Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19521-0

Towards efficient and robust face recognition through attention-integrated multi-level CNN

Aly Khalifa[1] · Ahmed A. Abdelrahman[1] · Thorsten Hempel[1] · Ayoub Al-Hamadi[1]

Received: 19 October 2023 / Revised: 24 March 2024 / Accepted: 26 May 2024

Abstract
The rapid advancement of deep Convolutional Neural Networks (CNNs) has led to remarkable progress in computer vision, contributing to the development of numerous face verification architectures. However, the inherent complexity of these architectures, often characterized by millions of parameters and substantial computational demands, presents significant challenges for deployment on resource-constrained devices. To address these challenges, we introduce RobFaceNet, a robust and efficient CNN designed explicitly for face recognition (FR). The proposed RobFaceNet optimizes accuracy while preserving computational efficiency, a balance achieved by incorporating multiple features and attention mechanisms. These features include both low-level and high-level attributes extracted from input face images and aggregated from multiple levels. Additionally, the model incorporates a newly developed bottleneck that integrates both channel and spatial attention mechanisms. The combination of multiple features and attention mechanisms enables the network to capture more significant facial features from the images, thereby enhancing its robustness and the quality of facial feature extraction. Experimental results across state-of-the-art FR datasets demonstrate that our RobFaceNet achieves higher recognition performance. For instance, RobFaceNet achieves 95.95% and 92.23% on the CA-LFW and CP-LFW datasets, respectively, compared to 95.45% and 92.08% for very deep ArcFace model. Meanwhile, RobFaceNet exhibits a more lightweight model complexity. In terms of computation cost, RobFaceNet has 337M Floating Point Operations Per Second (FLOPs) compared to ArcFace's 24211M, with only 3% of the parameters. Consequently, RobFaceNet is well-suited for deployment across various platforms, including robots, embedded systems, and mobile devices.

✉ Aly Khalifa
aly.khalifa@ovgu.de

Ahmed A. Abdelrahman
ahmed.abdelrahman@ovgu.de

Thorsten Hempel
thorsten.hempel@ovgu.de

Ayoub Al-Hamadi
ayoub.al-hamadi@ovgu.de

[1] Neuro-Information Technology, Otto-von-Guericke-University Magdeburg, 39106 Magdeburg, Germany

Published online: 05 June 2024

Keywords Face recognition · Deep metric learning · Attention module · Computer vision · CNN

1 Introduction

Convolutional Neural Networks (CNNs) have proven to be highly effective learning algorithms, particularly in discerning and capturing meaningful visual features [1, 13]. They exhibit remarkable performance across various visual recognition tasks and have brought about a revolution in face recognition (FR), significantly improving accuracy and opening new horizons in the field [12, 42].

Simultaneously, the domain of face recognition continues to evolve dynamically with ongoing research and development. Its applications span diverse sectors, encompassing security [30], video surveillance [4], and human-robot interaction [29, 44].

Moreover, the integration of face recognition technology into mobile devices has introduced fresh possibilities, enabling features like facial unlocking and enhancing overall user experiences [37]. However, it's crucial to acknowledge that current state-of-the-art (SOTA) face recognition methods often rely on deep CNNs [10, 14, 42], demanding substantial computational resources. Consequently, deploying face recognition in real-time applications or resource-limited systems such as mobile devices, self-driving cars, and robotics presents significant challenges.

Recent research efforts have been directed towards designing compact and efficient neural networks without compromising performance. Various approaches have been explored to achieve this objective, including network compression and acceleration techniques [33, 59]. These techniques involve methods such as network pruning [20, 35], mimic networks [48], knowledge distillation [16, 57], quantization [17], and depth-wise convolution [23, 24, 41].

However, the progress made in developing lightweight neural networks has not been comprehensively evaluated in the context of face recognition, as opposed to image classification and object detection tasks. Notably, only a limited number of studies have presented accurate lightweight architectures tailored specifically for face recognition purposes [15, 32, 37, 51, 52].

Furthermore, all of these previous works bypass low-level features with local details, favoring high-level features from the final convolution layer due to their larger receptive fields and semantic richness. Unfortunately, this inclination sacrifices local details, hindering the utilization of low-level information. This strategy, however, comes at the cost of lacking fine-grained, local information.

In this paper, we present RobFaceNet, an efficient and high-performance neural network designed for face recognition tasks on resource-limited systems. RobFaceNet has been crafted to leverage both low-level and high-level features, effectively harnessing the inherent diversity and essential information present across different facial regions. To ensure the network's lightweight nature, we employ the bottleneck residual block from MobileNetV2 [41] as our foundational building block. Furthermore, we enhance the network's discriminative capability by integrating an attention block within the bottleneck.

The primary contributions of our work can be summarized as follows:

- Multi-feature approach: We devise an efficient and accurate lightweight face recognition architecture that adopts a multi-feature approach. This strategy enables the extraction of comprehensive feature information, enhancing the network's face recognition capabilities while ensuring the feasibility of real-time processing.

- Enhanced bottleneck with attention: Our approach involves the incorporation of various attention blocks (Channel Attention and Squeeze-and-Excitation) within the bottleneck, tailored to specific layers. This augmentation significantly improves the discriminative power of RobFaceNet.
- Nonlinearity activation function: In RobFaceNet, we employ the h-swish function as the nonlinearity activation function, replacing Parametric Rectified Linear Unit (PReLU). This substitution significantly enhances the overall performance of the model with reduced computational cost.
- Comprehensive experimental evaluation: We evaluate our method on a series of popular face recognition benchmarks and demonstrate that our proposed model consistently outperforms other SOTA counterparts, even when compared to other models with similar or larger parameter sizes.

The structure of the paper is organized as follows. In Section 2, we review existing lightweight CNNs tailored for face recognition. Section 3 introduces the lightweight RobFaceNet architecture proposed for face recognition tasks. The outcomes of our experimental results and an in-depth ablation study are detailed in Section 4, followed by discussions and future work in Section 5. Finally, we conclude our work in Section 6.

2 Related work

Face recognition has gained immense popularity, especially in mobile device applications. This surge has amplified the need to strike a balance between model accuracy and computational cost in deep neural networks. While several advancements have been made to enhance face recognition systems' efficiency on mobile platforms, the challenge remains.

Various lightweight network architectures have been proposed for common visual tasks, including SqueezeNet [27], MobileNets [24], MobileNetV2 [41], ShuffleNet [60], MobileNetV3 [23], and MobileOne [46]. Moreover, efficient lightweight network architectures (Table 1) have adapted these networks for face recognition by designing compact convolution building blocks, such as SqueezerFaceNet [3], MobileFaceNets [9], AirFace [32], VarGFaceNet [52], ShuffleFaceNet [37], Mixfacenets [5], and PocketNet [7].

These models draw inspiration from the advancements in deep image classification models and the evolution of depthwise separable convolutions [41, 45, 58, 60]. These models address the specific challenge of the high number of parameters in fully connected (FC) layers. To overcome this issue, these efficient FR models replace FC layers with global depthwise convolutions (GDC). The GDC layer weights different units of the feature map differently,

Table 1 Proposed lightweight models in the literature for face recognition

Network	Vector Size	Base Architecture
MobileFaceNets [9]	256	MobileNetv2
AirFace [32]	512	MobileFaceNets
VarGFaceNet [52]	512	VarGNet
ShuffleFaceNet [37]	128	ShuffleNet
Mixfacenets [5]	512	MixNets
PocketNet [7]	128-256	PocketNet
RobFaceNet(Ours)	512	Mobile Networks

providing a more effective architecture for face recognition tasks. The GDC layer has a computational cost of only $W \times H \times C$, where W, H, and C represent the width, height, and channels of the input feature map. Moreover, this approach effectively reduces the parameter count while maintaining or even improving performance.

For instance, the MobileFaceNets architecture [9] is built upon the residual bottlenecks introduced by MobileNetV2 [41], incorporating approximately 1M parameters with 439M FLOPs. The authors finetuned the MobileNetV2 architecture by incorporating a GDC layer instead of a global average pooling (GAP) layer. Additionally, they opted to use the PReLU [21] function as the nonlinearity in all convolutional layers. These design choices have yielded improved performance in facial recognition tasks. While MobileFaceNets demonstrate improved performance across various datasets, they encounter limitations when applied to the MegaFace dataset, where accuracy experiences a slight decrease.

ShuffleFaceNet [37] proposed a compact FR model by finetuning ShuffleNetV2 [36]. This approach replaced the last GAP layer with a GDC layer and the Rectified Linear Unit (ReLU) activation function with PReLU. ShuffleFaceNet is slightly larger than MobileFaceNet; however, it offers better accuracy.

Following a similar pattern as in [9] and [37], the VarGFaceNet [58] and Mixfacenets [5] model architectures adopted the VarGNet [58] and MixNets [45], respectively. In [58], the authors introduced modifications to the VarGNet block, including incorporating a squeeze and excitation block (SE), replacing ReLU with PReLU, and introducing variable group convolutions before the FC layer. These modifications significantly reduced the number of parameters to 5M and the computational cost to 1G FLOPs. Recursive knowledge distillation was also employed to enhance the model's generalization capability. VarGFaceNet achieved an impressive accuracy of 99.85% on the Labeled Faces in the Wild dataset (LFW) with around 5M parameters and 1022M FLOPs. However, the computational cost of VarGFaceNet is still higher compared to ShuffelFaceNet and MobileFaceNet. Similarly, in [5], the authors introduced a family of efficient face recognition models (MixFaceNets) by incorporating the MixConv block [45] with a channel shuffle operation, which enhances the discriminative ability of the model. With 3.95M parameters and 626M FLOPs, the MixFaceNets model achieved an accuracy of 99.68% on the LFW dataset.

Inspired by the successes of MobileFaceNets [9], Li et al. [32] developed AirFace, a lightweight FR model based on the deeper MobileFaceNet(y2) [10] architecture. In their approach, they increased the network width and depth to further improve the model's performance. Additionally, they incorporated the Convolutional Block Attention Module (CBAM) [50] into every bottleneck within the network. With 4.23M parameters and 1000M FLOPs, the AirFace model achieved an accuracy of 99.27% on the LFW dataset. However, the model still expensive in terms of computational complexity.

In [7], Boutros et al. introduced a family of lightweight FR models called PocketNets. They utilized Neural Architecture Search (NAS) techniques to automatically discover an FR-specific lightweight architecture, optimizing it for performance and computational efficiency. In addition to the architectural design, Boutros et al. proposed a novel KD paradigm to address the challenges arising from the significant performance gap between the teacher and student models. Despite having only 0.92M parameters and 587M FLOPs, the PocketNets model achieved an accuracy of 99.58% on the LFW dataset.

Among the previously mentioned works, MobileFaceNets [9] and MixFaceNets [5] have particularly shined in achieving impressive accuracy with minimal computational costs. However, our proposed RobFaceNet architecture surpasses these by delivering superior results with even fewer computational demands. It stands as a potential solution of optimizing both accuracy and efficiency in face recognition tasks.

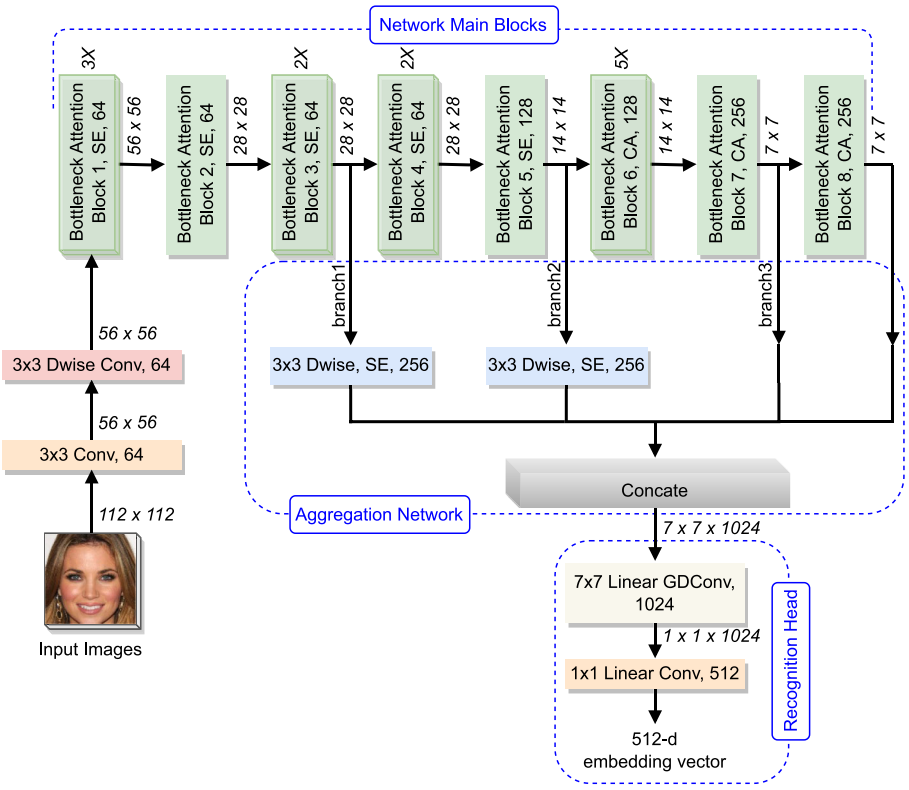


Fig. 1 Architecture of the proposed network. The RobFaceNet architecture incorporates multi-feature networks that consider both low-level and high-level features in the embedding process. This information is extracted from the middle blocks of the network

3 Proposed approach

The proposed RobFaceNet architecture, inspired by MobileFaceNets and MobileNetV2, is specifically designed for face recognition. Unlike conventional adaptations, RobFaceNet integrates novel enhancements such as attention-based bottlenecks and a multi-feature approach. Visually represented in Fig. 1 and detailed in Table 2, the RobFaceNet architecture comprises several convolutional layers. These layers incorporate innovative connections, merging low and high-level features. Such integration facilitates the extraction of a broader spectrum of informative features, significantly boosting the network's capability to capture and represent facial characteristics.

3.1 Enhanced bottleneck

In the context of using CNNs for facial feature extraction, giving the most recognizable face regions more weight is important. Similarly, the feature channels with the most distinguishing information should be assigned more weight [18]. To achieve superior performance, we intuitively combine them by introducing an attention-based enhanced bottleneck into RobFaceNet.

Table 2 The proposed network architecture

Input	Operator	n	s	c	Attention
$112^2 \times 3$	conv3 × 3	1	2	64	No
$56^2 \times 64$	depthwise conv3 × 3	1	1	64	No
$56^2 \times 64$	bottleneck	3	1	64	CA
$56^2 \times 64$	bottleneck	1	2	64	CA
$28^2 \times 64$	bottleneck	2	1	64	CA
$28^2 \times 64$	depthwise branch1	1	4	256	CA
$28^2 \times 64$	bottleneck	2	1	64	CA
$28^2 \times 64$	bottleneck	1	2	128	CA
$14^2 \times 128$	depthwise branch2	1	2	256	SE
$14^2 \times 128$	bottleneck	5	1	128	SE
$14^2 \times 128$	bottleneck	1	2	256	SE
$7^2 \times 256$	bottleneck	1	1	256	SE
$7^2 \times 1024$	linear GDConv7 × 7	1	1	1024	No
$1^2 \times 1024$	linear conv1 × 1	1	1	512	No

Each line describes a sequence of operators, repeated n times with stride s. All layers in the same sequence have the same number c of output channels

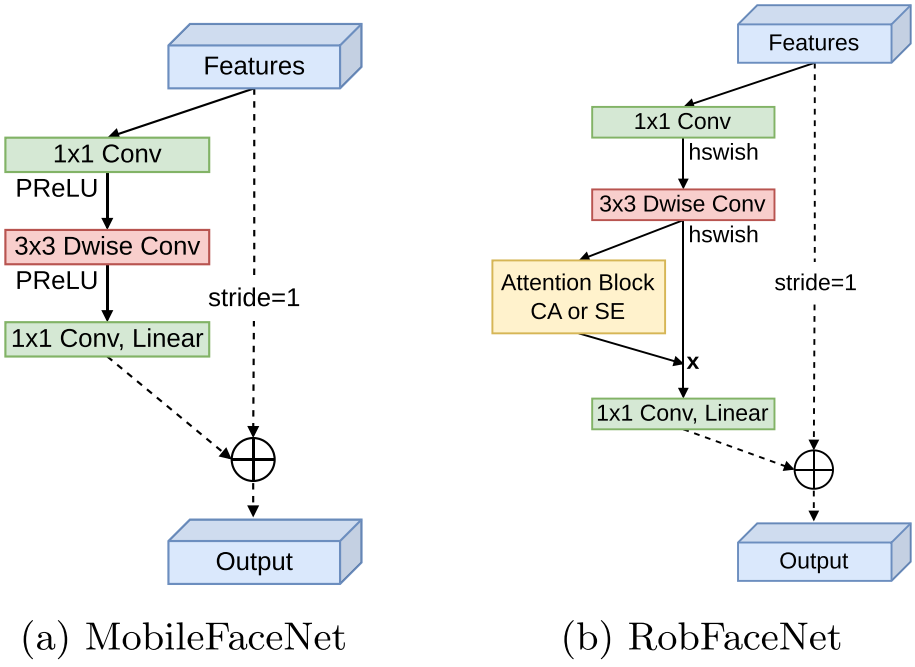


Fig. 2 Comparison of the bottlenecks used in (a) MobileFaceNet and (b) RobFaceNet. MobileFaceNet incorporates MobileNetV2 [41] bottlenecks, replacing ReLU with PReLU. RobFaceNet utilizes MobileNetV2 bottlenecks along with attention mechanisms, such as the Squeeze-and-Excite [25] block or coordinate attention [22]. Unlike [9, 41], we apply the squeeze and excite block or coordinate attention in the residual layer and use hswish as the nonlinearity. The dashed blocks are only applied when the stride is equal to 1

The enhanced attention-based bottleneck is an inverted bottleneck incorporating either a CA [22] or SE [25] attention module. As depicted in Fig. 2(b), we plug the attention blocks into the inverted residual block after the depthwise convolution layer, and the architecture of both SE and CA attention module is illustrated in Fig. 3. This fusion of attention modules improves the network's ability to interpret both channel and spatial features, thereby boosting its ability to discriminate between different facial characteristics.

The choice of attention module varies depending on the specific layer and is outlined in Table 2. In particular, we use the CA module in the initial layers to capture dependencies and correlations between different positions within a feature map. This helps the network effectively distinguish diverse facial attributes and structures. Conversely, we employ the SE module in the latter layers to capture channel-wise interactions and select the most suitable representation through channel weight recalibration.

This strategic deployment of attention modules across varying network layers enhances the model's overall performance and enables the network to focus on crucial features at different stages, thereby amplifying its capability for face recognition tasks. Additionally, by incorporating these attention mechanisms, RobFaceNet becomes more proficient at recognizing facial characteristics and achieves higher recognition accuracy without compromising computational efficiency.

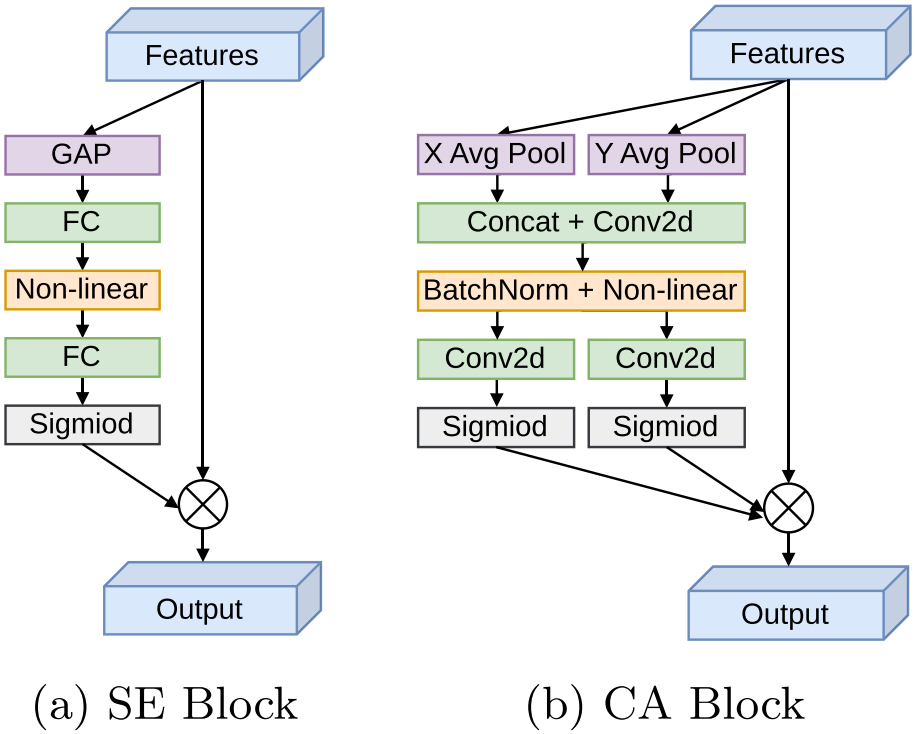


Fig. 3 Various attention mechanisms used in the proposed approach

3.2 RobFaceNet

Multi-feature CNN: Many existing methods rely solely on high-level features extracted from the last convolutional layer, overlooking the fact that representations from individual layers may lack comprehensiveness [55]. To address this gap, we introduce a novel multi-feature approach within our RobFaceNet.

Our approach precisely aggregates features from various network layers, extracting a comprehensive and informative feature pool. Consequently, RobFaceNet adeptly extracts, processes, and leverages a broad, insightful feature scope that is especially relevant for face recognition tasks, where recognizing and understanding fine-grained features and subtle variations are imperative for achieving high recognition precision.

Specifically, as illustrated in Fig. 1(aggregation network), we merge the outputs from the $block3$, $block5$, and $block7$ layers in RobFaceNet through separate network branches. To align the dimensions of the feature maps, we introduce $branch1$ and $branch2$ depthwise convolutional layers. Subsequently, the outputs of these layers are concatenated with the output of $block8$, resulting in feature maps of dimensions $7 \times 7 \times 1024$ that are fed into the following global depthwise convolution layer, $GDCConv7x7$.

By integrating both low-level and mid-level features besides the high-level features, our model becomes proficiently equipped for accurate and effective facial characteristic recognition. Moreover, our multi-feature approach adeptly balances between the richness of information and computational efficiency, ensuring the model's practical applicability across a spectrum of face recognition scenarios, thereby enhancing its versatility and overall effectiveness as a promising solution for various applications.

Nonlinearities: Traditional mobile and lightweight networks [24, 27, 41, 60] commonly employ the ReLU activation function [54] as their nonlinearity activation function. However, the ReLU function restricts activations to non-negative values, posing limitations, particularly when complex feature representations are needed. In the specialized context of lightweight face recognition networks [3, 5, 7, 9, 32, 37, 52], PReLU [21] is often favored over ReLU. This is because PReLU permits negative activation values, which has been shown to improve performance in face recognition tasks.

Unlike in related works, we chose to use the modified h-swish activation function, which was introduced in MobileNetV3 [23]. We made this choice for two main reasons. First, the h-swish activation function significantly reduces computational costs while still maintaining competitive performance in our RobFaceNet for face recognition tasks. Second, h-swish can be efficiently implemented as a piece-wise function, thereby minimizing memory access and substantially reducing latency costs [23].

By adopting h-swish, we aim to enhance both the efficiency and effectiveness of our network, making it ideally suited for face recognition tasks in resource-limited environments and embedded systems. The modified h-swish function is defined as:

$$H\text{-}Swish(x) = x\,\frac{ReLU6(x+3)}{6}, \tag{1}$$

where, $ReLU6(x) = min(max(x, 0), 6)$.

Embedding setting: In traditional lightweight networks, such as MobileNetV2, the GAP layer is often used to obtain an embedding vector. However, for face recognition tasks, this technique has been found to be sub-optimal [9, 10, 37, 51]. The central drawback of the GAP layer is its equal treatment of each unit in the output feature map, ignoring their varying discriminative power in face feature extraction.

In face recognition, different units in the feature map correspond to unique facial features, each with varying contributions to discriminative capability. A refined approach is needed to weight these units differently when forming the feature vector. Traditionally, replacing the GAP layer with a FC layer has been considered as a way to address this limitation, as it allows the network to learn specific weights for each unit. However, this approach significantly increases computational overhead and model size due to the added weights.

To mitigate these challenges, we follow an approach similar to that in MobileFaceNet [9]. Specifically, we replace MobileNetV2's GAP layer with a GDC layer in our proposed RobFaceNet. This GDC layer weights different units of the feature map differently, providing a more effective architecture for face recognition tasks. The incorporation of the GDC layer, together with strategic refinements across our model architecture, formulates a model proficiently balanced between computational efficiency and perceptive feature extraction, yielding a network finely tuned for face recognition tasks, especially within resource-constrained environments.

The GDC layer processes feature maps of dimensions $7 \times 7 \times 1024$ into $1 \times 1 \times 1024$ feature maps. Subsequently, a linear convolution of size $1 \times 1 \times 512$ is applied to generate the final 512-dimensional embedding vector.

4 Experiments and analysis

In this section, the experimental settings are presented first. After that, we visually demonstrate the efficiency of our RobFaceNet, then the experimental results for SOTA lightweight face recognition models are reported. Finally, we conduct an ablation study to analyze the impacts of various settings in RobFaceNet, focusing on how they influence accuracy and computational efficiency.

4.1 Experimental settings

Training datasets: Our RobFaceNet model was trained using the MS1MV2 dataset [10], and for the ablation study, we utilized the VGGFace2 dataset [8]. The MS1MV2 dataset is

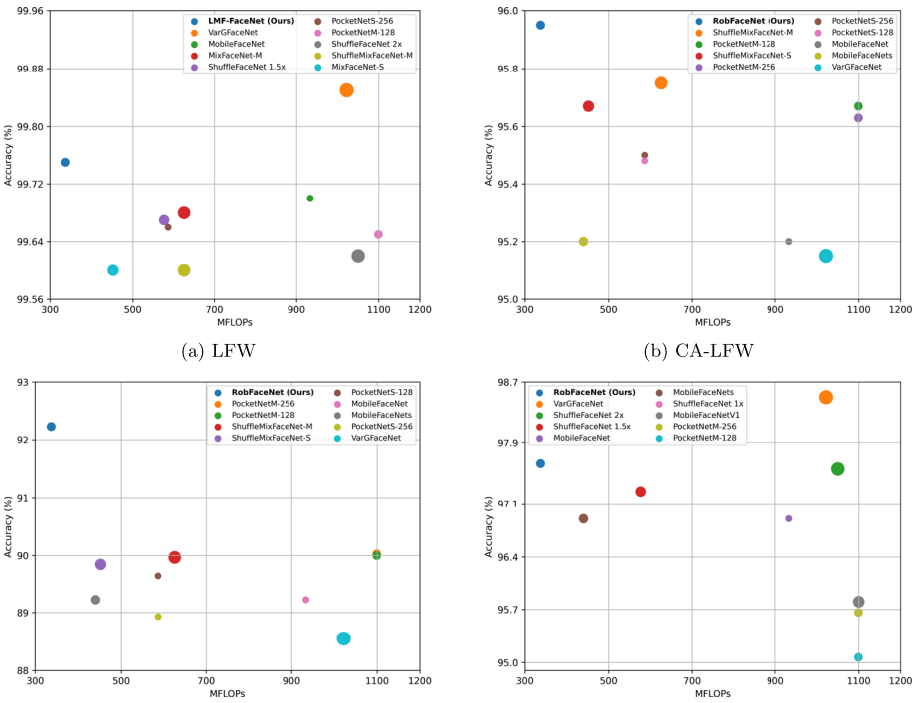
Table 3 Face datasets for training and testing

Input	# Identity	# Image/Videos	Task	Key features
MS1MV2 [10]	85K	5.8M/-	train	Unconstrained images
MS1MV3 [11]	91K	5.1M/-	train	Unconstrained images
LFW [26]	5,749	13,233/-	1:1	Unconstrained images
CFP-FP [43]	500	2,000/-	1:1	Cross-pose
AgeDB-30 [40]	568	16,488/-	1:1	Cross-age
CP-LFW [61]	3,968	11,652/-	1:1	Cross-pose
CA-LFW [62]	4,025	12,174/-	1:1	Cross-age
IJB-B [49]	1,845	21.8K/7,011	1:1	Large-scale,
			1:N	Full pose variation
IJB-C [39]	3,531	31.3K/11,779	1:1	Large-scale,
			1:N	Full pose variation

an improved version of MS-Celeb-1M [19], with around 5.8M images belonging to approximately 85k identities. On the other hand, VGGFace2 comprises 3.14M face images that cover a wide range of poses, ages, and ethnicities.

Validation and test datasets: We used the LFW [26], CFP-FP [43], and AgeDB-30 [40] datasets for validation purposes to assess the improvements achieved with different settings. Additionally, we used different benchmarks to evaluate the effectiveness of our proposed lightweight face model in various face recognition tasks, highlighting their main characteristics in Table 3. In addition to efficient face verification datasets like LFW, we also evaluated the performance of our lightweight networks on larger-scale image datasets, such as IJB-B [49] and IJB-C [39]. Furthermore, we extensively tested our models on cross-pose datasets, including CFP-FP [43] and CP-LFW [61], as well as cross-age datasets, such as AgeDB-30 [40] and CA-LFW [62]. These evaluations demonstrate the robustness and effectiveness of our lightweight face recognition RobFaceNet in various challenging scenarios.

Data preprocessing: For data preprocessing, we follow the commonly used approach as in recent works [6, 10, 63]: each face image is cropped to a size of 112×112, using a similarity transformation based on the five face landmarks (two eyes, a nose, and two mouth corners) detected by Multi-task Cascaded CNN (MTCNN) [56]. Finally, the RGB pixel values are normalized from $[0, 255]$ to $[-1, 1]$. The RobFaceNet model processes aligned



(a) LFW (b) CA-LFW

(c) CP-LFW (d) CFP-FP

Fig. 4 Computational complexity vs. performance vs. model size on (a) LFW, (b) CA-LFW, (c) CP-LFW, and (d) CFP-FP. The area of each model is proportional to the model size. Our RobFaceNet is marked with a blue circle and is placed repeatedly in the top left corner, proving a SOTA trade-off between FR performance, FLOPs, and compactness. Note that we visualize the top ten compact models that have performed best in recent literature on each benchmark

and cropped face images with dimensions of $112 \times 112 \times 3$ to produce 512-dimensional feature embeddings.

Training setup: The models introduced in this paper are implemented using PyTorch. For a fair performance comparison with other SOTA models, all models are trained using the ArcFace loss [10] with an angular margin of $m = 0.5$ and a feature scale of $s = 64$. During training, we set the batch size to 512 and utilized an NVIDIA Quadro RTX 8000 GPU. The optimization is performed using the Stochastic Gradient Descent (SGD) optimizer [28] with an initial learning rate of 1e-1, momentum of 0.9, and weight decay parameter of 5e-4. The learning rate is reduced by a factor of ten at 80k, 140k, 210k, and 280k training iterations. To monitor the model's performance during training, we evaluate it on LFW, CFP-FP, and AgeDB datasets after every 5000 training iterations. The training process is stopped after 300k iterations. For verification, we use the cosine distance between feature vectors in all experiments.

4.2 Performance vs. computational complexity

Figure 4 visually demonstrates the efficiency of our RobFaceNet by comparing the number of FLOPs with the verification performance achieved in Table 4. To offer a more comprehensive view, we have added the model size as an additional metric, visually represented by the marker size, which is directly proportional to the model size.

For a robust evaluation, we included the top ten compact models that have demonstrated exceptional performance in recent literature across each benchmark. Each model is represented by a colored circle, and its position on the graph correlates with its performance and computational complexity. The ideal model would occupy the upper-left corner, indicating high performance at a low computational cost.

Remarkably, as Fig. 4 illustrates, our RobFaceNet model consistently occupies this optimal position, highlighting an advantageous balance between model complexity and face recognition performance. This strong result emphasizes that RobFaceNet not only achieves high accuracy but also minimizes computational requirements, making it a highly efficient and practical solution for face recognition, particularly in resource-limited settings such as embedded and robotic systems.

4.3 Experimental results

In this section, we present the results obtained by RobFaceNet across different benchmarks. To ensure consistency and reproducibility, we follow the standard approach and adhere to the evaluation metrics employed in these benchmarks and also in previous works reporting on them.

Additionally, to ensure a fair comparison with SOTA models, RobFaceNet and all the models included in the comparison are trained on the MS1MV2 dataset and utilize the ArcFace loss function. This standardization of training datasets and loss functions allows for a direct and unbiased comparison of RobFaceNet against other SOTA models on these challenging benchmarks.

4.3.1 Performance comparison on public datasets

Table 4 presents the face recognition results achieved by our RobFaceNet on all evaluation benchmarks. It also includes a comparison between RobFaceNet and recent compact models

Table 4 RobFaceNet verification accuracies on 5 benchmarks and TAR at FAR 1e-4 on IJB-B and IJB-C

Model	#FLOPs (M)	#Params. (M)	Size (MB)	LFW (%)	Cross-Age		Cross-Pose		IJB	
					CA-LFW (%)	AgeDB-30 (%)	CP-LFW (%)	CFP-FP (%)	IJB-B (%)	IJB-C (%)
ArcFace (ResNet100) [10]	24211	65.2	261.22	99.82	95.45	98.15	92.08	98.40	94.20	95.60
MobileFaceNetV1 [38]	1100	3.40	13.1	99.40	94.47	96.40	87.17	95.80	92.00	93.90
PocketNetM-256 [7]	1099.15	1.75	7.0	99.58	95.63	97.17	**90.03**	95.66	90.74	92.70
PocketNetM-128 [7]	1099.02	1.68	6.7	99.65	**95.67**	96.78	90.00	95.07	90.63	92.63
ShuffleFaceNet 2x [37]	1050	4.5	18.0	99.62	-	97.28	-	97.56	-	-
VarGFaceNet [52]	1022	5	20.0	**99.85**	95.15	**98.15**	88.55	**98.50**	**92.90**	**94.70**
AirFace [32]	1000	4.23	-	99.27	-	93.25	-	94.11	-	-
MobileFaceNet [38]	933.30	2	4.0	**99.70**	95.20	**97.60**	89.22	96.90	**92.80**	**94.70**
ProxylessFaceNAS [38]	900	3.20	12.5	99.20	92.55	94.40	84.17	94.70	87.10	89.70
MixFaceNet-M [5]	626.1	3.95	15.8	99.68	-	97.05	-	-	91.55	93.42
ShuffleMixFaceNet-M [5]	626.10	3.95	15.8	99.60	**95.75**	96.98	**89.97**	94.96	91.47	93.5
PocketNetS-256 [7]	587.22	0.99	3.9	99.66	95.50	96.35	88.93	93.34	89.31	91.33
PocketNetS-128 [7]	587.11	0.92	3.7	99.58	95.48	96.10	89.63	94.21	89.44	91.62

Table 4 continued

Model	#FLOPs (M)	#Params. (M)	Size (MB)	LFW (%)	Cross-Age		Cross-Pose		IJB	
					CA-LFW (%)	AgeDB-30 (%)	CP-LFW (%)	CFP-FP (%)	IJB-B (%)	IJB-C (%)
ShuffleFaceNet 1.5x [37]	577.5	2.60	10.5	99.67	95.05	97.32	88.5	**97.26**	92.30	94.30
MixFaceNet-S [5]	451.7	3.07	12.28	99.60	-	96.63	-	-	90.17	92.30
ShuffleMixFaceNet-S [5]	451.7	3.07	12.28	99.58	95.67	97.05	89.85	94.10	90.94	93.08
MobileFaceNets [9]	439.8	0.99	8.2	99.55	95.20	96.07	89.22	96.90	-	-
ShuffleFaceNet 1x [37]	275.8	1.40	5.6	99.45	-	96.33	-	96.04	-	-
MixFaceNet-XS [5]	161.9	1.04	4.2	99.60	-	95.85	-	-	88.48	90.73
ShuffleMixFaceNet-XS [5]	161.9	1.04	4.2	99.53	94.93	95.61	86.93	91.25	87.86	90.43
GhostFaceNetV2-2 [2]	76.51	6.84	13.66	99.71	95.70	96.55	89.58	93.07	91.76	93.03
MobileFaceFormer [31]	-	0.98	-	99.41	95.50	96.75	96.75	94.29	-	-
RobFaceNet (Ours)	337.3	1.90	7.27	**99.75**	**95.95**	**97.42**	**92.23**	**97.63**	**92.08**	**93.86**

The table's first row shows the result achieved by the current SOTA ReNet100 models. The models are ordered based on the number of FLOPs. Results and the number of decimal points is reported as in the respective works. Our RobFaceNet consistently surpasses the SOTA performance on all evaluation benchmarks, demonstrating its superiority and effectiveness in face recognition tasks. Results are in % and higher values are better. The best performance in each category on each benchmark is in bold

proposed in the literature. The listed models are grouped based on their complexity measured in MFLOPs. The results are reported as in the related works. To provide a context of the current SOTA performance in larger-scale deep face recognition, we first report the results for the current SOTA ReNet100 models. The rest of the table is organized into three parts. The first part showcases the results for models with complexity above 1000M FLOPs, while the second and third parts present the results for models with less than 1000M and less than 500M FLOPs, respectively. This organization allows for a comprehensive comparison between RobFaceNet and models of varying complexities, highlighting RobFaceNet's efficiency and performance in the face recognition domain.

The top-performing reported result on the LFW benchmark (99.85% accuracy) is attributed to VarGFaceNet, which comes at a computational cost of 1022M FLOPs. In comparison, our RobFaceNet achieved a highly competitive result on LFW (99.75% accuracy) while utilizing 67% fewer FLOPs (337M). Similarly, similar results have been observed on the AgeDB-30 and CFP-FP benchmarks. Our RobFaceNet achieved remarkably close accuracy to the current SOTA models while adopting a more efficient model architecture that employs 62% fewer parameters.

For CA-LFW and CP-LFW, it not only achieves the best accuracy of 95.83% and 99.22%, respectively but also surpasses all SOTA models.

Notably, among all models with computational complexity less than 500M FLOPs, our RobFaceNet surpasses all the listed models, including MobileFaceNets. This underscores the efficiency and effectiveness of our proposed RobFaceNet architecture for face recognition tasks.

4.3.2 Evaluation on IJB-B and IJB-C

We evaluate RobFaceNet on popular large-scale face recognition benchmarks, namely IJB-B and IJB-C. The evaluation results detailed in Table 4 indicate that our RobFaceNet has achieved competitive performance when compared to numerous larger models. For instance, on the IJB-C benchmark, our RobFaceNet model, which has only 1.9M parameters and 337.3M FLOPs, achieves an impressive verification performance of 93.86% TAR at FAR of 1e-4. Notably, the best verification performance on this benchmark was achieved by MobileFaceNet [38] and VarGFaceNet [52], both with larger parameter counts of 2M and 5M, respectively. While MobileFaceNet and VarGFaceNet attain verification performances of 94.7%, they come at the cost of greater computational complexity, with 933.3M and 1022M FLOPs, respectively.

These results demonstrate that our lightweight RobFaceNet can perform on par with or even outperform larger models in face recognition tasks while having significantly fewer parameters. The competitive performance achieved by our model highlights its efficiency and effectiveness in handling real-world face recognition applications.

4.3.3 Evaluation on MegaFace

We also evaluated the performance of RobFaceNet on both the MegaFace dataset and its refined version, MegaFace(R) dataset, using FaceScrub as the probe set. The results presented in Table 5 indicate that RobFaceNet outperforms other methods across both MegaFace and MegaFace(R) datasets. Specifically, RobFaceNet achieves the highest Rank-1 face identification accuracy of 80.2% on MegaFace and the highest verification rate of 96.38% at a 10-6 FAR. On MegaFace(R), RobFaceNet achieves the highest Rank-1 accuracy of 96.25% and a verification rate of 97.17% at a 10-6 FAR.

Table 5 The achieved results on MegaFace and refined MegaFace(R) Challenge1

Model	#FLOPs (M)	#Params. (M)	Size (MB)	MegaFace		MegaFace(R)	
				Rank-1 (%)	Ver. (%)	Rank-1 (%)	Ver. (%)
VarGFaceNet [52]	1022	5	20.0	78.20	93.90	94.90	95.6
MobileFaceNet [38]	933.30	2	4.0	79.30	95.20	95.80	96.8
MixFaceNet-M [5]	626.1	3.95	15.8	78.20	94.26	94.95	95.83
ShuffleMixFaceNet-M [5]	626.10	3.95	15.8	78.13	94.24	94.64	95.22
PocketNetS-128 [7]	587.11	0.92	3.7	76.49	96.10	89.63	94.21
ShuffleMixFaceNet-S [5]	451.7	3.07	12.28	77.41	93.60	94.07	95.19
GhostFaceNetV1-1 [2]	215.66	4.09	8.17	79.32	96.20	95.94	96.9
RobFaceNet (Ours)	337.3	1.90	7.27	**80.20**	**96.38**	**96.25**	**97.17**

Results are in % and higher values are better. *Ver.* refers to the face verification given in TAR at 10-6 FAR. The best performance is in bold

4.4 Ablation study

4.4.1 Impact of attention modules

We investigate the efficacy of introducing attention modules into the inverted bottleneck configuration and present the results of our experimental evaluation in Table 6. We specifically compare the impact of two distinct attention modules: CA and SE. Our experimental methodology involves two phases. First, we apply a single type of attention module either CA or SE uniformly across all layers. Second, we explore a hybrid approach in which CA is used in the initial layers and SE in the later layers, and vice versa, to ascertain the most effective combination.

The obtained results in Table 6 clearly show that our proposed attention-based bottleneck architecture, which integrates CA modules in the initial layers and SE modules in the later layers, delivers superior recognition performance on all datasets. This high level of performance is also maintained across additional datasets featuring variations in age and pose. These results underscore the enhanced feature learning capabilities of the attention-based bottleneck structure and validate its utility in improving face recognition performance across a broad spectrum of scenarios and conditions.

Table 6 Effects of different attention modules

SE	CA	MFLOPs	LFW	AgeDB-30	CA-LFW	CP-LFW	CFP-FP
✗	✗	333.7	99.56	93.98	93.25	91.28	96.94
✓	✗	335.4	99.50	94.11	93.50	91.65	97.42
✗	✓	339.0	99.61	94.08	93.40	92.00	97.51
F	L	336.6	99.55	93.92	93.40	92.13	97.35
L	**F**	**337.3**	**99.65**	**94.53**	**93.66**	**92.33**	**97.79**

F and L denote the application of attention modules to the first and last layers of the network, respectively. All models are trained on VGGFace2 [8]. The last row indicates the RobFaceNet setting
Results are in % and higher values are better. The best performance is in bold

Table 7 Effects of different nonlinearities

ReLU	PReLU	HSwish	MFLOPs	LFW	AgeDB-30	CA-LFW	CP-LFW	CFP-FP
✓	✗	✗	341.4	99.60	94.21	93.40	92.20	97.27
✗	✓	✗	341.4	99.60	94.35	93.61	92.11	97.65
✓	✗	✓	340	99.53	94.13	93.26	92.17	97.20
✗	✓	✓	340	99.50	94.40	93.60	92.10	97.51
✘	✘	✔	**337.3**	**99.65**	**94.53**	**93.66**	**92.33**	**97.79**

All models are trained on VGGFace2 [8]. The last row indicates RobFaceNet settings
Results are in % and higher values are better. The best performance is in bold

4.4.2 Impact of nonlinearities

The effectiveness of nonlinearity selection is demonstrated in Table 7. We investigate various strategies for incorporating h-swish nonlinearities and assess the benefits of using mixed nonlinearities compared to using a single type of nonlinearity. Initially, we adopt the traditional approach of replacing ReLU with PReLU to improve face recognition performance. Subsequently, we experiment with substituting ReLU with h-swish. Finally, we examine the use of mixed nonlinearities, incorporating both h-swish and PReLU or ReLU at different sections of the network.

The results in Table 7 reveal the impact of these nonlinearity choices on face recognition performance. These insights are valuable for optimizing the network's nonlinear activation functions to achieve improved accuracy and efficiency. In our RobFaceNet architecture, we consistently employ the h-swish activation function across all network layers. This strategic decision results in a significant improvement in performance, along with a reduction in computational complexity when compared to alternative configurations. This finding underscores the benefits of employing h-swish nonlinearities, further enhancing the efficiency and effectiveness of our network.

5 Discussions

In this paper, we introduced a novel lightweight, efficient, and robust neural network explicitly tailored for face recognition tasks. Unlike existing lightweight FR networks such as MobileFaceNets [9], VarGFaceNet [52], and PocketNet [7], our proposed model, RobFaceNet, leverages both low-level and high-level features to enable the extraction of diverse and comprehensive feature information. This approach results in more robust and accurate face recognition performance across various conditions, including differences in lighting, poses, and occlusions.

Furthermore, RobFaceNet has been designed with a focus on efficiency without compromising performance. This is achieved through the careful selection of architectural components and attention mechanisms, allowing RobFaceNet to achieve competitive performance while maintaining low computational complexity.

To further demonstrate the effectiveness of the proposed network, we recorded the verification performance results, measured in terms of accuracy, on several popular benchmark datasets. Table 8 compares the performance of RobFaceNet against the SOTA deep FR baseline models and lightweight FR models ($<$ 500M FLOPs). The results indicate that our

Table 8 Comparison of very deep SOTA FR models, SOTA FR models with computation complexity under 500M FLOPs, and our proposed RobFaceNet

Model	#FLOPs (M)	#Params. (M)	Size (MB)	LFW (%)	Cross-Age		Cross-Pose		IJB		MegaFace	
					CA-LFW (%)	AgeDB-30 (%)	CP-LFW (%)	CFP-FP (%)	IJB-B (%)	IJB-C (%)	Rank-1 (%)	Ver. (%)
FaceNet [10]	451.7	3.07	-	99.63	-	-	-	-	-	-	70.49	86.47
SphereFace [34]	24211	65.2	261.22	99.42	90.30	92.88	81.40	-	-	-	72.73	85.56
CosFace [47]	24211	65.2	261.22	99.73	95.76	98.11	92.19	98.12	**94.80**	**96.36**	80.56	96.56
ArcFace [10]	24211	65.2	261.22	**99.82**	95.45	**98.15**	92.08	**98.40**	94.20	95.60	**81.03**	**96.98**
RobFaceNet (Ours)	337.3	1.90	7.27	99.75	**95.95**	97.42	**92.23**	97.63	92.08	93.86	80.20	96.38
MixFaceNet-S [5]	451.7	3.07	12.28	99.60	-	96.63	-	-	90.17	92.30	76.49	92.23
ShuffleMixFaceNet-S [5]	451.7	3.07	12.28	99.58	95.67	97.05	89.85	94.10	90.94	93.08	77.41	93.60
MobileFaceNets [9]	439.8	0.99	8.2	99.55	95.20	96.07	89.22	96.90	-	-	-	90.16
MixFaceNet-XS [5]	161.9	1.04	4.2	99.60	-	95.85	-	-	88.48	90.73	74.18	89.40
ShuffleMixFaceNet-XS [5]	161.9	1.04	4.2	99.53	94.93	95.61	86.93	91.25	87.86	90.43	73.85	89.24
GhostFaceNetV2-2 [2]	76.51	6.84	13.66	99.71	95.70	96.55	89.58	93.07	91.76	93.03	79.31	95.21
RobFaceNet (Ours)	337.3	1.90	7.27	**99.75**	**95.95**	**97.42**	**92.23**	**97.63**	**92.08**	**93.86**	**80.20**	**96.38**

The best performance in each category for each benchmark is highlighted in bold

proposed model outperforms the lightweight FR models in all evaluation datasets, highlighting the significant improvement achieved by incorporating multi-feature and attention mechanisms.

Moreover, our proposed model outperforms the very deep baseline models on two datasets, namely CA-LFW and CP-LFW, and achieves comparable results on the remaining benchmarks. This is accomplished with a more lightweight model complexity. For instance, in terms of computation cost, RobFaceNet has 337M FLOPs compared to ArcFace's 24211M, with only 3% of the parameters.

In addition to its performance benefits, RobFaceNet offers practical advantages by effectively addressing the limitations of the traditional lightweight networks and allowing for more reliable and accurate recognition.

A limitation of this paper is that the study mainly focused on standard face images and did not extend to low-resolution facial images. The lack of information in low-resolution images poses a challenge for effective recognition compared to standard face images. This limitation underscores the need for further research in developing network architectures that can effectively bridge the representation gap between low-resolution images and their high-resolution counterparts. Furthermore, although RobFaceNet has demonstrated robust performance on the cross-pose CP-LFW dataset, there remains scope for improvement in this domain. Future research opportunities could explore alternative network architectures capable of extracting more relevant features for cross-pose face recognition, thereby enhancing the model's ability to handle variations in pose.

6 Conclusions

Drawing inspiration from various mobile network design strategies, we introduce RobFaceNet, a novel lightweight, efficient, and robust network designed specifically for face recognition. Despite its simplicity, RobFaceNet demonstrates remarkable performance in both accuracy and computational efficiency. For instance, RobFaceNet achieves 95.95% and 92.23% accuracy on the CA-LFW and CP-LFW datasets, respectively, compared to 95.45% and 92.08% for the very deep ArcFace model. Meanwhile, RobFaceNet exhibits a more lightweight model complexity. In terms of computation cost, RobFaceNet has 337M FLOPs compared to ArcFace's 24211M, with only 3% of the parameters. Furthermore, when compared against lightweight FR models ($<$ 500M FLOPs), our proposed model outperforms the state-of-the-art in all evaluation datasets, achieving 99.75%, 97.42%, and 97.63% accuracy in the LFW, AgeDB-30, and CFP-FP datasets, respectively, compared to 99.55%, 96.07%, and 96.9% for MobileFaceNets [9] with a higher computational cost of 439.8M FLOPs.

The architecture of RobFaceNet employs a novel multi-feature approach, which combines features from different network levels and utilizes the modified h-swish activation function to reduce computational costs without compromising performance. Furthermore, we have designed an attention-enhanced bottleneck that improves the network's ability to discern crucial features at various levels, thereby boosting its face recognition capabilities. Through extensive experimentation across comprehensive public face verification benchmarks, we showcase RobFaceNet's effectiveness compared to deeper face recognition networks. These results underscore RobFaceNet's potential as a robust and efficient solution for face recognition tasks. These results highlight RobFaceNet's potential as a robust and efficient solution for face recognition tasks, especially in dynamic HRI environments where real-time process-

ing and interpretation of facial data are essential for facilitating meaningful and interactive engagements.

Author Contributions Conceptualization: Aly Khalifa; Methodology: Aly Khalifa; Software: Aly Khalifa, Ahmed A. Abdelrahman and Thorsten Hempel; Project heading: Ayoub Al-Hamadi; Funding acquisition: Ayoub Al-Hamadi; Formal analysis and investigation: Aly Khalifa and Ayoub Al-Hamadi; Writing - original draft preparation: Aly Khalifa; Writing - review and editing: Aly Khalifa, Ahmed A. Abdelrahman, Thorsten Hempel and Ayoub Al-Hamadi; Supervision: Ayoub Al-Hamadi.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the Federal Ministry of Education and Research of Germany (BMBF) project AutoKoWaT, no. 13N16336 and by the German Research Foundation (DFG) project AL 638/15-1, Al 638/14-1 and Al 638/13-1.

Data Availability All data generated or analyzed during this study are included in these published articles (CASIA-WebFace [53], VGGFace2 [8], MS-Celeb-1M [19], LFW [26], CFP-FP [43], CP-LFW [61], CA-LFW [62], AgeDB [40], IJB-B [49], IJB-C [39]).

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Abdelrahman AA, Hempel T, Khalifa A et al (2022) L2cs-net: fine-grained gaze estimation in unconstrained environments. arXiv:2203.03339
2. Alansari M, Hay OA, Javed S et al (2023) Ghostfacenets: Lightweight face recognition model from cheap operations. IEEE Access
3. Alonso-Fernandez F, Hernandez-Diaz K, Buades Rubio JM et al (2023) Squeezerfacenet: Reducing a small face recognition cnn even more via filter pruning. In: VIII International workshop on artificial intelligence and pattern recognition, IWAIPR
4. Bashbaghi S, Granger E, Sabourin R et al (2019) Deep learning architectures for face recognition in video surveillance. Deep Learn Object Detect Recognit 133–154
5. Boutros F, Damer N, Fang M et al (2021) Mixfacenets: Extremely efficient face recognition networks. In: 2021 IEEE international joint conference on biometrics (IJCB). IEEE, pp 1–8
6. Boutros F, Damer N, Kirchbuchner F et al (2022) Elasticface: Elastic margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1578–1587
7. Boutros F, Siebke P, Klemt M et al (2022) Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. IEEE Access 10:46,823–46,833
8. Cao Q, Shen L, Xie W et al (2018) Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp 67–74
9. Chen S, Liu Y, Gao X et al (2018) Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition. Springer, pp 428–438
10. Deng J, Guo J, Xue N et al (2019) Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 4690–4699
11. Deng J, Guo J, Zhang D, et al (2019) Lightweight face recognition challenge. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp 0–0

12. Ding C, Tao D (2017) Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE Trans Pattern Anal Mach Intell 40(4):1002–1014
13. Dong S, Wang P, Abbas K (2021) A survey on deep learning and its applications. Comput Sci Rev 40(100):379
14. Du H, Shi H, Zeng D et al (2022) The elements of end-to-end deep face recognition: A survey of recent advances. ACM Comput Surv (CSUR) 54(10s):1–42
15. Duong CN, Quach KG, Jalata I et al (2019) Mobiface: A lightweight deep learning face recognition on mobile devices. In: 2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS). IEEE, pp 1–6
16. Fard AP, Mahoor MH (2022) Facial landmark points detection using knowledge distillation-based neural networks. Comput Vis Image Underst 215(103):316
17. Gong R, Liu X, Jiang S et al (2019) Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 4852–4861
18. Guo MH, Xu TX, Liu JJ et al (2022) Attention mechanisms in computer vision: A survey. Comput Visual Media 8(3):331–368
19. Guo Y, Zhang L, Hu Y et al (2016) Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. Springer, pp 87–102
20. Han S, Mao H, Dally WJ (2015) Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv:1510.00149
21. He K, Zhang X, Ren S et al (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp 1026–1034
22. Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 13,713–13,722
23. Howard A, Sandler M, Chu G et al (2019) Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 1314–1324
24. Howard AG, Zhu M, Chen B et al (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861
25. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 7132–7141
26. Huang GB, Mattar M, Berg T et al (2008) Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition
27. Iandola FN, Han S, Moskewicz MW et al (2016) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv:1602.07360
28. Ketkar N, Ketkar N (2017) Stochastic Gradient Descent. A hands-on introduction. Deep Learn Python 113–132, https://doi.org/10.1007/978-1-4842-2766-4_8
29. Khalifa A, Abdelrahman AA, Strazdas D et al (2022) Face recognition and tracking framework for human-robot interaction. Appl Sci 12(11)
30. Kumar PM, Gandhi U, Varatharajan R et al (2019) Intelligent face recognition and navigation system using neural learning for smart security in internet of things. Clust Comput 22:7733–7744
31. Li J, Zhou L, Chen J (2024) Mobilefaceformer: a lightweight face recognition model against face variations. Multimedia Tools Appl 83(5):12,669–12,685
32. Li X, Wang F, Hu Q et al (2019) Airface: Lightweight and efficient model for face recognition. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp 0–0
33. Liang T, Glossner J, Wang L et al (2021) Pruning and quantization for deep neural network acceleration: A survey. Neurocomput 461:370–403
34. Liu W, Wen Y, Yu Z et al (2017) Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 212–220
35. Liu Z, Sun M, Zhou T et al (2018) Rethinking the value of network pruning. arXiv:1810.05270
36. Ma N, Zhang X, Zheng HT et al (2018) Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp 116–131
37. Martindez-Diaz Y, Luevano LS, Mendez-Vazquez H et al (2019) Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp 0–0
38. Martinez-Diaz Y, Nicolas-Diaz M, Mendez-Vazquez H et al (2021) Benchmarking lightweight face architectures on specific face recognition scenarios. Artif Intell Rev 1–44
39. Maze B, Adams J, Duncan JA et al (2018) Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 international conference on biometrics (ICB). IEEE, pp 158–165

40. Moschoglou S, Papaioannou A, Sagonas C et al (2017) Agedb: the first manually collected, in-the-wild age database. In: proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp 51–59
41. Sandler M, Howard A, Zhu M et al (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 4510–4520
42. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 815–823
43. Sengupta S, Chen JC, Castillo C et al (2016) Frontal to profile face verification in the wild. In: 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1–9
44. Strazdas D, Hintz J, Khalifa A et al (2022) Robot system assistant (rosa): Towards intuitive multi-modal and multi-device human-robot interaction. Sens 22(3):923
45. Tan M, Le QV (2019) Mixconv: Mixed depthwise convolutional kernels. arXiv:1907.09595
46. Vasu PKA, Gabriel J, Zhu J et al (2023) Mobileone: An improved one millisecond mobile backbone. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 7907–7917
47. Wang H, Wang Y, Zhou Z et al (2018) Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 5265–5274
48. Wei Y, Pan X, Qin H et al (2018) Quantization mimic: Towards very tiny cnn for object detection. In: Proceedings of the European conference on computer vision (ECCV). pp 267–283
49. Whitelam C, Taborsky E, Blanton A et al (2017) Iarpa janus benchmark-b face dataset. In: proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp 90–98
50. Woo S, Park J, Lee JY et al (2018) Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp 3–19
51. Wu X, He R, Sun Z et al (2018) A light cnn for deep face representation with noisy labels. IEEE Trans Inf Forensics Secur 13(11):2884–2896
52. Yan M, Zhao M, Xu Z et al (2019) Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp 0–0
53. Yi D, Lei Z, Liao S et al (2014) Learning face representation from scratch. arXiv:1411.7923
54. Zeiler MD, Ranzato M, Monga R et al (2013) On rectified linear units for speech processing. 2013 IEEE Int Conf Acoust. Speech and Signal Processing, IEEE, pp 3517–3521
55. Zhang H, Xu M (2020) Weakly supervised emotion intensity prediction for recognition of emotions in images. IEEE Trans Multimedia 23:2033–2044
56. Zhang K, Zhang Z, Li Z et al (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
57. Zhang L, Bao C, Ma K (2021) Self-distillation: Towards efficient and compact neural networks. IEEE Trans Pattern Anal Mach Intell 44(8):4388–4403
58. Zhang Q, Li J, Yao M et al (2019) Vargnet: Variable group convolutional neural network for efficient embedded computing. arXiv:1907.05653
59. Zhang Q, Zhang M, Chen T et al (2019) Recent advances in convolutional neural network acceleration. Neurocomput 323:37–51
60. Zhang X, Zhou X, Lin M et al (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 6848–6856
61. Zheng T, Deng W (2018) Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Tech Rep 5:7
62. Zheng T, Deng W, Hu J (2017) Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv:1708.08197
63. Zhong Y, Deng W, Hu J et al (2021) Sface: Sigmoid-constrained hypersphere loss for robust face recognition. IEEE Trans Image Process

Aly Khalifa

Ahmed A. Abdelrahman

Thorsten Hempel

Ayoub Al-Hamadi