92.81	95.13	99.61	93.07	94.86	99.20
MobileFaceNetV1	92.63	93.44	99.50	93.47	94.43	99.49

from the video is computed and the extracted features are compared using Cosine distance.

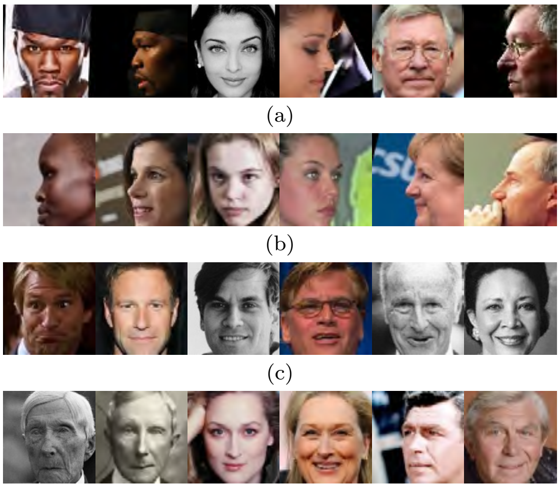
In Table 20 we present the TPR corresponding to 1e-5 and 1e-4 FPR values for the lightweight face models and the top-3 ranked methods of the competitions, where the first, second and third methods are "NothingLC", "Rhapsody"and "xfr", respectively. Similar to the case of DeepGlint-Light challenge, the best solutions for large-scale video face recognition, result from using lightweight face models guide by large teacher models through a knowledge distillation method, and including quality-aware methods to aggregate the features from different video frames. Among the lightweight face models considered in our study, VarGFaceNet and MobileFaceNet achieve the best verification results followed by ShuffleFaceNet, MobileFaceNetV1 and ProxylessFaceNAS.

Table 20: Verification accuracy (%) on the large-scale IQIYI-Video dataset.

Method	FPR=1e-5	FPR=1e-4
NothingLC [15]	**49.15**	**63.23**
Rhapsody [15]	46.78	61.87
xfr [15]	46.95	61.05
MobileFaceNet	31.13	46.93
VarGFaceNet	32.03	46.43
ShuffleFaceNet	30.14	44.55
MobileFaceNetV1	26.17	39.58
ProxylessFaceNAS	10.45	21.5

4.2.3 Cross-Factor Face Recognition

Due to the complex nonlinear facial appearance, some of its variations will be caused by people themselves, such as cross-pose and cross-age, which remain a major challenge in face recognition community. CFP-FP [78] and Cross-Pose LFW (CPLFW) [115] are used in this work to evaluate the performance of the face models in the cross-pose problem, while Cross-Age LFW (CALFW) [116] and AgeDB-30 [69] databases are used for evaluating the networks on age invariant face recognition. Figure 4 shows some examples face images of these databases.


(a)

(b)

(c)

(d) AgeDB-30

Fig. 4: Examples face images from (a) CFP-FP, (b) CPLFW, (c) CALFW and (d) AgeDB-30 databases containing pose and age variations.

Results on CFP-FP. The Frontal-Profile (FP) face verification experiment from CFP dataset [78] includes 350 same-person pairs and 350 different-person pairs for each of 10 splits. In Table 21 we present the verification results in terms of mean Accuracy, Equal Error Rate (EER) and Area Under Curve (AUC) for the lightweight models and state-of-the-art methods. It can be seen that although DDL method achieves the best performance, VarGFaceNet, MobileFaceNet and ShuffleFaceNet reach competitive results, outperforming methods specifically designed for the task at hand such as DR-GAN [86], strong very deep models like ResNet50-ArcFace [14] and other lightweight models including Seesaw-shuffleFaceNet [110] and AirFace [53]. MobileFaceNetV1 has a very similar performance, while ProxylessFaceNAS degrades a little bit, but its results are still better than those reported in previous works.

Table 21: Verification results (%) on CFP-FP database.

Method	Frontal-Profile		
	Accuracy	EER	AUC
HoG + Sub-SML [78]	77.3	22.2	86.0
FV + Sub-SML [78]	80.6	19.3	88.5
Deep features [78]	84.9	14.9	93.0
Human [78]	94.6	5.0	98.9
FV-DCNN [8]	91.9	8.0	97.7
Seesaw-shuffleFaceNet [110]	92.9	-	-
DR-GAN [86]	93.9	-	-
AirFace [53]	94.5	-	-
ResNet50-ArcFace [14]	95.6	-	-
ResNet100-ArcFace [14]	98.4	-	
DDL [42]	**98.5**	-	-
VarGFaceNet	96.9	**3.3**	**99.1**
MobileFaceNet	96.9	3.6	**99.1**
ShuffleFaceNet	96.3	4.1	99.0
MobileFaceNetV1	95.8	4.7	98.8
ProxylessFaceNAS	94.7	7.1	95.6

Results on CPLFW and CALFW. CPLFW [115] and CALFW [116] are recently introduced datasets that show higher pose and age variations, respectively, with same identities from LFW database. Similar to the original LFW dataset, both CPLFW and CALFW define an evaluation protocol with 10 individual subsets of image pairs. Each subset has 300 positive pairs and 300 negative pairs. Table 22 presents the comparison of face verification accuracy on CPLFW and CALFW datasets. As it can be appreciated, DDL and ResNet100-ArcFace models achieve the bests results for both datasets. Nonetheless, lightweight face models improve human results as well as well-established face models such as VGG-Face2 and SphereFace by an obvious margin. Among the selected lightweight face models, MobileFaceNet achieve the highest accuracy, followed by VarGFaceNet, ShuffleFaceNet, MobileFaceNetV1 and ProxylessFaceNAS.

Results on AgeDB-30. The AgeDB is an in-the-wild dataset with large variations in pose, expression, illuminations, and age [69]. It contains 16,488 images of 568 distinct subjects, such as actors/actresses, writers, scientists, and politicians. The minimum and maximum age is 1 and 101, respectively. The average age range for each subject is 50.3 years. There are four groups of test data with different year gaps (5, 10, 20 and 30 years, respectively) for age-invariant face verification. Each group has ten splits of face images, and each split contains 300 positive examples

Table 22: Face verification accuracy (%) on CPLFW and CALFW databases.

Method	CPLFW	CALFW
Human-Individual	81.21	82.32
Human-Fusion	85.24	86.50
CenterLoss [95]	77.48	85.48
SphereFace [56]	81.40	90.30
VGG-Face2 [6]	84.00	90.57
ResNet100-ArcFace [14]	92.08	**95.45**
DDL [42]	**93.43**	-
MobileFaceNet	89.22	95.15
VarGFaceNet	88.55	95.15
ShuffleFaceNet	88.50	95.05
MobileFaceNetV1	87.17	94.47
ProxylessFaceNAS	84.17	92.55

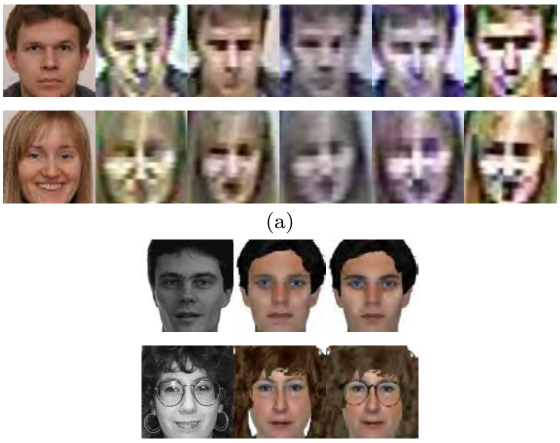
and 300 negative examples. The face verification evaluation metric used is Accuracy, as in LFW.In this paper, we only use the most challenging subset, AgeDB-30, to report the performance. In Table 23 we compare the verification accuracy obtained by the lightweight networks with state-of-the-art results reported in the literature. As we can see, MobileFaceNet is the top-ranked face recognition model, followed by VarGFaceNet and ShuffleFaceNet which obtain very close results.

Table 23: Verification accuracy (%) on AgeDB-30 database.

Method	Accuracy
VGG-Face [73]	85.1
CenterLoss [95]	90.7
ResNet50-ArcFace [14]	95.2
Marginal Loss [17]	95.7
Seesaw-shuffleFaceNet [110]	96.9
MobileFaceNet	**97.6**
VarGFaceNet	97.5
ShuffleFaceNet	97.3
MobileFaceNetV1	96.4
ProxylessFaceNAS	94.4

4.2.4 Heterogeneous Face Recognition

Heterogeneous face recognition refers to the problem of matching faces across different visual domains. The domain gap is mainly caused by differences in the origin of the face images, that can be due to sensory devices and cameras settings, or to the way of obtaining the images such as the facial sketches. Specifically, we asses the performance of the lightweight models on the low-resolution and the photo-sketch face recognition domains, by using the SCface database [26] and the extended UoM-SGFS [24], respectively. See Figure 5 for some examples images of these datasets.



(a)

(b)

Fig. 5: Examples face images from (a) SCface and (b) UoM-SGFS heterogeneous databases.

Results on SCface. The SCface database [26] contains images of 130 subjects, taken in uncontrolled indoor environments using five video surveillance cameras of various qualities. For each subject, there are 15 images taken at three different distances (five images at each distance), 4.2m (d1), 2.6m (d2) and 1.0m (d3), by five surveillance cameras, and one frontal mugshot image taken by a digital camera. For the experimental settings SCface defines face identification, where frontal mugshot images (high-resolution faces) are employed as gallery, and low-resolution face images taken by surveillance cameras at distance di; i = 1; 2; 3 are used as probes. Note that in this study we do not fine-tuned our lightweight models with target dataset. However, in order to be able for comparing our results with previous works, we follow the protocol of [60, 106] and report the identification performance for 80 subjects out of 130.

In Table 24 we compare the recognition rates at rank-1 of lightweight face models with state-of-the-art methods. For a fair comparison, we do not include deep models fine-tuned on the SCface training set. From the table we can observe that the best result for d1 (the largest distance) is obtained by MobileFaceNet followed by ProxylessFacceNAS, outperforming strong deeper models such as VGG-Face-FT [60] and ResNet50-ArcFace-FT [106] and ResNet100-ArcFace [14]. Although ShuffleFaceNet and MobileFaceNetV1 achieve lower results, they obtain a competitive accuracy, improving the performance of many deep models. For the other two distances which are relatively high-resolution images, most of the tested methods including the lightweight face architectures achieve much better results.

Results on UoM-SGFS. The University of Malta Software-Generated face Sketch database (extended UoM-SGFS) [24] is, to the best of our knowledge, the largest database for evaluating facial sketch recognition. It contains two sets of sketches from 600 subjects. The gallery is composed by a face image from every subject taken from the color FERET face database. Set A has 600 sketches created by using EFIT-V software, while Set B contains sketches in A with some image editions

Table 24: Identification results at rank-1 (%) on SCface database.

Method	d1 (4.2m)	d2 (2.6m)	d3 (1.0m)
DCA [29]	12.2	18.4	25.5
RICNN [108]	23.0	66.0	74.0
LDMDS [104]	62.7	70.7	65.5
LightCNN [99]	35.8	79.0	93.8
Center Loss [95]	36.5	81.8	94.3
VGG-Face [73]	41.3	75.5	88.8
ResNet50-ArcFace [14]	48.0	92.0	99.3
ResNet100-ArcFace [14]	58.9	**98.3**	99.5
FAN [106]	62.0	90.0	94.8
MobileFaceNet	**68.3**	97.0	**99.8**
ProxylessFaceNAS	67.3	95.3	98.0
VarGFaceNet	59.5	96.8	99.8
MobileFaceNetV1	57.0	95.3	99.8
ShuffleFaceNet	55.5	95.3	99.3

to emulate the process performed by law enforcement specialists. We follow the original protocol of the database where five random train/test set splits are obtained and the mean Recognition Rate (RR) at Rank N is reported. On each split, 450 subjects are used for training and 150 for testing. The results of the lightweight face models are compared with those obtained by the models reported in [68]. It can be seen from Table 25 that, in general, the lightweight models achieve state-of-the-art results, and perform even better than standard very deep models. They are only overcame by the DEEPS model [25], which is based on a fine-tuning of the standard VGG-Face, with a large number of photo-sketch pairs. Specifically, among the lightweight models ProxylessFaceNAS reach the highest recognition rates, follwed by VarGFaceNet, MobileFaceNetV1, ShuffleFaceNet and MObileFaceNet. Nonetheless, we can see that the recognition rates are still low for the task at hand.

Table 25: Identification results at different ranks on UoM-SGFS database.

Method	SetA			SetB		
	Rank-1	Rank-10	Rank-50	Rank-1	Rank-10	Rank-50
VGG-Face [68]	20.40	55.33	81.33	30.13	66.67	88.47
Dlib [68]	28.27	74.93	97.33	42.13	84.53	99.47
DEEPS [25]	**78.40**	**97.73**	**99.47**	**82.40**	**98.80**	**99.87**
MobileFaceNet	30.13	70.53	93.60	53.87	89.73	97.60
ShuffleFaceNet	30.13	71.60	94.40	49.40	88.93	98.67
MobileFaceNetV1	33.33	75.33	95.60	51.73	92.27	99.33
VarGFaceNet	34.26	73.33	94.67	52.53	88.13	98.40
ProxylessFaceNAS	35.47	76.27	95.60	55.20	91.20	99.20

4.2.5 Active Authentication on Mobile Devices

The University of Maryland Active Authentication Dataset (UMDAA-01) [23] is a benchmark for evaluating the effectiveness of face recognition techniques when used for face-based continuous authentication. The benchmark dataset contains 750 face videos captured by the front camera (and also screen touch data) of 50 distinct users, with three different illumination settings. The Session 1 was captured with artificial lighting, videos in Session 2 were obtained without any illumination and in Session 3 under natural sunlight. For each subject five videos are available in each session: one with different face poses used for enrollment and four test videos captured while the user was performing a specific activity, such as looking at a window popup, scrolling test, taking a picture or working on a document. Figure 6 shows some face images from UMDAA-01 for the different activities.

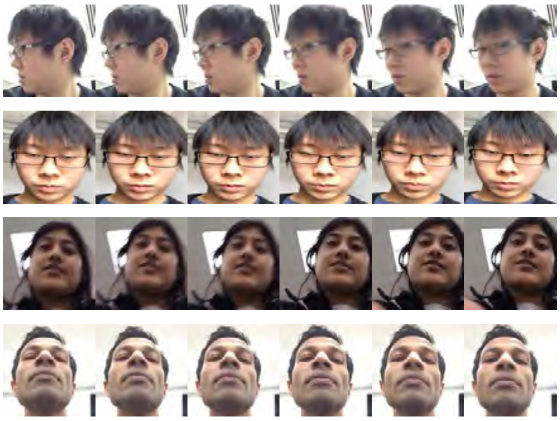


Fig. 6: Examples face frames from UMDAA-01 database.

Two evaluation protocols are provided to reflect some of the challenges of a typical face-based active authentication system. In this work, we follow protocol 1, which is the most difficult one. Under this protocol, the gallery is composed by the enrollment videos from one session (e.g. Session 1) and the methods are tested on the four non-enrollment video clips from the other two sessions (e.g. Sessions 2 and 3). Hence, there are six available scenarios for this protocol considering all pairs of session combinations. The Rank-1 Recognition Rates (RR) for all session combinations and their average values (Avg.) are presented in Table 26. It can be seen that the five lightweight models achieve impressive results, outperforming by a great margin the traditional methods evaluated on this protocol before.

Recently, in [59] it was evaluated the benefits of using just those frames with the most relevant information from a video sequence instead of using all video frames. For this, a quality value is estimated for each frame by measuring four parameters: pose, eyes, mouth and blur. The authors evaluated three different CNN models with a softmax function as classifier and obtained the best results by selecting

Table 26: Rank-1 recognition rates on UMDAA-01 database.

Method	S1-S2	S1-S3	S2-S1	S2-S3	S3-S1	S3-S2	Avg.
FF [23]	54.48	45.27	25.52	56.80	24.77	56.01	43.80
SRC [23]	52.79	51.18	44.18	58.58	17.64	51.95	46.05
MSSRC [23]	47.21	46.15	43.06	60.36	17.64	45.85	43.38
MobileFaceNet	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
ShuffleFaceNet	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
VarGFacenet	99.00	99.50	**100.0**	99.00	**100.0**	**100.0**	99.58
MobileFaceNetV1	99.50	99.50	**100.0**	99.00	**100.0**	**100.0**	99.67
ProxylessFaceNAS	97.50	**100.0**	**100.0**	97.50	**100.0**	**100.0**	99.17

the best ten frames. In this work, we use their proposed quality measure and test the lightweight models for the best ten frames. Table 27 shows the obtained RR at Rank-1 for each session combination and compare them with the results of the three models evaluated in [59]. The average (Avg.) recognition rate for each model through all sessions is also presented in this table. It can be seen that MobileFaceNet and ShuffleFaceNet have a very similar performance and their results are very close to those of the deeper ResNet50 model used in [59]. Although the recognition rates of VarGFaceNet, MobileFaceNetV1 and ProxylessFaceNAS are lower, they are competitive. Notice that, this strategy of using the best ten frames degrades a little bit the recognition rates, however, it is supposed to be more efficient than using all the video frames.

Table 27: Rank-1 recognition rates on UMDAA-01 dataset, using the best 10 frames.

Method	S1-S2	S1-S3	S2-S1	S2-S3	S3-S1	S3-S2	Avg.
Dlib [59]	93.18	84.09	88.64	81.82	90.91	88.64	87.88
MobileNet [59]	92.44	98.26	93.02	94.19	97.09	90.70	94.28
ResNet50 [59]	98.26	99.42	**100.0**	**100.0**	**100.0**	**100.0**	**99.61**
MobileFaceNet	**100.0**	**99.50**	**100.0**	98.50	99.50	**100.0**	99.58
ShuffleFaceNet	**100.0**	**99.50**	99.50	97.50	**100.0**	**100.0**	99.42
VarGFaceNet	97.50	99.50	**100.0**	98.00	**100.0**	99.50	99.08
MobileFaceNetV1	98.67	98.67	98.84	98.26	97.67	97.67	98.30
ProxylessFaceNAS	97.00	96.00	95.50	94.00	96.50	94.50	95.58

4.3 Ablation study: Effect of using common lightweight architectures

In this section we investigate the effect of using common lightweight architectures designed for general computer vision tasks, without modifying them for the case of face recognition. Specifically, we use the original networks of the lightweight face models considered in our study and compare their performance on the specific face recognition scenarios. For a fair comparison, we train all the original models (VarGNet [112], MobileNetV2 [76], ShuffleNetV2 [61], MobileNetV1 [36] and

ProxylessNAS [5]) under the same training settings used for their face versions as it was described in Section 4.1. In the experiments we will use * to refer to the original models implemented by ourselves.

The verification accuracy on the efficient validation datasets LFW, CFP-FP and AgeDB-30 is shown in Table 28. We can see that, in general, the lightweight face models achieve better results than their original versions. Specifically, in the LFW dataset, both the original and the face models perform very similar, due to that in this database the recognition performance is almost saturated. However, in front of pose (results on CFP-FP) and age (AgeDB-30) variations, we can observe greater improvements by the face architectures, which obtain the best accuracy results.

Table 28: Verification accuracy of original lightweight models and their corresponding face versions on LFW, CFP-FP and AgeDB-30 databases.

Method	LFW	CFP-FP	AgeDB-30
VarGNet*	98.0 ± 0.7	85.5 ± 1.5	89.6 ± 2.0
VarGFaceNet	**99.7 ± 0.3**	**96.9 ± 0.8**	**97.5 ± 0.7**
MobileNetV2*	99.3 ± 0.5	90.5 ± 1.6	92.3 ± 1.9
MobileFaceNet	**99.7 ± 0.3**	**96.9 ± 0.8**	**97.6 ± 0.6**
ShuffleNet*	99.5 ± 0.4	96.2 ± 1.1	94.8 ± 1.1
ShuffleFaceNet	**99.7 ± 0.3**	**96.9 ± 0.7**	**97.3 ± 0.8**
MobileNetV1*	99.1 ± 0.5	93.7 ± 1.2	93.9 ± 1.2
MobileFaceNetV1	**99.4 ± 0.4**	**95.8 ± 0.6**	**96.4 ± 1.1**
ProxylessNAS*	96.3 ± 0.9	83.6 ± 1.4	82.6 ± 1.6
ProxylessFaceNAS	**99.2 ± 0.5**	**94.7 ± 1.7**	**94.4 ± 1.5**

In order to give more evidences of the superiority of lightweight face models, we assess the performance of the original versions in several of the benchmark datasets used covering all the specific recognition scenarios that have been addressed in this study. Specifically, we report the identification rates at Rank-1 on MegaFace and IJB-B databases for Image FR; the verification accuracy on YTF and the Rank-1 on COX Face (setting V1-V2) for Video FR; the verification accuracy on CPLFW and CALFW for Cross-Factor FR; the identification rate at Rank-1 on SCface (d1) for Heterogeneous FR; and the average accuracy on UMDAA-01 database (using the best ten frames) for Active Authentication (AA) on Mobile Devices (UMDAA-10).

Table 29 and Table 30 compare the results obtained by the original models and their corresponding face versions on each dataset of the different FR scenarios. We can observe that, in general, all the lightweight face architectures boost the performance of their original versions in all the face recognition scenarios, especially for the case of ProxylessFaceNAS. The bigger improvements are achieved for large-scale image FR (results on MegaFace and IJB-B), and heterogeneous FR (results on SCface). These results show that modifications introduced in common lightweight architectures such as applying PReLU instead of ReLU and replacing GAP layer, are more suitable for face recognition tasks since they allow us to extract more essential information and achieve better performance.

Table 29: Performance comparison between original lightweight architectures and their face version in image and video FR scenarios.

	Image FR		Video FR	
	MegaFace (Id@Rank-1)	IJB-B (Rank-1)	YTF (Acc.)	COX/V1-V2 (Rank-1)
VarGNet*	72.0	83.0	92.7	94.3
VarGFaceNet	**78.2**	**94.0**	**96.0**	**98.0**
MobileNetV2*	71.4	83.4	93.9	91.8
MobileFaceNet	**79.3**	**94.0**	**96.2**	**96.8**
ShuffleNetV2*	69.7	89.0	93.3	93.4
ShuffleFaceNet	**77.4**	**93.6**	**95.7**	**97.8**
MobileNetV1*	71.0	89.4	92.1	93.1
MobileFaceNetV1	**76.0**	**93.2**	**95.2**	**96.9**
ProxylessNAS*	31.1	68.5	89.2	67.8
ProxylessFaceNAS	**69.7**	**90.7**	**94.4**	**96.9**

Table 30: Performance comparison between original lightweight architectures and their face version in cross-factor and heterogeneous FR, as well as active authentication FR scenarios.

	Cross-Factor FR		Heterogeneous FR	AA FR
	CPLFW (Acc.)	CALFW (Acc.)	SCface-d1 (Rank-1)	UMDAA-10 (Avg. Acc.)
VarGNet*	81.0	91.2	47.5	96.4
VarGFaceNet	**88.6**	**95.2**	**59.5**	**99.1**
MobileNetV2*	80.8	90.6	46.8	90.1
MobileFaceNet	**89.2**	**95.2**	**68.3**	**99.6**
ShuffleNetV2*	86.0	93.2	50.3	94.4
ShuffleFaceNet	**88.5**	**95.1**	**55.5**	**99.4**
MobileNetV1*	84.5	92.2	52.0	96.0
MobileFaceNetV1	**87.2**	**94.5**	**57.0**	**98.3**
ProxylessNAS*	73.2	93.8	33.5	71.0
ProxylessFaceNAS	**84.2**	**92.6**	**67.3**	**95.6**

4.4 Computational efficiency assessment

With the emergence of mobile phones, tablets and augmented reality, FR has been applied in mobile devices. Due to computational limitations, the recognition tasks in these devices need to be carried out in a light but timely fashion. In this section we aim at demonstrate the great advantages of the benchmarked lightweight face models to be deployed in real-time applications or on resource-limited devices.

In Table 31 we present the computational requirements for each of the analyzed lightweight architecture, and compare them with their original versions, as well as with some state-of-the-art face recognition models in terms of the model size in Megabytes (MB), Floating Point Operations Per Second (FLOPs), and number of parameters (#Param.). Besidees, the inference time per image by using different hardware settings including an Intel i7-7700HQ Laptop CPU (4 cores, 8 threads, 2.80GHz base freq.), a Nvidia Quadro P2000 Desktop GPU (5Gb GDDR5, 1024 CUDA cores), a Nvidia GeForce GTX 1050Ti Laptop GPU (4Gb GDDR5, 768

CUDA cores) and a Nvidia GeForce GTX 1660Ti Desktop GPU (6Gb GDDR5, 1536 CUDA cores), are presented.

Table 31: Comparison of storage space, complexity and inference time on different devices of the lightweight face architectures with their original versions as well as state-of-the-art face recognition models.

Method	Storage and complexity			Inference time (ms)			
	Model size	FLOPs (G)	#Param. (M)	Intel i7-7700HQ	Quadro P2000	1050 Ti	1660 Ti
VGG-Face [73]	526	15	138	1,523.4	25.2	35.3	108.9
VGG-Face2 [6]	165	4.0	25.6	433.4	19.6	21.6	14.2
ResNet100-ArcFace [14]	250	24.2	65.2	285.6	48.1	59.6	13.0
Light CNN-4 [99]	26	1.5	4.1	2653.8	41.0	55.5	14.1
Light CNN-9 [99]	32	1.0	5.6	2106.7	40.9	56.2	15.8
Light CNN-29 [99]	125	3.9	12.6	126.7	7.9	7.9	2.4
MobileNetV2*	**7.5**	**0.5**	**1.8**	103.7	11.1	14.8	4.5
ShuffleNetV2*	10.1	0.6	2.5	33.0	5.3	12.3	2.8
MobileNetV1*	12.6	1.1	3.2	99.0	5.6	20.0	1.8
VarGNet*	21.6	1.0	5.5	95.3	5.5	13.0	4.0
ProxylessNAS*	12.2	0.9	3.1	150.2	6.0	8.3	2.9
MobileFaceNet	8.2	0.9	2.0	62.4	5.5	7.3	3.3
ShuffleFaceNet	10.5	0.6	2.6	**29.1**	**4.7**	**4.7**	1.9
MobileFaceNetV1	13.1	1.1	3.4	53.5	4.9	8.2	**1.6**
VarGFaceNet	20.0	1.0	5.0	126.6	5.1	27.1	3.5
ProxylessFaceNAS	12.5	0.9	3.2	157.3	6.4	9.0	3.1

Regarding the original lightweight architectures (MobileNetV2*, ShuffleNetV2*, MobileNetV1* and ProxylessNAS*), we can appreciate that although they have a slightly lower memory footprint and less number of parameters than their corresponding face versions, their inference times are bigger. For example, in the cases of MobileNetV1* and MobileNetV2* the CPU runtimes are reduced by a factor of 2×, which is of paramount importance for real-time applications. In general, the best inference times are obtained by ShuffleFaceNet, followed by the MobileFaceNetV1.

On the other hand, compared with state-of-the-art face models, lightweight architectures exhibit significant improvements especially in the inference speed and storage requirements. In particular, we can see in the experiments above that ResNet100-ArcFace [14] is one of the best performing state-of-the-art models in the different evaluated scenarios, however, it demands high computational resources. For example, the biggest difference in accuracy between ResNet100-ArcFace and MobileFaceNet, is 8% in the very large-scale DeepGlint-Image dataset (one of the most challenging databases), while in the remaining databases is less than 3%. But regarding the computational complexity, ResNet100-ArcFace requires 19× more storage space and involves 26× more FLOPs and 32× more parameters than MobileFaceNet. Moreover, MobileFaceNet is at least 4× faster than ResNet100-ArcFace in a Laptop CPU Intel i7-7700HQ.

5 Summary and discussion

In this paper, we presented a comprehensive analysis and evaluation of lightweight face networks specifically designed for face recognition under five specific scenarios including image FR, video FR, cross-factor FR (e.g. pose and age) and heterogeneous FR (e.g. low-resolution and photo-sketch), as well as active authentication in mobile devices. First, we reviewed different strategies that have been proposed in the literature in order to design efficient network architectures and we see that using lightweight face recognition models is a topic that has gained the attention of the research community in the last years. Therefore, three successful lightweight deep face models recently introduced specifically for face recognition, namely VarGFaceNet [103], MobileFaceNet [9] and ShuffleFaceNet [63], were selected to extract facial image representations. In addition, motivated by the strategies used on these architectures, we propose to modify the MobileNetV1 [36] and ProxylessNAS [5] architectures to enhance its discriminative ability for the specific case of face recognition, resulting two new lightweight face models named MobileFaceNetV1 and ProxylessFaceNAS, respectively. Then, various well-known face benchmarks were employed for the experiments to analyze the overall performance of these five lightweight face models and the impact of different variations present on the five specific scenarios.

The conducted experimental evaluation shows that, in general, lightweight face networks provide promising results for face recognition. They are able to perform very similar to state-of-the-art very deep face models in most of the face recognition scenarios. These models share common design aspects in their architectures that we consider to be the key of their success in face recognition. These aspects include using PReLU non-linear activation function instead of ReLU to increase the discriminative ability of the networks, and replacing the GAP layer as the feature output layer to avoid the decrease of essential information. Among the tested lightweight face models, MobileFaceNet and VarGFaceNet achieved the best performance, followed by ShuffleFaceNet, MobileFaceNetV1 and ProxylessFaceNAS. In the case of MobileFaceNet it enhances the discriminative ability of MobileFaceNetV1 by using an inverted bottleneck structure. Meanwhile, VarGFaceNet fixes the number of channels in a group convolution, instead of fixing the total group numbers, as it is in ShuffleFaceNet. The general effectiveness of these modifications is demonstrated when comparing these models with their original lightweight networks without changing them for face recognition, where the improvement on their performance is significant. On the other hand, an analysis of the computational complexity of the different lightweight face models is included, by comparing them each other and with respect to state-of-the-art approaches. Overall, by analyzing the obtained results it is possible to determine the best solution for practical FR systems.

Overall, although lightweight face models have achieved a significant progress in face recognition, it remains a challenge the study of new emerged techniques in order to enhance their performance specially on those scenarios, where the achieved performance levels are not still as high as those from the state-of-the-art methods. We review some issues that still remain to be addressed, as well as some general remarks.

5.1 Lightweight FR Accuracy

From the benchmarking lightweight face models in specific face recognition scenarios, it is shown that this kind of architectures are able to obtain a performance as good as state-of-the-art methods based on complex deep learning models. High levels of accuracy are observed for efficient image (e.g. LFW, CFP-FF) and video (e.g. YTF) verification datasets, as well as for active authentication in mobile devices (results on UMDAA-01). However, for very large-scale image (results on Trillions-Pairs) and video (results on IQIYI-Video) datasets, as well as in front of several pose variations (results on CPLFW), the performance of lightweight face models drop, which suggest that in these scenarios additional techniques need to used for increase the performance of this kind of models.

On the other hand, when we tested the lightweight face models on heterogeneous scenarios, specifically in low-resolution (results on SCface) and facial sketch recognition (results on UoM-SGFS) domains, we reveal that the performance of these models are very similar and in some cases better (e.g. MobileFaceNet and ProxylessFaceNAS) than all existing methods that, like our models, have not been fine-tuned for these specific scenarios. However, the obtained accuracy levels are still lower and the main cause of this phenomenon is the large gap between the training and test data. This is why state-of-the-art results are usually based on training the networks with different modalities samples. In this sense, we asses the performance of the MobileFaceNet by fine-tuning (FT) on the SCface database. We follow the protocol used in [60, 106] where 50 out of 130 subjects are randomly chosen for fine-tuning the networks and the rest 80 subjects are used for testing. There is no identity overlap between the training and test sets. Table 32 compares the identification results for the three distances of MobileFaceNet-FT with state-of-the-models also fine-tuned on SCface training set. We can observed that after fine-tuning, all methods improve their accuracies considerably. MobileFaceNet-FT achieve the best performance for d1, which is the hardest. Although deep networks are robust to a degree of low-resolution, they depend on real surveillance data which are rather limited in size. It would be interesting to explore different transfer learning strategies but this is still a challenge not only for lightweight but also for deep models in general.

Table 32: Identification results at rank-1 (%) on SCface database.

Method	d1 (4.2m)	d2 (2.6m)	d3 (1.0m)
VGG-Face-FT [60]	46.3	78.5	91.5
LightCNN-FT [60]	49.0	83.8	93.5
Center Loss-FT [60]	54.8	86.3	95.8
ResNet50-ArcFace-FT [106]	67.3	93.5	98.0
DCR-FT [60]	73.3	93.5	98.0
FAN-FT [106]	77.5	95.0	98.3
ResNet100-ArcFace-FT [42]	80.5	98.0	99.5
DDL-FT [42]	93.2	99.2	98.5
MobileFaceNet-FT	**95.3**	**100.0**	**100.0**

5.2 Practical Applications

With the emergence of mobile phones, tablets and augmented reality, the deployment of face recognition models in these devices imposes to carry out this task in a light but timely fashion. In this work, we have shown that the usage of lightweight face models allow us not only recognizing faces in different scenarios with a competitive accuracy to state-of-the-art models but also with lower computational complexity. In particular, ShuffleFaceNet presents the lowest inference times, specially for lower-capacity devices (e.g. Laptop Intel i7), where it is 2× faster than MobileFaceNet and 4× faster than VarGFaceNet. However, MobileFaceNet and VarGFaceNet are the most accurate models in all the scenarios. Thus, depend on the application at hand, we recommended to use MobileFaceNet to those scenarios in which face recognition demands higher accuracy, while ShuffleFaceNet is indicated when it is possible to give up a little precision in exchange for a lower computational cost.

5.3 Future Insights

Recently, several approaches have emerged aim at improving the accuracy of lightweight deep networks. Among them, several works mainly have focused on how to design a more effective loss function (e.g. Li-ArcFace [53] and UniformFace [20]). However, we have included these proposals in the comparison of lightweight face models with state-of-the-art methods in the different scenarios (results on LFW, CFP-FP, MegaFace, YTF) and we observed that, in general, these new loss functions provide little improvement in accuracy; since their main robustness lie in a better convergence during training, leading with high locality and unbalance in feature distribution. On the other hand, knowledge distillation is being actively investigated due to its architectural flexibility and effectiveness for training lightweight models. Some works [21,103] have applied knowledge distillation during the training to enhance the interpretation ability of lightweight networks, while others focus on new distillation loss functions [42]. Although the great success in practice, there are not too many works on either the theoretical or empirical understanding of knowledge distillation. The teacher-student is the most common learning architecture to transfer the knowledge and how to select or design proper structures of teacher and student is very important but difficult problem. Besides, in the particular case of video-based face recognition, quality-aware aggregation methods [15] have been shown to be useful to improve the accuracy of the models.

Regarding to additionally improve the efficiency of lightweight networks, techniques such as pruning [33,52,58] and quantization [74,38,51], can be explored. However, this topic have not been fully covered in the literature, and we think that this deserves further attention.

6 Conclusion

In this paper we have presented a comprehensive survey of face recognition methods based on lightweight deep learning architectures. We have used five lightweight face models to conduct an extensive experimentation on 16 face recognition datasets grouped in five different scenarios. The obtained results show that the lightweight face architectures are able to achieve state-of-the-art performance in most of the evaluated scenarios. In some cases, such as low-resolution face recognition, further developments are needed. We showed that finetuning can helps to boost the performance on those cases, but also we provide some insights to additional techniques than can be explored in order to improve the efficiency and accuracy of lightweight models for some specific problems. In addition, the five lightweight face models were compared with their original versions in terms of accuracy and efficiency. It was shown that the face models retain the simplicity of their original lightweight versions, while significantly improve their accuracy on all the evaluated face recognition scenarios. We conclude that the models should be selected depending on the practical application at hand. From the evaluated models, MobileFaceNet and VargFaceNet are in general the most accurate ones, while ShuffleFaceNet is recommended when the computational resources are limited. The presented evaluation and analysis, can serve as a baseline for future research on this topic.

References

1. Trillionpairs. `http://trillionpairs.deepglint.com/overview` (2019). Accessed: 2020-07-23
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence (12), 2037–2041 (2006)
3. Balaban, S.: Deep learning and face recognition: The state of the art. In: Biometric and Surveillance Technology for Human and Activity Identification XII, vol. 9457, p. 94570B. International Society for Optics and Photonics (2015)
4. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on pattern analysis and machine intelligence **19**(7), 711–720 (1997)
5. Cai, H., Zhu, L., Han, S.: Proxylessnas: Direct neural architecture search on target task and hardware. CoRR **abs/1812.00332** (2018). URL `http://arxiv.org/abs/1812.00332`
6. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE (2018)
7. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: European conference on computer vision, pp. 109–122. Springer (2014)
8. Chen, J.C., Zheng, J., Patel, V.M., Chellappa, R.: Fisher vector encoded deep convolutional features for unconstrained face verification. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 2981–2985. IEEE (2016)
9. Chen, S., Liu, Y., Gao, X., Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: J. Zhou, Y. Wang, Z. Sun, Z. Jia, J. Feng, S. Shan, K. Ubul, Z. Guo (eds.) Biometric Recognition, pp. 428–438 (2018)
10. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z.: Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. CoRR **abs/1512.01274** (2015). URL `http://arxiv.org/abs/1512.01274`
11. Cheng, J., Wang, P.s., Li, G., Hu, Q.h., Lu, H.q.: Recent advances in efficient computation of deep convolutional neural networks. Frontiers of Information Technology & Electronic Engineering **19**(1), 64–77 (2018)
12. Courbariaux, M., Bengio, Y.: Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1. CoRR **abs/1602.02830** (2016). URL `http://arxiv.org/abs/1602.02830`
13. Courbariaux, M., Bengio, Y., David, J.: Binaryconnect: Training deep neural networks with binary weights during propagations. CoRR **abs/1511.00363** (2015). URL `http://arxiv.org/abs/1511.00363`

14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
15. Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S.: Lightweight face recognition challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)
16. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. CoRR **abs/1905.00641** (2019). URL `http://arxiv.org/abs/1905.00641`
17. Deng, J., Zhou, Y., Zafeiriou, S.: Marginal loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 60–68 (2017)
18. Denil, M., Shakibi, B., Dinh, L., Ranzato, M., De Freitas, N.: Predicting parameters in deep learning. In: Advances in neural information processing systems, pp. 2148–2156 (2013)
19. Ding, C., Tao, D.: Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
20. Duan, Y., Lu, J., Zhou, J.: Uniformface: Learning deep equidistributed representation for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3415–3424 (2019)
21. Duong, C.N., Luu, K., Quach, K.G., Le, N.: Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks. CoRR **abs/1905.10620** (2019). URL `http://arxiv.org/abs/1905.10620`
22. Duong, C.N., Quach, K.G., Le, N., Nguyen, N., Luu, K.: Mobiface: A lightweight deep learning face recognition on mobile devices. arXiv preprint arXiv:1811.11080 (2018)
23. Fathy, M.E., Patel, V.M., Chellappa, R.: Face-based active authentication on mobile devices. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1687–1691. IEEE (2015)
24. Galea, C., Farrugia, R.A.: Matching software-generated sketches to face photographs with a very deep cnn, morphed faces, and transfer learning. IEEE Transactions on Information Forensics and Security **13**(6), 1421–1431 (2017)
25. Galea, C., Farrugia, R.A.: Matching software-generated sketches to face photographs with a very deep cnn, morphed faces, and transfer learning. IEEE Transactions on Information Forensics and Security **13**(6), 1421–1431 (2018). DOI 10.1109/TIFS.2017.2788002
26. Grgic, M., Delac, K., Grgic, S.: Scface–surveillance cameras face database. Multimedia tools and applications **51**(3), 863–879 (2011)
27. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Computer Vision and Image Understanding p. 102805 (2019)
28. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. CoRR **abs/1607.08221** (2016). URL `http://arxiv.org/abs/1607.08221`
29. Haghighat, M., Abdel-Mottaleb, M.: Low resolution face recognition in surveillance systems using discriminant correlation analysis. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 912–917. IEEE (2017)
30. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149 (2015)
31. Harandi, M.T., Sanderson, C., Shirazi, S., Lovell, B.C.: Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In: CVPR, p. 27052712 (2011)
32. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
33. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: International Conference on Computer Vision (ICCV), pp. 1389–1397 (2017)
34. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
35. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1314–1324 (2019)

36. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
37. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141 (2018)
38. Hu, Q., Wang, P., Cheng, J.: From hashing to cnns: Training binary weight networks via hashing. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
39. Huang, G., Liu, S., Van der Maaten, L., Weinberger, K.Q.: Condensenet: An efficient densenet using learned group convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2752–2761 (2018)
40. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708 (2017)
41. Huang, G.B., Ramesh, M., Berg, T., Learned-miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments (2007)
42. Huang, Y., Shen, P., Tai, Y., Li, S., Liu, X., Li, J., Huang, F., Ji, R.: Improving face recognition from hard samples via distribution distillation loss. arXiv preprint arXiv:2002.03662 (2020)
43. Huang, Z., Shan, S., Wang, R., Zhang, H., Lao, S., Kuerban, A., Chen, X.: A benchmark and comparative study of video-based face recognition on cox face database. IEEE Transactions on Image Processing **24**(12), 5967–5981 (2015)
44. Huang, Z., Wang, R., Shan, S., Chen, X.: Learning euclidean-to-riemannian metric for point-to-set classification. In: CVPR, p. 16771684 (2014)
45. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
46. IBUG: Lightweight face recognition challenge & workshop (ICCV 2019). `https://ibug.doc.ic.ac.uk/` (2019). Accessed: 2019-05-09
47. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
48. Jain, A.K., Li, S.Z.: Handbook of face recognition, vol. 1. Springer (2011)
49. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4873–4882 (2016)
50. Lebedev, V., Ganin, Y., Rakhuba, M., Oseledets, I., Lempitsky, V.: Speeding-up convolutional neural networks using fine-tuned cp-decomposition. arXiv preprint arXiv:1412.6553 (2014)
51. Li, F., Zhang, B., Liu, B.: Ternary weight networks. arXiv preprint arXiv:1605.04711 (2016)
52. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710 (2016)
53. Li, X., Wang, F., Hu, Q., Leng, C.: Airface:lightweight and efficient model for face recognition. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (2019)
54. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
55. Liu, H., Simonyan, K., Yang, Y.: DARTS: differentiable architecture search. CoRR **abs/1806.09055** (2018). URL `http://arxiv.org/abs/1806.09055`
56. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212–220 (2017)
57. Liu, Y., Peng, B., Shi, P., Yan, H., Zhou, Y., Han, B., Zheng, Y., Lin, C., Jiang, J., Fan, Y., et al.: iqiyi-vid: A large dataset for multi-modal person identification. arXiv preprint arXiv:1811.07548 (2018)
58. Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C.: Learning efficient convolutional networks through network slimming. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2736–2744 (2017)
59. Llanes, N.M., Castillo-Rosado, K., Méndez-Vázquez, H., Khellat-Kihel, S., Tistarelli, M.: Face recognition on mobile devices based on frames selection. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings, pp. 316–325 (2019)

60. Lu, Z., Jiang, X., Kot, A.: Deep coupled resnet for low-resolution face recognition. IEEE Signal Processing Letters **25**(4), 526–530 (2018)
61. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. arXiv preprint arXiv:1807.11164 (2018)
62. Mandal, B., Lim, R.Y., Dai, P., Sayed, M.R., Li, L., Lim, J.H.: Trends in machine and human face recognition. In: Advances in Face Detection and Facial Image Analysis, pp. 145–187. Springer (2016)
63. Martindez-Diaz, Y., Luevano, L.S., Mendez-Vazquez, H., Nicolas-Diaz, M., Chang, L., Gonzalez-Mendoza, M.: Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (2019)
64. Martínez-Díaz, Y., Chang, L., Hernández, N., Méndez-Vázquez, H., Sucar, L.E.: Efficient video face recognition by using fisher vector encoding of binary features. In: ICPR, pp. 1436–1441 (2016)
65. Martínez-Díaz, Y., Hernandez, N., Biscay, R.J., Chang, L., Mendez-Vazquez, H., Sucar, L.E.: On fisher vector encoding of binary features for video face recognition. Journal of Visual Communication and Image Representation **51**, 155–161 (2018)
66. Martínez-Díaz, Y., Méndez-Vázquez, H., López-Avila, L., Chang, L., Enrique Sucar, L., Tistarelli, M.: Toward more realistic face recognition evaluation protocols for the youtube faces database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 413–421 (2018)
67. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB), pp. 158–165. IEEE (2018)
68. Mndez-Vzquez, H., Becerra-Riera, F., Morales-Gonzlez, A., Lpez-Avila, L., Tistarelli, M.: Local deep features for composite face sketch recognition. In: 2019 7th International Workshop on Biometrics and Forensics (IWBF), pp. 1–6 (2019). DOI 10.1109/IWBF.2019.8739212
69. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: The first manually collected, in-the-wild age database. pp. 1997–2005 (2017). DOI 10.1109/CVPRW.2017.250
70. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 343–347. IEEE (2014)
71. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3967–3976 (2019)
72. Parkhi, O.M., Simonyan, K., Vedaldi, A., Zisserman, A.: A compact and discriminative face track descriptor. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1693–1700 (2014)
73. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference, vol. 1, pp. 1–12 (2015)
74. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: Xnor-net: Imagenet classification using binary convolutional neural networks. In: European Conference on Computer Vision, pp. 525–542. Springer (2016)
75. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
76. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
77. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823 (2015)
78. Sengupta, S., Chen, J., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9 (2016). DOI 10.1109/WACV.2016.7477558
79. Sepas-Moghaddam, A., Pereira, F., Correia, P.L.: Face recognition: A novel multi-level taxonomy based survey. arXiv preprint arXiv:1901.00713 (2019)
80. Sun, Y., Liang, D., Wang, X., Tang, X.: Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
81. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)

82. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Closing the gap to human-level performance in face verification. deepface. In: IEEE Computer Vision and Pattern Recognition (CVPR), vol. 5, p. 6 (2014)
83. Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V.: Mnasnet: Platform-aware neural architecture search for mobile. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2820–2828 (2019)
84. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
85. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. arXiv preprint arXiv:1503.01817 (2015)
86. Tran, L.Q., Yin, X., Liu, X.: Representation learning by rotating your faces. IEEE transactions on pattern analysis and machine intelligence (2018)
87. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1365–1374 (2019)
88. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of cognitive neuroscience **3**(1), 71–86 (1991)
89. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
90. Wang, M., Deng, W.: Deep face recognition: A survey. arXiv preprint arXiv:1804.06655 (2018)
91. Wang, P., Cheng, J.: Accelerating convolutional neural networks for mobile applications. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 541–545 (2016)
92. Wang, Q., Guo, G.: Benchmarking deep learning techniques for face recognition. Journal of Visual Communication and Image Representation **65**, 102,663 (2019)
93. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification. In: CVPR, p. 24962503 (2012)
94. Wang, X., Wang, S., Zhang, S., Fu, T., Shi, H., Mei, T.: Support vector guided softmax loss for face recognition. arXiv preprint arXiv:1812.11317 (2018)
95. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision, pp. 499–515. Springer (2016)
96. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 90–98 (2017)
97. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: CVPR, pp. 529–534 (2011)
98. Wu, B., Wan, A., Yue, X., Jin, P., Zhao, S., Golmant, N., Gholaminejad, A., Gonzalez, J., Keutzer, K.: Shift: A zero flop, zero parameter alternative to spatial convolutions. arXiv preprint arXiv:1711.08141 (2017)
99. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13**(11), 2884–2896 (2018)
100. Xie, S., Zheng, H., Liu, C., Lin, L.: SNAS: stochastic neural architecture search. CoRR **abs/1812.09926** (2018). URL `http://arxiv.org/abs/1812.09926`
101. Xie, W., Shen, L., Zisserman, A.: Comparator networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 782–797 (2018)
102. Xie, W., Zisserman, A.: Multicolumn networks for face recognition. arXiv preprint arXiv:1807.09192 (2018)
103. Yan, M., Zhao, M., Xu, Z., Zhang, Q., Wang, G., Su, Z.: Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (2019)
104. Yang, F., Yang, W., Gao, R., Liao, Q.: Discriminative multidimensional scaling for low-resolution face recognition. IEEE Signal Processing Letters **25**(3), 388–392 (2017)
105. Yang, J., Ren, P., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), In press (2017)
106. Yin, X., Tai, Y., Huang, Y., Liu, X.: Fan: Feature adaptation network for surveillance face recognition and normalization. arXiv preprint arXiv:1911.11680 (2019)

107. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
108. Zeng, D., Chen, H., Zhao, Q.: Towards resolution invariant face recognition in uncontrolled scenarios. In: 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2016)
109. Zhang, J.: Seesaw-net: Convolution neural network with uneven group convolution. arXiv preprint arXiv:1905.03672 (2019)
110. Zhang, J.: Seesawfacenets: sparse and robust face verification model for mobile platform. arXiv preprint arXiv:1908.09124 (2019)
111. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
112. Zhang, Q., Li, J., Yao, M., Song, L., Zhou, H., Li, Z., Meng, W., Zhang, X., Wang, G.: Vargnet: Variable group convolutional neural network for efficient embedded computing. arXiv preprint arXiv:1907.05653 (2019)
113. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. arXiv preprint arXiv:1707.01083 (2017)
114. Zhang, X., Zou, J., He, K., Sun, J.: Accelerating very deep convolutional networks for classification and detection. IEEE transactions on pattern analysis and machine intelligence **38**(10), 1943–1955 (2015)
115. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying crosspose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Tech. Rep pp. 18–01 (2018)
116. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)
117. Zhi-Peng, F., Yan-Ning, Z., Hai-Yan, H.: Survey of deep learning in face recognition. In: 2014 International Conference on Orange Technologies, pp. 5–8. IEEE (2014)