

MixFaceNets: Extremely Efficient Face Recognition Networks

Fadi Boutros^{1,2}, Naser Damer^{1,2}, Meiling Fang^{1,2}, Florian Kirchbuchner¹, Arjan Kuijper^{1,2}

¹Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany

²Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany

Email: fadi.boutros@igd.fraunhofer.de

Abstract

In this paper, we present a set of extremely efficient and high throughput models for accurate face verification, MixFaceNets which are inspired by Mixed Depthwise Convolutional Kernels. Extensive experiment evaluations on Label Face in the Wild (LFW), Age-DB, MegaFace, and IARPA Janus Benchmarks IJB-B and IJB-C datasets have shown the effectiveness of our MixFaceNets for applications requiring extremely low computational complexity. Under the same level of computation complexity ($\leq 500M$ FLOPs), our MixFaceNets outperform MobileFaceNets on all the evaluated datasets, achieving 99.60% accuracy on LFW, 97.05% accuracy on AgeDB-30, 93.60 TAR (at FAR1e-6) on MegaFace, 90.94 TAR (at FAR1e-4) on IJB-B and 93.08 TAR (at FAR1e-4) on IJB-C. With computational complexity between 500M and 1G FLOPs, our MixFaceNets achieved results comparable to the top-ranked models, while using significantly fewer FLOPs and less computation overhead, which proves the practical value of our proposed MixFaceNets. All training codes, pre-trained models, and training logs have been made available <https://github.com/fdbtr/mixfacenets>.

1. Introduction

Recent advancements in the development of deep convolutional neural networks and the novelty of margin-based Softmax loss functions have significantly improved the State-Of-the-Art (SOTA) performance of face recognition to an extraordinary level [6], even under challenging conditions [5, 2]. However, deploying deep learning-based face recognition models in the embedded domains and other use-cases constrained by the computational capabilities and high throughput requirements is still challenging [19, 7].

Recently, a great progress in designing efficient face recognition solutions has been achieved by utilizing lightweight deep learning model architecture designed for common computer vision tasks such as MobileNetV2

[24], ShuffleNet [16], VarGNet [31] for face recognition [4, 15, 18, 29]. MobileFaceNet [4] was one of the earliest works that proposed an efficient face recognition model with around 1M of parameters and 439M FLOPs. MobileFaceNet architecture is inspired by MobileNetV2 [24]. The architecture of MobileNetV2 is based on an inverted residual structure and the depthwise separable convolution [10]. AirFace [15], ShuffleFaceNet [18] and VarGFaceNet [29] model architectures are built from MobileNetV2 [24], ShuffleNetV2 [16] and VarGNet [31], respectively, reaching high levels of accuracy using compact models with around 1G FLOPs computation complexity. ShuffleNetV2 utilizes channel shuffle operation proposed by ShuffleNetV1, achieving an acceptable trade-off between accuracy and computational efficiency. VarGNet [31] proposed to fix the number of input channels in each group convolution instead of fixing the total group numbers in an effort to balance the computational intensity inside the convolutional block. VarGFaceNet was the winner of the DeepGlint-Light track of ICCV Lightweight Face Recognition (LFR) Challenge (2019) [7]. The deepglint-light track of the LFR challenge targets face recognition for environment constraint by the computational complexity of 1G FLOPs and memory footprint of 20M (around 5M of trainable parameters). Very recently, MixNets proposed new convolution building block (MixConv) by extending vanilla depthwise convolution [10] with multiple kernel sizes. Using neural architecture search, MixNets developed highly efficient networks. MixNets outperformed previous mobile models including MobileNets [24, 10], ShuffleNetV2 [16] on image classification and object detection tasks.

In this work, we propose a set of extremely efficient architectures for accurate face verification and identification, namely the MixFaceNets. We opt to use MixNets as a baseline network structure to develop our MixFaceNets. We carefully designed tailored head and embedding settings that are suitable for face recognition. We also extend the MixConv block with a channel shuffle operation aiming at increasing the discriminative ability of MixFaceNets. With computation complexity of 451M FLOPs, our MixFaceNet-

S and ShuffleMixFaceNet-S achieved 99.60 and 99.56 % accuracies on Labeled Faces in the Wild (LFW) [12] and 92.23 and 93.60 TAR (at FAR1e-6) on MegaFace [14] which are significantly higher than the ones achieved by MobileFaceNets [4] with a comparable level of computational complexity (99.55% accuracy on LFW and 90.16% TAR (at FAR1e-6) on MegaFace). Also, our MixFaceNets achieve comparable results to the SOTA solutions that have computation complexity of thousands of MFLOPs.

2. Related Works

Common deep learning-based face recognition models are typically computationally expensive for deployment on low computational power devices and high throughput processes. This challenge has received increased attention in the literature in the last few years [19, 7]. In this section, we list out and discuss the recent efforts on designing an efficient deep learning model for face recognition. While there are a wide range of evaluation datasets for face recognition and each of the listed previous works reported the performance on a different set of evaluation datasets, we present in this section the reported accuracy by these previous works on the LFW [12] as it is the most commonly reported benchmark. We also account for the computational complexity of the list of works in terms of the number of trainable parameters and the FLOPs when it is feasible.

The Light CNN [28] was one of the earliest works that presented 3 network architectures for learning compact representation on a large-scale database. The proposed architectures- Light CNN-4, Light CNN-9, and Light CNN-29, contain 4, 5.5, and 12.6m trainable parameters and FLOPs of 1.5G, 1, and 3.9G, respectively. The best performance among these three architectures on the LFW dataset (99.33%) is achieved by the Light CNN-29. Compared to the recent efficient face recognition models [19], Light CNN architecture is considered computational expensive. ShiftNet [27] proposed shift-based modules as an alternative to spatial convolutions and it is adopted for the face recognition task. The presented face model contains 0.78m parameters and it achieved 96% accuracy on the LFW dataset. MobileFaceNets [4] is based on MobileNetV2 [24] and it achieved a high accuracy on LFW dataset with around one million of trainable parameter and 443M FLOPs. MobileFaceNets model architecture is based on the residual bottlenecks proposed by MobileNetV2 [24] and depth-wise separable convolutions layer, which allows building CNN with a smaller set of parameters compared to standard CNNs. Different from MobileNetV2 architecture, MobileFaceNet uses Parametric Rectified Linear Unit (PReLU) [9] as the non-linearity in all convolutional layers and replaces the last global average pooling with linear global depth-wise convolution layer as a feature output layer. The best-reported accuracy on LFW by Mobile-

FaceNets (1m parameters and 433M FLOPs) was 99.55%.

Several recent works utilized lightweight models designed for a common computer vision task to learn face representation [18, 29]. ShuffleFaceNet [18] is a compact face recognition model based on ShuffleNet [16]. Similar to MobileFaceNet [4], ShuffleFaceNet replaces the last global average pooling layer with a global depth-wise convolution layer and a Rectified Linear Unit (ReLU) with PReLU. ShuffleFaceNet presented four models with different complexity levels. The best-reported accuracy on LFW was 99.67% (2.6M parameters and FLOPs of 577.5M). VarGFaceNet [29] deployed variable group convolutional network proposed by VarGNet [31] to design a compact face recognition model with 5m trainable parameters and 1G FLOPs. VarGFaceNet adds squeeze and excitation (SE) block on the VarGNet block, replaces ReLU with PReLU, and uses variable group convolution along with pointwise convolution as the feature output layer. VarGFaceNet achieved 99.683 % accuracy on LFW. This accuracy is increased to 99.85 % by training VarGFaceNet with recursive knowledge distillation. However, the computational cost of VarGFaceNet is higher than the ones of ShuffleFaceNet and MobileFaceNet. AirFace [15] proposed to increase the MobileFaceNet network width and the depth and adding attention module, achieving 99.27% accuracy on LFW. The work also presented a loss function named Li-ArcFace which is based on ArcFace. Li-ArcFace demonstrates better converging and performance than ArcFace loss on low dimensional features embedding. The proposed model by AirFace has a computational cost of 1G FLOPs.

In a recent survey by Martinez-Diaz et al. [19], the computational requirements and the verification performance of five lightweight model architectures are analyzed and evaluated. The evaluated models are MobileFaceNet (0.9G FLOPs and 2.0m parameters), VarGFaceNet [29] (1G FLOPs and 5m parameters), ShuffleFaceNet [18] (577.5M FLOPs and 2.6m parameters), MobileFaceNetV1 (1.1G FLOPs and 3.4m parameters), and ProxylessFaceNAS (0.9G FLOPs and 3.2m parameters). MobileFaceNetV1, ProxylessFaceNAS and MobileFaceNet are extended versions of MobileNetV1 [10], ProxylessNAS [3], and MobileFaceNets [4], respectively. The evaluated models in [19]- ShuffleFaceNet, VarGFaceNet, MobileFaceNet achieved very close accuracy on LFW (99.7%) and other evaluated datasets, while MobileFaceNetV1 and ProxylessFaceNAS achieved slightly lower accuracy.

Among the previous listed works, MobileFaceNets [4] is the only architecture that achieved high accuracy with less than 500M FLOPs. With the almost same number of FLOPs as in MobileFaceNets, our MixFaceNets outperform MobileFaceNets and achieved competitive results to other models with fewer FLOPs using extremely efficient architecture as shown in Tables 1 and 2.

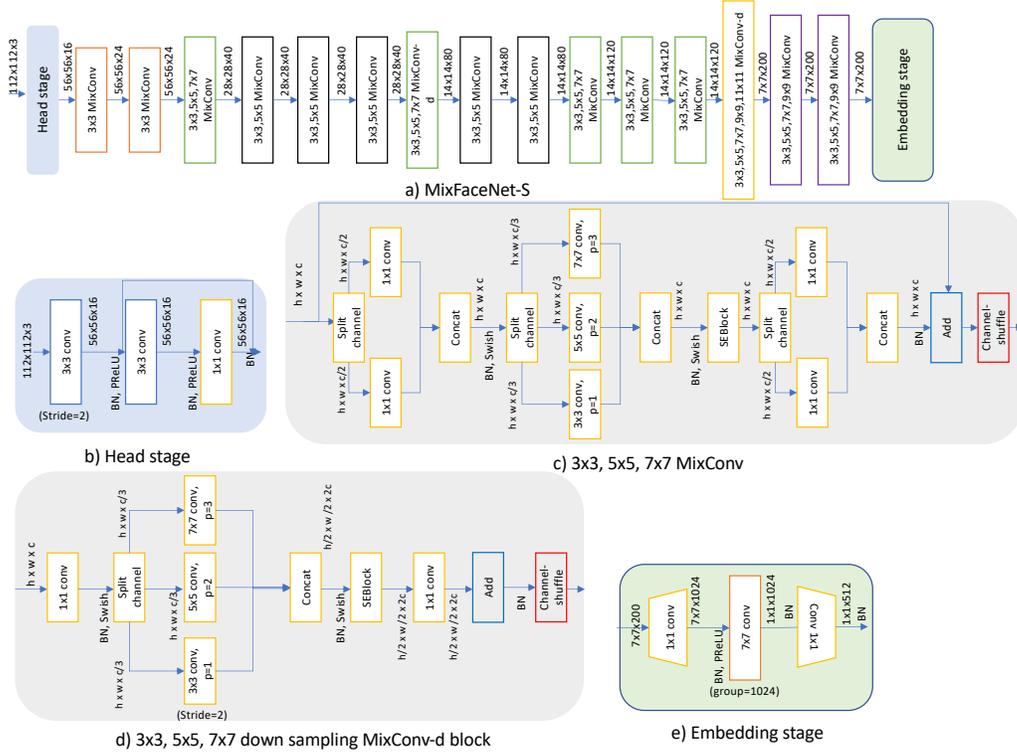


Figure 1: An overview of our proposed MixFaceNet-S network architecture inspired by MixNets [25]. The input of MixFaceNet-S has size of $112 \times 112 \times 3$ and the output is a face embedding of dimension $512 - d$. b) illustrates the head setting of MixFaceNet-S. We downsample the input on the first convolution (stride=2) then we add one residual block. d) is the MixConv block with multiple kernel sizes ($(3, 3)$, $(5, 5)$, $(7, 7)$) and channel shuffle operation. All MixConv blocks have the same structure as in (c) and the reduced blocks have the same structure as in (d). e) shows the embedding setting of MixFaceNet-S. We expand the channel from 200 to 1024, then we apply global depthwise convolution to obtain a $512 - d$ embedding. The input and output size, kernel size, stride, and padding (p) are shown for each convolution layer.

3. Approach

This section presents the architecture of our efficient MixFaceNets designed for accurate face verification. Figure 1 illustrates the architecture of the MixFaceNet, partially inspired by MixNets [25]. To improve the accuracy and the discriminative ability of MixNet, we (a) implement different head settings, (b) introduce channel shuffle operation to the MixConv block, and (c) propose different embedding settings. We discuss in this section the MixConv as an inspiration to our work. Then, we present the detailed architecture of our MixFaceNet.

3.1. Mixed Depthwise Convolutional Kernels

Depthwise Convolution is one of the most popular building block for mobile models [10, 24, 16]. Depthwise Convolutional applies a single convolution filter over each channel of input, thus it reduces the number of parameters and achieves computational efficiency while maintaining the discriminative ability of the convolution [10]. Mixed Depthwise Convolutional Kernels (MixConv) [25] extends vanilla depthwise convolution by using multiple

kernel sizes in a single convolution. MixConv depends on mixing up multiple kernel sizes in a single convolution by splitting up the input of convolution into groups and applying different kernel sizes to each group. Different from vanilla depthwise convolution, MixConv can capture different patterns from convolution input at various resolutions. Also, it requires fewer parameters and it is more computationally efficient than using a single kernel e.g. using multiple kernels of size $[(3, 3), (5, 5), (7, 7)]$ is more computation efficient than using a single kernel of size 7×7 . For example, given a convolution input of size $w \times h \times c$ and multiple kernels of size $[(3, 3), (5, 5), (7, 7)]$, MixConv split the input into 3 groups, each of them has a dimension of $w \times h \times c/3$. Then it uses different kernels for each of these groups. Finally, the three outputs are concatenated to produce the final convolution output. An example of MixConv and downsample-MixConv blocks are shown as part of Figure 1. Unlike manually designed mobile models [16, 18, 24], MixConv utilized neural architecture search to develop new series of MixConv-based networks, namely MixNets (MixNet-S, MixNet-M, and MixNet-L). MixNet-

S and MixNet-M are developed using neural architecture search, while MixNet-L is obtained by scaling up the number of channels in each block by a factor of 1.3. For details about the network structure and search space, we refer to the original work [25].

3.2. MixFaceNet Architecture

We deploy MixNets [25] as baseline network structure to develop our proposed MixFaceNets. Figure 1 illustrates the network architecture for MixFaceNet-S. For the network head, we apply fast down-sampling in the first 3×3 convolution (stride=2) followed by batch normalization [13] and PReLU non-linearity [9]. Then, we use one residual block as shown in Figure 1.b. For the MixFaceNet-S network, we use the same global structure as MixNet-S. However, different from MixNet-S, we did not apply down-sampling at the first convolution after the head stage to reserve as much as possible information at the earliest stage of the network. We propose to mix up both channels and kernels to increase the discriminative ability of MixFaceNet and improve the model accuracy. We achieve that by introducing shuffle operation to the MixConv block. The channel shuffle operation is proposed by [18] to enabling information flowing between different groups of channels. We apply a channel shuffle operation with a group value of 2 after each MixConv block. Thus, MixFaceNet can capture high and low-resolution patterns at different scales and it also enables information communication between various groups of channels. Figures 1.c and 1.d show the detailed structure of MixConv and down-sampling MixConv blocks with the channel shuffle operation. All MixConv blocks uses swish as an activation function [23] followed by batch normalization. MixConv includes also squeeze-and-excitation (SEBlock) [11] at the end of each block. Finally, to obtain the feature embedding of the input face image, we replace the last global average pooling layer with global depth-wise convolution as presented in the next section. We propose 3 network architectures, MixFaceNet-XS, MixFaceNet-S, and MixFaceNet-M. MixFaceNet-S network architecture is illustrated in Figure 1.a. MixFaceNet-XS is obtained by scaling up MixFaceNet-S with a depth multiplier of 0.5. MixFaceNet-M has the same global network architecture as MixNet-M [25] with the same strategies applied to MixFaceNet-S i.e. head setting and embedding settings. All proposed MixFaceNets are trained and evaluated with/without channel shuffle operation. The models trained with channel shuffle operation will be noted as ShuffleMixFaceNet-XS, ShuffleMixFaceNet-S, and ShuffleMixFaceNet-M.

3.2.1 Embedding Setting

MixNets use global average pooling before the classification layer as a feature output layer. This is common choice

for most of the classical compact deep learning models [10, 24, 16]. However, previous works in [6, 4] observed that CNNs with a fully connected layer (FC) or global depthwise convolution are more accurate than the ones with global average pooling for face recognition and verification. A fully connected layer has been used in many of the recent deep face recognition models to obtain face representations [6]. However, using FC on top of the last convolutional layer will add a large number of parameters to the model. And thus, it extremely increases the memory footprint and reduces the throughput. For example, giving the last convolutional layer of CNN with a kernel size of 7×7 (as in MixNet) and output feature maps of size 200, the output of this layer, in this case, has a size of $7 \times 7 \times 200$. Using FC of size $512 - d$ on top of the previous layer will add additional 5M parameters to the network ($7 \times 7 \times 200 \times 512$). Even for small FC, $128 - d$, the number of additional parameters causes by FC will be 1.2M. Thus, using FC is not the optimal choice for an efficient face recognition model. Using global depthwise convolution is a common choice for most of the previous works proposing efficient face recognition models as it contains fewer parameters than FC and it can lead to higher verification performance than using global average pooling [4]. Therefore, we replace the global average pooling with global depthwise convolution. Specifically, we first add 1×1 convolutional layer (Conv1) with stride=1 and zero paddings followed by batch normalization [13] and PReLU non-linearity [9]. In Conv1, we expand the channel from 200 to 1024. Then, we use 7×7 convolution layer (stride=1, padding=0 and grouping=1024) followed by batch normalization. Finally, we use 1×1 convolution with 512 output channels followed by batch normalization to obtain the final feature embedding which is of size $512 - d$, as shown in Figure 1.e.

4. Experimental Setup

Dataset: We use the MS1MV2 dataset [6] to train our MixFaceNet models. The MS1MV2 is a refined version of the MS-Celeb-1M [8] by [6] and it contains 5.8M images of 85K identities. The Multi-task Cascaded Convolutional Networks (MTCNN) solution [30] is used to detect and align face images. The MixFaceNet models process aligned and cropped face images of the size $112 \times 112 \times 3$ to produce $512 - d$ feature embeddings. We evaluate our MixFaceNets on the widely used LFW [12] and on the AgeDB-30 [21] datasets. Also, we report the performance of the MixFaceNets on large scale evaluation datasets including MegaFace [14], IJB-B [26] and IJB-C [20].

MixFaceNets Training Setup: The proposed models in this paper are implemented using Pytorch. All models are trained using ArcFace loss [6]. We set the margin value of ArcFace loss to 0.5 and the feature scale to 64. We set the batch size to 512 and train our model using dis-

Method	FLOPs (M)	# Params. (M)	LFW (%)	AgeDB-30 (%)
ArcFace (LResNet100E-IR) [6]	24211	65.2	99.83	98.15
AirFace [15]	1000	-	99.27	-
ShuffleFaceNet 2× [18]	1050	4.5	99.62	97.28
ShuffleFaceNet 1.5× [18]	577.5	2.6	99.67	97.32
VarGFaceNet [29]	1022	5	99.68	98.10
MobileFaceNet [19]	933	2.0	99.7	97.6
MobileFaceNetV1 [19]	1100	3.4	99.4	96.4
ProxylessFaceNAS [19]	900	3.2	99.2	94.4
MixFaceNet-M (ours)	626.1	3.95	99.68	97.05
ShuffleMixFaceNet-M (ours)	626.1	3.95	99.60	96.98
MobileFaceNets [4]	439.8	0.99	99.55	96.07
ShuffleFaceNet 0.5× [18]	66.9	0.5	99.23	93.22
MixFaceNet-S (ours)	451.7	3.07	99.60	96.63
ShuffleMixFaceNet-S (ours)	451.7	3.07	99.58	97.05
MixFaceNet-XS (ours)	161.9	1.04	99.60	95.85
ShuffleMixFaceNet-XS (ours)	161.9	1.04	99.53	95.62

Table 1: MixFaceNets verification accuracies on LFW and AgeDB-30 datasets. The first row of the table show the achieved result by the current SOTA ReNet100 models. The table is divided into two parts. The first part of the table shows the achieved result by models that have computational complexity between 500 and 1000M FLOPs. The second part of the table shows the achieved by models that have computational complexity less than 500M FLOPs. The number of decimal points is reported as in the related works.

tributed Partial-FC algorithm [1] on one machine with 4 Nvidia GeForce RTX 6000 GPUs to enable faster training on a single node. All models are trained with Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of $1e-1$. We set the momentum to 0.9 and the weight decay to $5e-4$. The learning rate is divided by 10 at 80k, 140k, 210k, and 280k training iterations. During the training, we evaluate the model on LFW and AgeDB after each 5650 training iterations. The training is stopped after 300k iterations. During the testing, the classification layer is removed and the feature is extracted from the last layer, which is of the size $512 - d$. We used euclidean distance between feature vectors in all experiments for comparison.

5. Results

This section presents the achieved result by the MixFaceNets on different benchmarks. We acknowledge the evaluation metrics in the ISO/IEC 19795-1 [17] standard. However, for the sake of comparability and reproducibility, we follow the evaluation metrics used in the utilized benchmarks and the previous works reporting on them.

5.1. Result on LFW and AgeDB-30

LFW [12] is one of the widely used datasets for unconstrained face verification. The dataset contains 13,233 images of 5749 different identities. The result on LFW is reported as verification accuracy (as defined in [12]) following the unrestricted with labeled outside data protocol using the standard 6000 comparison pairs defined in [12].

AgeDB [21] is common used in-the-wild dataset for evaluating age-invariant face verification. It contains 16,488 images of 568 different identities. We report the performance as verification accuracy for AgeDB-30 (years gap 30) as it is the most challenging subset of AgeDB. Also, it is the commonly reported set of AgeDB by the recent SOTA face recognition models. Similar to the LFW, we followed the standard protocol provided by AgeDB to evaluate our models on AgeDB-30 [21]. Table 1 shows the achieved result on LFW and AgeDB-30. We first report the result for one of the top-ranked face recognition models, ArcFace (LResNet100E-IR) [6], to give an indication of the current SOTA performance on LFW (99.83 %) and AgeDB-30(98.15%). Although, the ArcFace (LResNet100E-IR) model [6] is far from being considered an efficient model, in comparison to lightweight models, with 24211M FLOPs and 65.2m parameters. Then, the second section of Table 1 presents the achieved result by the recent lightweight models that have computational complexity between 500 and 1000M FLOPs. The best-reported result on LFW (99.70% accuracy) is achieved by the MobileFaceNet [19] (933M FLOPs). Our MixFaceNet-M achieved a competitive result on LFW (99.68% accuracy) using 38% fewer FLOPs (626M). A similar result has been achieved on AgeDB-30. Our MixFaceNets achieved very close accuracy to the current SOTA models using a more efficient model architecture with almost the same number of parameters. Among all models that have computational complexity less than 500M FLOPs, our MixFaceNet models outperform all listed mod-

Method	MFLOPs	Params (M)	MegaFace		MegaFace (R)		IJB	
			Rank-1 (%)	TAR at FAR1e-6	Rank-1 (%)	TAR at FAR1e-6	IJB-B	IJB-C
ArcFace (LResNet100E-IR) [6]	24211	65.2	81.03	96.98	98.35	98.48	94.2	95.6
AirFace [15]	1000	-	80.80	96.52	98.04	97.93	-	-
MobileFaceNet [19]	933	2.0	79.3	95.2	95.8	96.8	92.8	94.7
ShuffleFaceNet [19, 18]	577.5	2.6	77.4	93.0	94.1	94.6	92.3	94.3
MobileFaceNetV1 [19]	1100	3.4	76.0	91.3	91.7	93.0	92.0	93.9
VarGFaceNet [19, 29]	1022	5.0	78.20	93.9	94.9	95.6	92.9	94.7
ProxylessFaceNAS [19]	900	3.2	69.7	82.8	82.1	84.8	87.1	89.7
MixFaceNet-M (ours)	626.1	3.95	78.2	94.26	94.95	95.83	91.55	93.42
ShuffleMixFaceNet-M (ours)	626.1	3.95	78.13	94.24	94.64	95.22	91.47	93.5
MobileFaceNets [4]	439.8	0.99	-	90.16	-	92.59	-	-
MixFaceNet-S (ours)	451.7	3.07	76.49	92.23	92.67	93.79	90.17	92.30
ShuffleMixFaceNet-S (ours)	451.7	3.07	77.41	93.60	94.07	95.19	90.94	93.08
MixFaceNet-XS	161.9	1.04	74.18	89.40	89.35	91.04	88.48	90.73
ShuffleMixFaceNet-XS (ours)	161.9	1.04	73.85	89.24	88.823	91.03	87.86	90.43

Table 2: The achieved results on large-scale evaluation datasets- MegaFace, IJB-B, and IJB-C. The results on MegaFace and MegaFace (R) [6] using FaceScrub as probe set are reported as face identification (Rank-1 %) and verification (TAR at FAR1e-6) for different lightweight models. The last two columns of the table show 1:1 verification TAR (at FAR=1e-4) on IJB-B and IJB-C. The first row reports the evaluation result using SOTA face recognition model- ArcFace (LResNet100E-IR) which contains 65.2M parameters and 24211M FLOPs. The rest of the table is organized in two parts- models with computational complexity between 500 and 1000M FLOPs and models with less than 500M FLOPs. The number of decimal points is reported as in the related works. Considering the computation complexity, our MixFaceNet models are evaluated as ones of top-ranked models.

els including MobileFaceNets [4] on LFW and AgeDB-30. Similar conclusion can be seen in the Figure 2a and 2b. It can be clearly noticed that our MixFaceNet achieved the highest accuracies on LFW and AgeDB-30 when considering the same level of computational complexity.

5.2. Result on IJB-B and IJB-C

The IARPA Janus Benchmark-B ((IJB-B) face dataset consists of 1,845 subjects of 21,798 still images and 55,026 frames from 7,011 videos [26]. The IJB-B verification protocol provides a list of 10,270 genuine comparisons and 8M impostor comparisons. The IARPA Janus Benchmark-C (IJB-C) face dataset is an extension of IJB-B by increasing the database variability and size with additional 1,661 new subjects [20]. The IJB-C consists of 31,334 still images and 117,542 frames from 11,779 videos of 3531 subjects. The IJB-C verification protocol provides a list of 19,557 genuine comparisons and 15,638,932 impostor comparisons. The result on IJB-C and IJB-B is reported in terms of true accepted rates (TAR) at false accepted rates (FAR) (as defined in [26]) equal to 1e-4 to provide a comparable result with the previous works evaluated on these datasets. The achieved verification performances on IJB-B and IJB-C by our MixFaceNet models are reported as part of the Table 2. Our MixFaceNet-M and ShuffleMixFaceNet-M models achieved close results to the top-ranked models using significantly fewer FLOPs.

5.3. Result on MegaFace

The evaluation protocol of MegaFace includes gallery (1m images from Flickr) and probe (FaceScrub and FGNe)

sets. In this paper, we use MegaFace [14] as a gallery set and FaceScrub [22] as the probe set to provide a comparable result with the previous works evaluated on this dataset. The MegaFace [14] contains 1m images of 690K different identities and the FaceScrub contains 100K images of 530 identities [22]. The result on MegaFace is reported as identification (Rank-1) and verification (TAR at FAR=1e-6) to be compatible with the previous works evaluated on this dataset [19]. Also, we report the result on the refined version of the MegaFace presented in [6]. The face verification and identification results on the MegaFace and the refined version of MegaFace (noted as MegaFace (R)) are presented in Table 2. For all evaluated models that have computational complexity between 500 and 1000M FLOPs, our MixFaceNet-M outperformed ProxylessFaceNAS [19], VarGFaceNet [19], MobileFaceNetV1 [19], and ShuffleFaceNet [19, 18]. And it achieved very close verification and identification results to the top-ranked models- AirFace [15] and MobileFaceNet [19] using less than half the number of FLOPs. Also, when the considered computational cost is less than 500M FLOPs, in the third section of Table 2, our ShuffleMixFaceNet-S achieved the highest verification and identification performances.

5.4. Performance vs. Computational Complexity

To present the achieved results in terms of the trade-off between the verification performance and the computation complexity (represented by the number of FLOPs), we plot the number of FLOPs vs. the verification performance of our MixFaceNets and the SOTA solutions. The plots for the comparisons on the LFW, AgeDB-30, MegaFace, IJB-B,

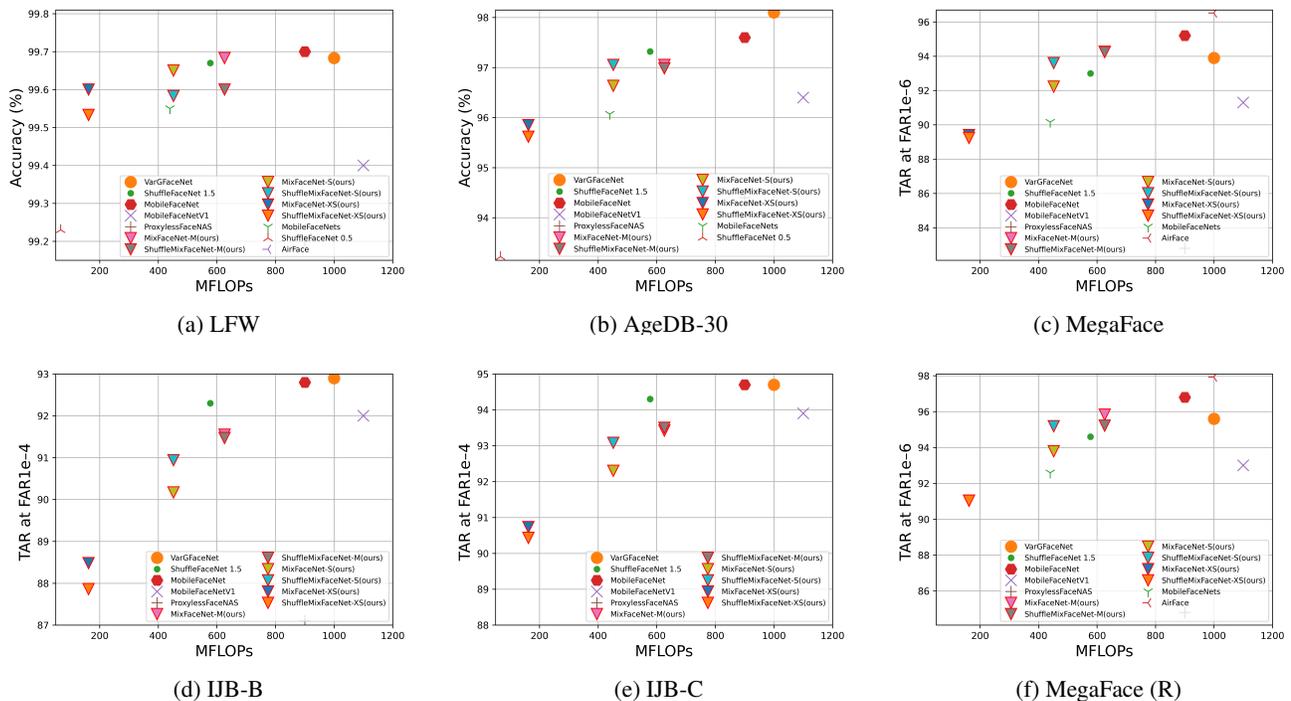


Figure 2: FLOPs vs. performance on LFW (accuracy), AgeDB-30 (accuracy), MegaFace (TAR at FAR1e-6), IJB-B (TAR at FAR1e-4), IJB-C (TAR at FAR1e-4) and refined version of MegaFace, noted as MegaFace (R), (TAR at FAR1e-6). Our MixFaceNet models are highlighted with triangle marker and red edge color.

IJB-C and MegaFace(R) benchmarks are presented in Figures 2 (a), (b), (c), (d), (e) and (f). Each of the reported solutions is presented by an indicator on the plot, where an ideal solution will tend to be placed on the top left corner (high performance and low complexity). In most ranges of the number of FLOPs, and on the 6 benchmarks, different versions of our MixFaceNets achieved the highest verification performance. Similar conclusions can be made by analyzing the presented values in Table 2.

6. Conclusion

We presented in this paper accurate and extremely efficient face recognition models, MixFaceNets. We conducted extensive experiments on popular, publicly available, datasets including LFW, AgeDB-30, MegaFace, IJB-B, and IJB-C. The overall evaluation results demonstrate the effectiveness of our proposed MixFaceNets for applications associated with low computational complexity requirements. Our MixFaceNet-S and ShuffleMixFaceNet-S outperformed MobileFaceNets [4] under the same level of computation complexity ($\leq 500M$ FLOPs). Also, our MixFaceNet-M is showed to be one of the top-ranked performing models, while using significantly fewer FLOPs than the SOTA models.

Acknowledgment: This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

- [1] X. An, X. Zhu, Y. Xiao, L. Wu, M. Zhang, Y. Gao, B. Qin, D. Zhang, and F. Ying. Partial fc: Training 10 million identities on a single machine. In *Arxiv 2010.05222*, 2020.
- [2] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Unmasking face embeddings by self-restrained triplet loss for accurate masked face recognition. *CoRR*, abs/2103.01716, 2021.
- [3] H. Cai, L. Zhu, and S. Han. Proxyllessnas: Direct neural architecture search on target task and hardware. In *7th ICLR, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [4] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings*, volume 10996 of *Lecture Notes in Computer Science*, pages 428–438. Springer, 2018.
- [5] N. Damer, F. Boutros, M. Süßmilch, F. Kirchbuchner, and A. Kuijper. Extended evaluation of the effect of real and simulated masks on face recognition performance. *IET Biometrics*, n/a(n/a).

- [6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [7] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *2019 IEEE/CVF ICCV, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 2638–2646. IEEE, 2019.
- [8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV 2016, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, Lecture Notes in Computer Science, pages 87–102. Springer, 2016.
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE ICCV, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034. IEEE Computer Society, 2015.
- [10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.
- [11] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE CVPR, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7132–7141. IEEE Computer Society, 2018.
- [12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, Uni. of Massachusetts, Amherst, October 2007.
- [13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015.
- [14] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *2016 IEEE CVPR, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4873–4882. IEEE Computer Society, 2016.
- [15] X. Li, F. Wang, Q. Hu, and C. Leng. Airface: Lightweight and efficient model for face recognition. In *2019 IEEE/CVF ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 2678–2682.
- [16] N. Ma, X. Zhang, H. Zheng, and J. Sun. Shufflenet V2: practical guidelines for efficient CNN architecture design. In *ECCV 2018, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 122–138. Springer, 2018.
- [17] A. Mansfield. Information technology–biometric performance testing and reporting—part 1: Principles and framework. *ISO/IEC*, pages 19795–1, 2006.
- [18] Y. Martínez-Díaz, L. S. Luevano, H. M. Vazquez, M. Nicolás-Díaz, L. Chang, and M. González-Mendoza. Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In *2019 IEEE/CVF ICCV, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 2721–2728. IEEE, 2019.
- [19] Y. Martínez-Díaz, M. Nicolás-Díaz, H. Méndez-Vázquez, L. S. Luevano, L. Chang, M. Gonzalez-Mendoza, and L. E. Sucar. Benchmarking lightweight face architectures on specific face recognition scenarios. *Artificial Intelligence Review*, pages 1–44, 2021.
- [20] B. Maze, J. C. Adams, J. A. Duncan, N. D. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. IARPA janus benchmark - C: face dataset and protocol. In *2018 ICB, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pages 158–165. IEEE, 2018.
- [21] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotzia, and S. Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE CVPRW, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1997–2005. IEEE Computer Society, 2017.
- [22] H. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE ICIP, ICIP 2014, Paris, France, October 27-30, 2014*, pages 343–347. IEEE, 2014.
- [23] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. In *6th ICLR, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018.
- [24] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4510–4520. IEEE Computer Society, 2018.
- [25] M. Tan and Q. V. Le. Mixconv: Mixed depthwise convolutional kernels. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 74. BMVA Press, 2019.
- [26] C. Whitelam, E. Taborisky, A. Blanton, B. Maze, J. C. Adams, T. Miller, N. D. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. IARPA janus benchmark-b face dataset. In *2017 IEEE CVPRW, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 592–600. IEEE Computer Society, 2017.
- [27] B. Wu, A. Wan, X. Yue, P. H. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer. Shift: A zero flop, zero parameter alternative to spatial convolutions. In *2018 IEEE CVPR, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9127–9135, 2018.
- [28] X. Wu, R. He, Z. Sun, and T. Tan. A light CNN for deep face representation with noisy labels. *IEEE Trans. Inf. Forensics Secur.*, 13(11):2884–2896, 2018.
- [29] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *2019 IEEE/CVF ICCV, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 2647–2654.
- [30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE SPL*, 23(10):1499–1503, 2016.
- [31] Q. Zhang, J. Li, M. Yao, L. Song, H. Zhou, Z. Li, W. Meng, X. Zhang, and G. Wang. Vargnet: Variable group convolutional neural network for efficient embedded computing. *CoRR*, abs/1907.05653, 2019.