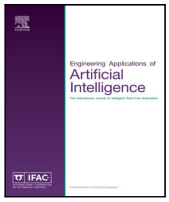
Engineering Applications of Artificial Intelligence 137 (2024) 109114

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

MixQuantBio: Towards extreme face and periocular recognition model compression with mixed-precision quantization

Jan Niklas Kolf [a,b,*], Jurek Elliesen [a,b], Naser Damer [a,b], Fadi Boutros [a]

[a] Fraunhofer Institute for Computer Graphics Research IGD, Fraunhoferstr. 5, Darmstadt, 64283, Hessia, Germany
[b] Technical University Darmstadt, Karolinenplatz 5, Darmstadt, 64289, Hessia, Germany

ARTICLE INFO

MSC:
68T07
68T10

Keywords:
Deep learning
Face recognition
Model quantization
Mixed-precision
Periocular recognition

ABSTRACT

Current periocular and face recognition approaches utilize computationally costly deep neural networks, achieving notable recognition accuracies. Deploying such solutions in applications with limited computational resources requires minimizing their computational demand while maintaining similar recognition accuracies. Model compression techniques like model quantization can be used to reduce the computational costs of deep models. This approach is widely studied and applied to different machine-learning tasks, however it is understudied and investigated for biometrics. We propose in this work to reduce the computational cost of face and periocular recognition models using fixed- and mixed-precision model quantization. Specifically, we first quantize the full-precision models to fixed 8 and 6 bits, reducing the required memory footprint by 5x while maintaining, to a very large degree, the recognition accuracies. However, our achieved results demonstrated that by quantizing the models to extremely low b bits, e.g., below 6 bits, the accuracies significantly dropped, which motivated our investigation on mixed-precision quantization. Hence, we propose to utilize an iterative mixed-precision quantization scheme. In each iteration, the least important parameters are selected based on their weight magnitude and quantized to low b-bit precision and the model is fine-tuned. This approach is repeated until all parameters are quantized to low b-bit precision, achieving extreme reduction in memory footprint, e.g., 16x times, without significant loss in the model accuracies. The effectiveness of mixed- and fixed-precision quantization for biometric recognition models is studied and proved for two modalities, face and periocular, using three different deep network architectures and using different b bit precision.

1. Introduction

The use of biometric recognition of people based on their physical or behavioral characteristics is becoming more and more common in everyday life (Jain et al., 2004). The most commonly used physical characteristics are the fingerprint, the face or parts of the face like the iris or the periocular region (Jain et al., 2004; Spolaor et al., 2016; Boutros et al., 2022d). Especially for mobile devices, using biometric recognition is an easy way to balance device security and usability (Spolaor et al., 2016; Sharma and Ross, 2022). Modern smartphones are equipped with front-facing cameras, allowing easy capture of the face or its regions (Spolaor et al., 2016). When face recognition (FR) is used, the entire facial image is used in the recognition task. To successfully perform recognition if the face is occluded, e.g. by masks (Boutros et al., 2021b), the periocular region can be used instead (Zanlorensi et al., 2020a). This region around the eye includes the eye itself, the eyebrow, the eyelid, and parts of the cheekbone (Sharma and Ross, 2022; Kumari and Seeja, 2019). This region contains enough discrepancy and uniqueness to perform periocular recognition (PR) successfully (Zanlorensi et al., 2020a; Kolf et al., 2022, 2023b).

The success of current state-of-the-art FR and PR systems is based on over-parameterized deep neural networks (DNN) (Boutros et al., 2022a; Deng et al., 2019a; Krizhevsky et al., 2012; Wang and Deng, 2021). However, these DNN are associated with high computational cost and a large memory footprint (Boutros et al., 2022a; Deng et al., 2019a; Cao et al., 2018). Especially in battery-powered mobile devices with shared resources, the use of DNN is limited by these restrictions (Cai et al., 2020; Deng et al., 2019b; Martínez-Díaz et al., 2021).

In order to utilize the high performance of DNN-based PR and FR systems on mobile devices, various techniques are used (Kolf et al., 2023a; Caldeira et al., 2024; Boutros, 2022). Efficient, small DNN architectures are being developed to reduce the number of computational operations and the amount of memory required while maintaining similar performance levels compared to conventional DNN designs (Boutros

* Corresponding author at: Fraunhofer Institute for Computer Graphics Research IGD, Fraunhoferstr. 5, Darmstadt, 64283, Hessia, Germany.
E-mail address: jan.niklas.kolf@igd.fraunhofer.de (J.N. Kolf).

https://doi.org/10.1016/j.engappai.2024.109114
Received 13 March 2024; Received in revised form 24 June 2024; Accepted 4 August 2024
Available online 26 August 2024

et al., 2022e; Li et al., 2020, 2019; Zhang et al., 2022). For example, this technique is used by ShuffleFaceNet (Martindez-Diaz et al., 2019), MobileFaceNets (Chen et al., 2018), and MixFaceNets (Boutros et al., 2021a), among others. Knowledge Distillation is a different approach, which uses a complex and high performing teacher network to train a compact student network (Boutros et al., 2020, 2022c; Liu et al., 2022). This method reduces the number of computational operations, but the performed calculations still use floating-point numbers that rely on special, dedicated hardware that also incurs a high computational effort (Boutros et al., 2022b).

To minimize the hardware-based computational complexity and the memory footprint, model quantization is used. In this approach, the calculations are not performed with floating-point, but with integer values. The integer values used have a reduced bit width compared to the floating-point numbers, which reduces both the computational cost and the memory requirements of the used DNN (Gholami et al., 2021). Model quantization also speeds up inference, i.e. the execution of DNN, by a factor between 2 to 4, as integer operations are performed faster based on the underlying hardware (Fei et al., 2021; Wu et al., 2016; Krishnamoorthi, 2018; Paszke et al., 2019). Typically, fixed bit widths are used for all integer parameters, such as 8, 6 or 4 bit (Krishnamoorthi, 2018; Gong et al., 2018; Jacob et al., 2018; Jin et al., 2022; Banner et al., 2019; Yao et al., 2021; Zhuang et al., 2018). This technique was successfully applied for FR (Boutros et al., 2022b; Bunda et al., 2022). Previous works (Kolf et al., 2022, 2023b) proposed the usage of model quantization for PR. These works have shown that model quantization with 8, 6 and 4 bit integers can be successfully used for FR and 8 and 6 bit integers for PR to reduce the computational cost of inference while maintaining recognition performance at a similar competitive level when compared to full-precision floating-point models. However, considering the FR and PR task, the performance drops rapidly when quantizing the utilized models with the applied method to low bit widths, e.g. 4 bit (Kolf et al., 2022; Boutros et al., 2022b) or 2 bit (Bunda et al., 2022). This indicates that the immediate reduction of the parameter's value range of the considered models is not sufficient to ensure a high recognition performance on FR and PR tasks.

Having all parameters of a model quantized to the same low precision value might not be optimal in terms of the trade-off between model memory footprint and its performance. This could be due to the fact that certain parameters might a greater influence on the performance of the DNN than other parameters (Liu et al., 2024; Cheng et al., 2023). In such a case, it is advantageous to assign a more meaningful value range with a higher bit width to the more influential parameters and to reduce the bit width for less influential parameters (Li et al., 2023). In this way, the performance of the DNN can be maintained to a large degree while the computational effort and memory footprint are reduced (Gholami et al., 2021; Liu et al., 2024). This approach is achieved through mixed-precision quantization (Rakka et al., 2022). In this approach, the bit width is not set the same for all integer values, but for each parameter the optimal bit width is determined. This minimizes the loss of information and allows the DNN to be better adapted to the reduced bit widths. Different methods use linear programming (Ma et al., 2023), reinforcement learning (Wang et al., 2019) or iterative techniques (Li et al., 2023) to calculate the bit widths per parameter. While these techniques have been successfully used in various computer vision tasks, these techniques have been applied with limited evaluation for FR (Bunda et al., 2022) and have not yet been applied and evaluated for PR in previous work (Sharma and Ross, 2022; Kolf et al., 2022, 2023b; Boutros et al., 2022b; Raja et al., 2019; Reddy et al., 2018a; Zhang et al., 2018a).

This work is an extension of our previous work (Kolf et al., 2022) published at the IEEE International Joint Conference on Biometrics (IJCB 2022) (Anon, 2022). In this paper, we extend our previous work on model quantization for PR (Kolf et al., 2022) and show that mixed-precision quantization can be successfully applied for PR. Furthermore, as an additional contribution, we investigate the performance of mixed-precision quantization on FR and extend the evaluation significantly beyond that of previous work (Bunda et al., 2022). We apply the MPQ framework (Li et al., 2023) and quantize FR and PR DNN iteratively to a minimum bit width of 2. In a wide range of test scenarios for FR and PR, for different backbone networks, the success of mixed-precision quantization is shown. E.g., our 2 bit quantized models are outperforming their respective 8 bit fixed-precision quantized and FP32 model counterparts on PR.

Section 3 presents the framework used, including the applied fixed- and mixed-precision quantization methodologies. In Section 4, the experimental setup for PR and FR including the respective datasets, evaluation metrics, and the quantization details are presented. The results of the experiments are discussed in Section 5 and the effects of different quantization levels are analyzed. The final conclusion is set out in Section 6.

2. Related work

In recent studies, various techniques were applied to achieve efficient DNN as well as efficient PR and FR models, specifically aimed for mobile and embedded devices.

Face Recognition: The current aim of efficient state-of-the-art DNN based FR models is to design lightweight models (Boutros et al., 2021a), which require less computational costs than larger and more complex DNNs (Boutros et al., 2022e). Examples of lightweight FR models include MobileFaceNet (Chen et al., 2018), SqueezeFaceNet (Yu and Li, 2019), VarGFaceNet (Yan et al., 2019), MixFaceNet (Boutros et al., 2021a), ShuffleFaceNet (Martindez-Diaz et al., 2019), and GhostFaceNets (Alansari et al., 2023). MobileFaceNet (Chen et al., 2018) adopts MobileNetV2 (Sandler et al., 2018) architecture and utilize fast down sampling strategies and global depth-wise convolutions to create a compact network with under 1M parameters. SqueezeFaceNet (Yu and Li, 2019) use SqueezeNet (Iandola et al., 2016) as base architecture and modify it to have reduced element-wise operations, resulting in a model with 0.74M parameters. VarGNet (Zhang et al., 2019) is adopted by VarGFaceNet (Yan et al., 2019) that uses variable group convolutional network, creating a FR model with 5M parameters. MixFaceNet (Boutros et al., 2021a) modifies MixConv (Tan and Le, 2019) block, adding channel shuffle operations, resulting in a compact FR model with 3.07M parameters. ShuffleFaceNet (Martindez-Diaz et al., 2019) adapts ShuffleNet (Ma et al., 2018) and utilizes global depth-wise convolutional layers instead of global average pooling, creating a model with 2.6M parameters. GhostNet (Han et al., 2020) is utilized by GhostFaceNets (Alansari et al., 2023) to create a compact FR model that contains 4.09M parameters. These lightweight architectures find a balance between computational complexity and accuracy. While these models have reduced the computational time through architecture design, their execution remains computational challenging, as the inference with floating-point parameters requires dedicated hardware to be executed efficiently. With the aim to create efficient and small DNN architecture specifically tailored to FR, Neural Architecture Search and AutoML techniques were successfully applied (Boutros et al., 2022e; Savchenko et al., 2023).

Apart from efficient architectures, model quantization was also examined. Fixed-precision model quantization was applied in Boutros et al. (2022b) and Bunda et al. (2022). Later, solution presented in the 2023 Efficient FR competition (Kolf et al., 2023a) adapted this approach to achieve compact FR models. In their approach, the authors quantized different backbone architectures to small bit width, reducing computational complexity and model footprint while maintaining competitive performance. Mixed-precision quantization was shortly evaluated in a limited experimental setup in Bunda et al. (2022), lacking evaluation on various benchmarks and bit widths.

From the previous works, model quantization proved to largely reduce the memory footprint of FR models while maintaining the

performance (Boutros et al., 2022b; Bunda et al., 2022). Based on underlying hardware and processors, speedups from up to two folds can be observed when quantizing a floating-point model to 8 bit. However, it was reported (Boutros et al., 2022b; Bunda et al., 2022) that quantizing FR models to extremely low precision levels (to 4 and 2 bits) impacts the model convergence during the fine tuning process, resulting in extremely low performances. In order to stabilize convergence at low bit widths, other optimizers and hyper parameters were used in Bunda et al. (2022). However, this still results in a low FR performance with quantization to 2 bits. Previous works have shown the fundamental success of model quantization. In order to further enhance the deployability of DNN and reduce the computational effort, the main motivation of this work is to reduce the required bit widths of the DNN parameters even further while maintaining recognition performance to a large degree.

Periocular Recognition: In Park et al. (2009) the authors conducted a feasibility study using the periocular region with Local Binary Patterns (LBP) for recognition, which was followed up and adapted by similar approaches (Bharadwaj et al., 2010; Xu et al., 2010). Additionally, Scale Invariant Feature Transformation (SIFT) has been applied (Ross et al., 2012; Alonso-Fernandez and Bigun, 2014; Ahuja et al., 2016; Raja et al., 2014). SIFT adopted for mobile devices was introduced in Raja et al. (2014). Leung-Mallik filters used to acquire multiple features of the periocular region were utilized in Tan and Kumar (2012).

Deep learning mechanisms have also been used to perform PR on mobile devices. In Nie et al. (2014) a convolutional version of Restricted Boltzman Machines was applied to learn features in an unsupervised approach which were then fused into a metric learning system. Features of the iris and periocular region extracted from a convolutional neural network (CNN) were fused for PR recognition in Zhang et al. (2018a). Raja et al. (2016) used deep learning to create features from the periocular region for recognition. Feature learning with a convolutional autoencoder in an unsupervised learning approach was proposed in Reddy et al. (2019), outperforming a supervised ResNet50. In Proença and Neves (2017) a data augmentation pipeline to modify the periocular area was developed, increasing recognition performance. To reduce within-class variability in unconstrained PR, Zanlorensi et al. (2020b) proposed to utilize generative models. In Alonso-Fernandez et al. (2022) several feature extractors were combined and validated with cross-spectrum experiments between visible and near-infrared images, achieving good recognition rates. Various algorithms were evaluated in a competition (Rattani et al., 2016) targeting mobile devices under different conditions.

To reduce the computational effort for PR on mobile devices, OcularNet was proposed in Reddy et al. (2018b). It uses multiple small CNNs to extract features from six overlapping patches from the periocular region, and combined them into a feature vector. This model outperformed a ResNet50 model, despite being 15.6x smaller. Pruning techniques were applied in Almadan and Rattani (2021) to create a small and efficient NN for ocular recognition on mobile devices. Knowledge distillation was applied in recent works (Boutros et al., 2020, 2022c) to transfer the knowledge learned by larger networks to smaller ones, either in a traditional manner (Boutros et al., 2020) or on the template level (Boutros et al., 2022c). A recent work (Kolf et al., 2022) was the first to propose to use fixed-precision model quantization for PR. Parameters were quantized to 8, 6 and 4 bit, which yielded efficient and compact models that retained their recognition performance when quantized to 8 and 6 bit. The quantization to 4 bit resulted in a significant decrease in performance, suggesting that direct quantization to a low bit width can lead to a significant drop in performance. Such as for FR, PR model quantization proved to be a very efficient technique to achieve high performances with models having much smaller memory footprint. However, as stated for PR and FR, going to a very low quantization precision (lower than 6 bits) resulted in a huge loss in performance.

This performance loss, when quantizing FR and PR models to very low precision levels might be due to the need of higher precision parameters to achieve accepted performances. However, these required higher precision parameters are not necessary all the parameters of the network. Thus, optimizing the quantization process depending on the precision needs in different networks parts might be the solution that enables extreme model compression. Based on previous success of mixed-precision quantization on different tasks (Li et al., 2023; Zhou et al., 2016), this work proposes and investigates the use of mixed-precision quantization for PR and FR tasks to overcome the current limitations of fixed-precision quantization in these domains.

3. Methodology

This section presents our approach for reducing the computational cost and memory footprint of DNN used for PR and FR by utilizing model quantization. Towards that, we first propose to quantize the model weights and activations using fixed-precision quantization. As the quantization model accuracies (as will be demonstrated in Section 5) significantly dropped when we utilized a very small low bit value, we theorize that this quantization effect on the performance is not necessarily the same for all quantized parameters. Therefore, as one of the first works, we propose to apply mixed-precision quantization for PR and FR (Kolf et al., 2022, 2023b,a). This aims at quantizing each parameter to a different b bit value based on the contribution of this parameter to the model prediction. An overview of the proposed approach is shown in Fig. 1. This section first presents additional theoretical background to model quantization. Then, this section presents the utilized approach for fixed-precision quantization and the conceptual differences between fixed and mixed-precision quantization. Finally, we present the utilized approach for mixed-precision quantization.

3.1. Theoretical background

The core concept of model quantization is to replace the computationally challenging floating-point parameters, often represented with 32 bit (FP32), with b bit integer parameters (Gholami et al., 2021). Typical bit widths for b are 8, 6, 4, 2, or 1 (Gholami et al., 2021; Li et al., 2023). The advantages of this approach are twofold, since storing fewer bits reduces the memory footprint during inference and, depending on the underlying hardware, integer operations are performed much more efficiently than floating point operations, increasing DNN throughput and decreasing its computational cost (Krishnamoorthi, 2018; Paszke et al., 2019; Weng, 2021). The weights and activation functions of a DNN are typically quantized to a fixed bit precision. However, the impact of quantization on different parameters may vary. Consequently, it is beneficial to utilize mixed-precision quantization, where different parameters are quantized to different precision levels, to achieve a more optimal trade-off between model memory footprint and performance. In fixed-precision quantization, all parameters and activation functions are quantized to b_p and b_a bit, respectively (Jacob et al., 2018; Guo, 2018). In mixed-precision quantization, the bit width b is individually set for each parameter and activation function (Gholami et al., 2021).

To maintain the performance of the FP32 DNN, quantizing a parameter to b bit must result in a behavior similar to that of the full-precision parameter. However, quantizing FP32 values with a continuous value range to a discrete, lower-dimensional value range of integers is associated with a loss of information. Therefore, in practice, the quantized DNN has to be adapted, i.e. calibrated, and fine-tuned, to the reduced value range with additional training steps (Gholami et al., 2021). This procedure is called Quantization-Aware Training (QAT), which is explained in more detail in this section.

The range of values of an FP32 parameter r is $r \in [-3.4 \times 10^{38}, 3.4 \times 10^{38}]$ (Van Der Walt et al., 2011) and is defined in the respective IEEE standard (Zuras et al., 2008). While FP32 numbers have a wide range of values, in practice DNN do not fully utilize this range. This is because

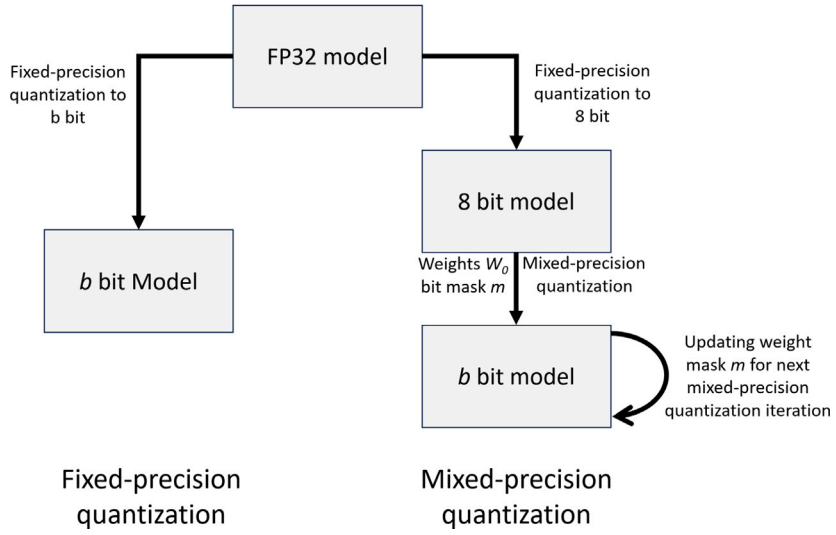


Fig. 1. Overview of the proposed fixed- and mixed-precision quantization approach. In fixed-precision quantization, an FP32 model is quantized to b bit (8, 6, and 4 bit). In mixed-precision quantization, an FP32 model is first quantized to 8 bit using symmetric quantization and the weights are stored. In each mixed-precision quantization iteration, the initial weights are loaded and the model is trained until convergence. For p percent of parameters with the smallest weight magnitude, the bit width is halved, reducing average bit width of the model. The iterative quantization ends when all parameters are quantized to 2 bit.

various activation functions have certain fixed value ranges. Also, the regularization and normalization terms in the optimization procedure prevent individual parameters from growing too big or too small.

For an FP32 parameter r within a DNN, let $r \in [\alpha, \beta]$ (Gholami et al., 2021). The value range of a signed b bit integer r_Q is defined as $r_Q \in [-2^{b-1}, 2^{b-1} - 1]$ and contains 2^b different values. A suitable mapping function is required to map the values from r to r_Q. The mapping can be performed uniformly, where the values are equally distributed to $[-2^{b-1}, 2^{b-1} - 1]$, or non-uniformly, where the distances between individual values in r in the new value range are different. Uniform distribution is due to its simple implementation the most prevalent method (Gholami et al., 2021; Guo, 2018).

A further distinction can be made between symmetric and asymmetric quantization (Gholami et al., 2021). In symmetric quantization, the boundaries of the FP32 value range r are set to equal values, i.e. $\alpha = -\beta$. While in asymmetric quantization, the boundaries α and β are determined during a calibration process and are based on the current FP32 value range of a single parameter or parameter group. Therefore, in asymmetric quantization $\alpha \neq -\beta$. In this work, we follow the asymmetric quantization scheme of Kolf et al. (2022), Boutros et al. (2022b) for the fixed-precision quantization (Cai et al., 2020). For mixed-precision quantization, we follow and adapt the procedure of Li et al. (2023), which uses the symmetric quantization method (Zhou et al., 2016). Both techniques, mixed and fixed-precision quantization, are explained in more detail in the following section.

Most quantization schemes do not directly replace FP32 parameters with integer parameters (e.g. int8, int16), but simulate integer quantization (Gholami et al., 2021). As modern DNN frameworks such as Pytorch (Paszke et al., 2019) only train DNN that utilize floating point values, the parameters are quantized from FP32 to b bit integers and are immediately dequantized from integer to FP32 during training. This ensures that the associated loss of information incurred by quantization is included in the training process. Therefore, the network is adapted to this loss of information, and it can be trained using deep learning accelerators such as Pytorch (Paszke et al., 2019).

3.2. Fixed-precision quantization

The adapted fixed-precision quantization technique (Cai et al., 2020) uses an asymmetric and uniform mapping between the FP32 and the integer value range. This quantization procedure is defined as (Gholami et al., 2021; Jacob et al., 2018):

$$r_{\text{Q-Int}} = Q_{A,b}(r) = round\left(\frac{r_{\text{FP32}}}{S}\right) - Z, \tag{1}$$

where the operation $round(\cdot)$ rounds the floating-point number to the nearest integer. The b bit signed integer value Z is the zero-point. It specifies the zero-point shift of the mapping from the range of values of r_{FP32} and that of $r_{\text{Q-Int}}$. The constant S is the scaling factor and divides the value range into equal-sized partitions. Since the range of values from r_{FP32} can be asymmetric ($-\alpha \neq \beta$) S and Z are defined as (Jacob et al., 2018)

$$S = \frac{\beta - \alpha}{2^b - 1}, \tag{2}$$

$$Z = round\left(\beta \cdot \frac{2^b - 1}{\beta - \alpha} - 2^b - 1\right). \tag{3}$$

To reconstruct the original FP32 value from a quantized value, there is the dequantization operation (Gholami et al., 2021; Jacob et al., 2018):

$$r_Q = D(r_{\text{Q-Int}}) = S \cdot (r_{\text{Q-Int}} + Z). \tag{4}$$

The range of r, limited by α and β, can be calculated statically or dynamically. If it is calculated dynamically, the minimum and maximum values for individual parameters are stored during retraining, and after a specified number of calibration steps this calculation is frozen and the values are set as α and β (Gholami et al., 2021). In this work, dynamic calibration is used to obtain a better parameter mapping.

The parameter ranges can also be calculated for different groups of parameters, which is called quantization granularity. In CNN, layers consists of multiple channels (i.e. filters), that each can have different value ranges, which can therefore be quantized differently. There are three different methodologies with increasing granularity (Gholami et al., 2021; Xiao et al., 2023): Layerwise, groupwise, and channelwise quantization. In layerwise quantization, all convolution channels of an entire layer are considered in order to determine α and β. Groupwise quantization increases the granularity, as the channels of a layer are divided into groups for which the range of values is determined. Channelwise quantization increases the granularity even further. There, the range of values is determined for individual channels of a convolutional layer. Due to its high popularity and good performance, channelwise quantization is used for asymmetric quantization in this work (Huang et al., 2021; Zhang et al., 2018b).

3.3. Mixed-precision quantization

While previous works have shown that model quantization to 8, 6 and 4 bit for FR (Boutros et al., 2022b; Bunda et al., 2022) and 8 and 6 bit for PR (Kolf et al., 2022) can be applied without a major performance impact, quantizing models directly to low bit width can result in major performance degradation (Gholami et al., 2021), as shown in Kolf et al. (2022), Boutros et al. (2022b), Bunda et al. (2022). One way to balance accuracy and model size is to use mixed-precision quantization. Parameters that have a large impact on accuracy are quantized to integers with a larger bit width, while parameters with less influence are reduced to smaller bit widths, resulting in a significant decrease in the number of bits needed on average, while still providing performance comparable to FP32 models.

We adopted the approach presented by Li et al. (2023) for mixed-precision quantization. It is based on an iterative quantization method, resulting in models with an average low bit width, but with similar performance as the respective full-precision DNN.

At the start, a bit mask m is created that sets the desired start bit width b for all parameters W in the DNN. Each parameter $w_i \in W$ receives its individual bit width $b_i \in m$. At the beginning of the initial training, all b_i are set to b. The individual bit widths in m are used to quantize all parameters W and the DNN is trained until it converges. After the initial training, the weights W_0 of the DNN are stored and used as the starting weights for each iteration.

In each iteration of the quantization procedure, the weights W_0 and bit mask m are loaded and each parameter w_i is quantized to bit width $b_i \in m$. The DNN is again trained until convergence and its weights W are saved for evaluation. To reduce the average bit width of the parameters, p percent of parameters that have a current bit width > 2 and that have the lowest absolute weight value, i.e. $|w_i|$, are selected. For these parameters, their respective bit width $b_i \in m$ is updated as $b_i = \lfloor \frac{b_i}{2} \rfloor$, reducing the average bit width of all parameters. The updated bit mask m is saved and loaded in the next iteration, the stored parameters W_0 used at the start of every iteration remain unchanged. This iterative process is repeated until all parameters are set to 2 bit.

To quantize the weights of a model to a desired bit width b during each iteration, the DoReFa method introduced by Zhou et al. (2016) is used, following Li et al. (2023). It is a parameter-free, symmetric, and uniform quantization method that relies only on the desired bit width b. An FP32 value $r_{\text{FP32}} \in [0, 1]$ is quantized to 2^b different values using the symmetric mapping function (Li et al., 2023; Zhou et al., 2016)

$$r_{\text{Q-FP32}} = Q_{S,b}(r_{\text{FP32}}) = \frac{1}{2^b - 1} round((2^b - 1) r_{\text{FP32}}), \tag{5}$$

where the operation $round(\cdot)$ rounds the floating-point number to the nearest integer. The function $Q_{S,b}$ simulates the influence of integer quantization while $r_{\text{Q-FP32}} \in [0, 1]$ is an FP32 value.

As in general $r_{\text{FP32}} \notin [0, 1]$ for a parameter r_{FP32} in a pre-trained FP32 DNN, an additional scaling operation is performed using $\tanh(\cdot)$ that scales each input to the value range $r_{\text{FP32}} \in [0, 1]$. The full quantization method quantizes an FP32 parameter to 2^b different values and outputs the dequantized FP32 value $r_{\text{Q}} \in [-1, 1]$ (Li et al., 2023; Zhou et al., 2016):

$$r_{\text{Q}} = 2 \times Q_{S,b}\left(\frac{\tanh(r_{\text{FP32}})}{2 \cdot \max(|\tanh(r_{\text{FP32}})|)} + \frac{1}{2} \right) - 1 \tag{6}$$

To quantize the activation functions, the PACT methodology introduced by Choi et al. (2018) is utilized. It learns an adaptive clip threshold α per layer in the network. The function is defined as (Li et al., 2023; Choi et al., 2018)

$$a_{\text{FP32}} = \text{PACT}(r_{\text{FP32}}) = \begin{cases} 0, & r_{\text{FP32}} \in (-\infty, 0) \\ r_{\text{FP32}}, & r_{\text{FP32}} \in [0, \alpha) \\ \alpha, & r_{\text{FP32}} \in [\alpha, +\infty) \end{cases} \tag{7}$$

Similar to DoReFa quantization, the activation value a_{FP32} is quantized and immediately dequantized using the formula (Li et al., 2023; Choi et al., 2018)

$$a_Q = round\left(a_{\text{FP32}} \cdot \frac{2^b - 1}{\alpha} \right) \cdot \frac{\alpha}{2^b - 1}. \tag{8}$$

3.4. Quantization-aware training

The mappings utilized between floating-point and integer numbers involve a loss of representativeness of the DNN, as the learned parameters of the DNN are not trained to deal with the lower expressiveness of each parameter. This results in a loss of performance due to quantization. To counteract this, the DNN is fine-tuned to adjust the quantized parameters (Gholami et al., 2021). This process is called quantization-aware training (QAT).

4. Experimental setup

This section presents the experimental setups followed in this work to train and evaluate full-precision, fixed-precision, and mixed-precision FR and PR models.

4.1. Dataset and evaluation protocols

Face Recognition: The baseline FR models are trained on the MS1MV2 dataset (Guo et al., 2016; Deng et al., 2019a). It is a refined version of the MS-Celeb-1M (Guo et al., 2016) dataset. MS1MV2 contains 5.8M images of 85k identities that are cropped by the dataset authors to 112×112 pixels using the procedure described in Deng et al. (2019a).

The full-precision and quantized FR models are evaluated on Labeled Faces in the Wild (LFW) (Huang et al., 2007), Celebrities in Frontal-Profile in the Wild (CFP-FP) (Sengupta et al., 2016), Cross-age LFW (CALFW) (Zheng et al., 2017), AgeDB30 (Moschoglou et al., 2017), and Cross-Pose LFW (CPLFW) (Zheng and Deng, 2018) datasets. Following the respective evaluation protocol of each benchmark, we report the verification accuracy in percent.

AgeDB (Moschoglou et al., 2017) is an in-the-wild dataset for the evaluation of age-invariant face verification, containing 16,488 images from 568 identities. We present the results for AgeDB-30 (30 year age difference), which is the most widely reported and difficult subset of AgeDB.

The Labeled Faces in the Wild (LFW) dataset (Huang et al., 2007) is an unrestricted face verification dataset. It consists of 13,233 images of 5749 identities collected from the web. The performance of the LFW is measured using the "unrestricted with labeled outside data" protocol and the standard 6000 comparison pairs, 3000 genuine and 3000 imposter pairs (Huang et al., 2007).

The Celebrities in Frontal-Profile in the Wild (CFP-FP) dataset (Sengupta et al., 2016) focuses on the comparison between frontal and profile faces. It consists of 7,000 images of 500 different identities, with 10 frontal and 4 profile pictures per identity. The evaluation protocol for CFP-FP includes 3,500 genuine pairs and 3,500 imposter pairs.

The Cross-age LFW (CALFW) dataset (Zheng et al., 2017) is based on the Labeled Faces in the Wild LFW dataset, with a focus on comparison pairs that have an age gap, though not as large as AgeDB-30. The age gap distribution of CALFW is described in Zheng et al. (2017). It consists of 3000 genuine comparisons, and the negative pairs are chosen to be of the same gender and race in order to reduce the influence of attributes.

The Cross-Pose LFW (CPLFW) dataset (Zheng and Deng, 2018) is a collection of 3000 genuine comparison pairs with pose differences, taken from the LFW dataset. Negative pairs are chosen to be of the same gender and race.

Periocular Recognition: To evaluate, train, and quantize DNN for PR, the UFPR-Periocular dataset (Zanlorensi et al., 2020a) is selected.

It is one of the largest periocular databases available for research and is composed of 33,660 images of periocular regions, taken from both sides of the face, captured from 1,122 subjects in 3 sessions with 196 different mobile devices. These images were not taken in an ideal setting and have variations in blur, obstruction, and lighting to demonstrate real-world capturing scenarios. The images are aligned and cropped by the authors of the dataset, the procedure is described in Zanlorensi et al. (2020a). The images are cropped to 224 × 224 pixels. The UFPR dataset is selected as it is the most diverse dataset for periocular images with a large and diverse set of capturing devices.

The dataset contains three different training and evaluation protocols. The first protocol is the closed world protocol and represents an identification scenario. There, all identities contained in the training set are also contained in the validation and test set. The other two protocols are for the verification scenario. In both verification protocols, the test set does not contain images of the identities that are included in the training set. These protocols are noted as the open world scenario. The difference between the two verification protocols is that in the first protocol, the identities from the training set are included in the validation set, which is called open world/closed validation. In the verification protocol, there are different identities in the training, validation, and test set (i.e. identity-disjoint). This protocol is called open world/open validation. Each protocol is divided into three different folds, for each of which there is a training, validation, and test set. For a more detailed description of the dataset, please refer to the work of Zanlorensi et al. (2020a). We follow the procedure of Uzair et al. (2013) and flip images containing the left eye horizontally so that they have the same orientation as images with the right eye. Images from the left and right eyes are considered to have the same identity in our experimental setup. All images are scaled to size 224 × 224. We follow the process described in Zanlorensi et al. (2020a) and train and evaluate the models and quantization for both the closed world identification protocol and the open world/closed validation verification protocol. The benchmark in each protocol is run and evaluated for all three folds. As described in Zanlorensi et al. (2020a), the rank 1 and rank 5 metrics, calculated from the pairwise cosine similarity between the embedding vectors, as well as the area under the curve (AUC), the equal error rate (EER) and FNMR at different FMR are calculated for the identification protocol. The metrics are explained in more detail in the following Section. For the evaluation of the verification scenario, the AUC, EER, and FNMR at different FMR operation points are used.

4.2. Metrics

The rank metric used for identification scenarios indicates in what percentage of the cases examined the identity of a person in the database is the 1st place (rank 1) or is within the first five rankings (rank 5). The ranking is created by the cosine similarity between the input and the database, the test set, and is sorted in descending order according to the similarity score. The Equal Error Rate (EER) is defined as the operation point where the false match rate (FMR) is equal the false non-match rate (FNMR), based on the ISO/IEC 19795-1 (Mansfield, 2006) standard. For further evaluation, different FNMR at FMR are given. The receiver operating characteristics (ROC) curve gives the individual 1 − FNMR and FMR for more operation points (Jain et al., 2004). The area under curve (AUC) metric specifies how good the verification performance is. A 1.0 metric means that the system can perfectly distinguish comparisons between the identical identity and foreign identities.

4.3. Backbone architectures and baselines

We follow the previous works (Kolf et al., 2022; Boutros et al., 2022b) and utilize the DNN architectures ResNet18 (He et al., 2016a), ResNet50 (He et al., 2016a) and MobileFaceNet (Chen et al., 2018). All models are trained with a Stochastic Gradient Descent optimizer with a momentum of 0.9 and a weight decay of $5e-4$.

Face Recognition: Following Boutros et al. (2022b), the FP32 models are trained using Arcface (Deng et al., 2019a) loss and a margin value of $m = 0.5$ and a feature scale of $s = 64$ for 180k iterations on MS1MV2 dataset. The initial learning rate is 0.1. It is reduced by a factor of 10 after 100 k and 160 k iterations. A mini-batch size of 512 is chosen for each optimization step.

Periocular Recognition: We train all baseline FP32 models with a learning rate of 0.1 for 20 epochs with Arcface loss (Deng et al., 2019a), following Kolf et al. (2022). The model with the best performance on the validation set is selected. For a better comparison with mixed-precision quantized models, additional training is performed with 20 epochs using a learning rate of 0.01, which is divided by 10 after 4 epochs. The Arcface margin value is set to $m = 1.0$ and the feature scale is set to $s = 64$, following our previous work (Kolf et al., 2022). We follow Boutros et al. (2022a), and decrease the learning rate by a factor of 10 after the 8th and 15th epochs. A mini-batch size of 32 is chosen for each optimization step.

4.4. Fixed-precision model quantization

The pre-trained models are quantized using the asymmetric quantization scheme and quantization-aware training, introduced in Section 3.3 and Section 3.4, respectively.

Face Recognition: We follow Boutros et al. (2022b) and quantize the weights and activations of the FP32 models to 8 and 6 bit. Quantized models are fine-tuned for 11k iterations and a learning rate of 10^{-4} on the MS1MV2 (Guo et al., 2016; Deng et al., 2019a).

Periocular Recognition: The models for both the verification and identification task are quantized to fixed bit levels. During the quantization, the weights and activations are quantized to the identical bit level. The models are separately quantized 8, 6 and 4 bits. Quantization-aware training is conducted for 20 epochs using the UFPR periocular dataset. The learning rate for the 8 and 6 bit quantization is 0.1, for 4 bit quantization, the learning rate is 0.01. A batch size of 8 was used for ResNet18 and ResNet50 and a batch size of 16 for MobileFaceNet, following Kolf et al. (2022). After the 8th and 15th epochs, the learning rate is reduced by a factor of 10.

4.5. Mixed-precision model quantization

All activations of the FP32 models are quantized to 8 bit using the PACT quantization scheme, following Li et al. (2023). The weights of the models are quantized using the symmetric DoReFa quantization, which is described in Section 3.2. The starting bit width for the weights is 8 bit, as previous work (Kolf et al., 2022; Boutros et al., 2022b) has not shown a major performance loss for this bit width. The bit width of the weights is reduced after each iteration. After each step, the bit width is halfed for 50% of the weights with the least importance. The quantization algorithm is performed for 12 iterations. In the final iteration, the bit widths of all weights are set to 2 bit.

Face Recognition: The quantized models are fine-tuned until convergence for 8 epochs and a learning rate of 0.005 on the MS1MV2 dataset.

Periocular Recognition: The quantized verification and identification models are fine-tuned for 20 epochs on the respective UFPR training dataset, following Kolf et al. (2022). A learning rate of 0.01 is used, which is reduced to 0.001 after the fourth epoch.

Table 1
Comparison of Fixed and Mixed-Precision Quantized ResNet18, ResnNet50 and MobileFaceNet Models at Different Bit Levels on the Utilized Face Recognition Benchmarks LFW, AgeDB30, CALFW, CPLFW and CFP-FP. The Verification Accuracy is Given in Percent. Models Quantized To Bit Widthx Below 6 Bit Using Fixed-Precision Quantization do not Converge During the Fine-Tuning Process. Applying Mixed-Precision Quantization Results in Quantized Models with Low Bit Widths Not Possible With the Fixed-Precision Quantization Scheme. The Low Bit Width Models Maintain Their Verification Accuracy To a Large Degree, when Compared to the Respective FP32 model.

Model	Precision	Avg. bits	Size (MB)	LFW (↑)	AgeDB30 (↑)	CALFW (↑)	CPLFW (↑)	CFP-FP (↑)
ResNet18	Fixed (Boutros et al., 2022b)	32.00	96.10	99.67	97.13	95.70	89.73	94.47
		8.00	24.03	99.63	97.03	95.72	89.48	94.46
		6.00	18.02	99.52	96.55	95.58	88.37	93.23
	Mixed	8.00	24.03	99.48	96.28	95.27	88.22	92.29
		6.00	18.02	99.52	96.03	95.20	88.82	92.04
		4.61	13.85	99.40	95.63	94.73	87.55	91.44
		3.99	11.97	99.52	95.48	94.70	87.23	90.84
		3.30	9.91	99.27	95.02	94.63	87.28	90.53
		2.99	8.99	99.40	95.13	94.63	87.32	90.19
		2.65	7.96	99.38	94.57	94.45	87.07	90.04
		2.50	7.50	99.38	94.98	94.25	86.45	89.79
		2.32	6.98	99.42	95.38	94.58	86.07	90.21
		2.25	6.75	99.37	94.97	94.63	86.47	90.10
		2.09	6.27	99.38	94.90	94.47	86.10	89.84
		2.00	6.01	99.27	93.98	94.47	86.47	89.46
ResNet50	Fixed (Boutros et al., 2022b)	32.00	174.36	99.80	98.08	96.10	92.43	98.01
		8.00	43.59	99.78	98.00	96.00	92.17	97.70
		6.00	32.69	99.70	97.17	95.87	90.17	95.00
	Mixed	8.00	43.59	99.77	97.43	95.93	91.27	94.94
		6.00	32.69	99.72	97.65	95.82	90.90	94.81
		4.60	25.06	99.70	96.95	95.58	90.68	93.99
		3.99	21.73	99.68	96.80	95.52	89.83	93.76
		3.29	17.93	99.62	96.85	95.75	89.95	94.06
		2.99	16.31	99.58	97.05	95.67	90.28	94.03
		2.65	14.43	99.62	96.90	95.52	89.83	93.83
		2.50	13.60	99.65	97.03	95.57	89.88	93.39
		2.32	12.66	99.48	96.32	95.42	89.62	93.97
		2.25	12.25	99.62	96.73	95.62	89.87	93.64
		2.09	11.38	99.65	96.95	95.62	89.55	93.76
		2.00	10.90	99.62	96.75	95.62	89.53	93.33
MobileFaceNet	Fixed (Boutros et al., 2022b)	32.00	4.01	99.47	95.62	95.15	87.98	91.59
		8.00	1.00	99.43	95.47	95.05	87.95	91.40
		6.00	0.75	98.87	93.03	93.30	84.57	87.69
	Mixed	8.00	1.00	99.27	94.33	94.18	85.82	89.57
		6.00	0.75	99.20	93.70	93.82	85.77	89.14
		4.64	0.58	98.27	89.17	92.03	77.90	81.36
		3.97	0.50	98.05	89.67	91.50	79.38	82.27
		3.28	0.41	97.50	85.97	89.12	73.40	77.40
		2.97	0.37	97.32	86.37	89.12	75.72	78.56
		2.64	0.33	97.30	85.35	89.98	73.78	78.31
		2.48	0.31	97.75	85.70	89.52	72.05	77.70
		2.32	0.29	97.35	86.42	90.18	74.33	78.56
		2.24	0.28	96.88	84.90	88.38	73.02	77.64
		2.09	0.26	97.12	85.27	88.60	71.88	76.84
		2.00	0.25	96.88	86.17	89.12	73.42	76.96

5. Results

Table 1 shows the achieved face verification performance and Tables 2 and 3 show for comparison recent efficient FR results. In Table 4 and Table 5, the verification and identification results on the UFPR periocular dataset are shown, respectively. The results are grouped by the backbone model (ResNet18, ResNet50, MobileFaceNet) and the quantization type (fixed and mixed-precision). When quantizing an FP32 model to 8 bits, the model size is reduced by 75%, leading to a much smaller memory footprint and a decrease in the computational cost of model inference. Quantizing to 4 bits reduces the model size by 87.5%, and to 2 bits by 93.75%. In the following, we analyze in more detail different aspects of the presented results.

5.1. Impact of mixed-precision quantization

The goal of the proposed solution is to achieve extremely low precision quantization, and thus memory footprint, while maintaining as much performance as possible. The results of our extensive evaluation setting, explained in this Section, demonstrate that mixed-precision quantization is indeed achieving remarkable results in low-bit width scenarios.

Face Recognition: From Table 1 it is noticeable that fixed-precision quantized models maintain the performance of the respective full-precision model when quantized to 8 bits, with a very limited performance drop. This performance drop is increased substantially when moving to 6 bit quantization. As stated in Boutros et al. (2022b), moving to fixed-precision quantization lower than 6 bit showed that none of the models converged during the fine-tuning process. The aim of the proposed mixed-precision quantization is to achieve a lower average quantization precision than 6 bit, while still preserving a significant amount of performance.

The 8 bit mixed-precision quantized models have a slightly lower performance than the 8 bit fixed-precision quantized models, as shown in Table 1. This difference might result from the applied quantization method and is not a major concern when evaluating the performance, since the focus of the work is the lower bit width scenarios. At quantization levels of 6 bit, the mixed-precision quantization method achieves

Table 2
Comparison of our Mixed-Precision Quantized MobileFaceNet Model to Recent Efficient Face Recognition Models Evaluated During the Efficient Face Recognition Competition 2023 (Kolf et al., 2023a) on the Utilized Face Recognition Benchmarks LFW, AgeDB30, CALFW, CPLFW and CFP-FP. For Each Benchmark, the Verification Accuracy is Given in Percent. Using the Proposed Quantization Schemes to Reduce the Bit Width of the Model Yields Comparable Results to other Evaluated FP32 Models, While Significantly Reducing the Memory Footprint and Computational Overhead of the Quantized Model.

Model	Avg. bits	Size (MB)	LFW	AgeDB30	CALFW	CPLFW	CFP-FP
MobileFaceNet Q8-bit (Boutros et al., 2022b)	8.0	1.10	99.43	95.47	95.05	87.95	91.40
MobileFaceNet Q6-bit (Boutros et al., 2022b)	6.0	0.79	98.87	93.03	93.30	84.57	87.69
PocketNetM-256 (Boutros et al., 2022e)	32.0	7.0	99.58	97.17	95.63	90.03	95.66
PocketNetM-128 (Boutros et al., 2022e)	32.0	6.74	99.65	96.78	95.67	90.00	95.07
Idiap EdgeFace-XS(γ = 0.6) (Kolf et al., 2023a)	32.0	7.17	99.68	95.72	95.25	91.88	94.46
Idiap EdgeFace-XXS-Q (Kolf et al., 2023a)	8.0	1.73	99.50	93.77	94.68	89.65	93.11
Mobilenet_V2_visteam (Kolf et al., 2023a)	32.0	3.38	98.58	83.65	88.63	82.90	89.39
SAM-MFaceNet eHWS V1 (Kolf et al., 2023a)	32.0	4.40	99.55	95.57	95.10	91.35	95.01
SAM-MFaceNet eHWS V2 (Kolf et al., 2023a)	32.0	4.40	99.65	95.72	94.90	91.28	94.73
SQ-HH (Kolf et al., 2023a)	32.0	4.55	98.07	84.28	87.17	84.13	91.60
ShuffleNetv2 × 0.5 (Kolf et al., 2023a)	32.0	0.77	97.72	80.33	86.00	83.48	87.76
ShuffleNetv2 × 1.5 (Kolf et al., 2023a)	32.0	7.90	98.95	88.78	91.08	89.73	93.44
MobileFaceNet [Our]	8.00	1.00	99.27	94.33	94.18	85.82	89.57
	6.00	0.75	99.20	93.70	93.82	85.77	89.14
	4.64	0.58	98.27	89.17	92.03	77.90	81.36
	3.97	0.50	98.05	89.67	91.50	79.38	82.27
	3.28	0.41	97.50	85.97	89.12	73.40	77.40
	2.97	0.37	97.32	86.37	89.12	75.72	78.56
	2.64	0.33	97.30	85.35	89.98	73.78	78.31
	2.48	0.31	97.75	85.70	89.52	72.05	77.70
	2.32	0.29	97.35	86.42	90.18	74.33	78.56
	2.24	0.28	96.88	84.90	88.38	73.02	77.64
	2.09	0.26	97.12	85.27	88.60	71.88	76.84
	2.00	0.25	96.88	86.17	89.12	73.42	76.96

Table 3
Comparison of our Mixed-Precision Quantized MobileFaceNet Model to Fixed-Quantization QMobileFaceNet Model proposed by Bunda et al. Bunda et al. (2022) on LFW (Huang et al., 2007) and AgeDB30 (Moschoglou et al., 2017) Benchmarks. The Comparison is Limited Based on Results Provided in Bunda et al. (2022). For Mixed-Precision Quantization, the Authors Provide no Results. It is Evident That Our proposed Mixed-Quantization Approach is Outperforming the Fixed-Precision Quantization Approach utilized in Bunda et al. (2022).

Model	Avg. bits	Size (MB)	LFW	AgeDB30
QMobileFaceNet (Bunda et al., 2022)	32	4.01	98.85	90.38
	8.00	1.00	98.68	88.79
	4.00	0.50	98.63	88.20
	2.00	0.25	93.45	75.62
MobileFaceNet [Our]	32	4.01	99.47	95.62
	8.00	1.00	99.27	94.33
	6.00	0.75	99.20	93.70
	4.64	0.58	98.27	89.17
	3.97	0.50	98.05	89.67
	3.28	0.41	97.50	85.97
	2.97	0.37	97.32	86.37
	2.64	0.33	97.30	85.35
	2.48	0.31	97.75	85.70
	2.32	0.29	97.35	86.42
	2.24	0.28	96.88	84.90
	2.09	0.26	97.12	85.27
	2.00	0.25	96.88	86.17

competitive results to the fixed-precision quantization method. At bit width below 6 bit, the mixed-precision quantization method achieve competitive results to the fixed-precision quantization when considering the recognition performance together with memory requirements of the model. It is noteworthy that the mixed-precision models converge even at low bit widths, while fixed-precision quantization does not converge at low bit widths (Boutros et al., 2022b).

As expected, quantizing the model using mixed-precision to average bit width of 4-bit or lower, the verification error increases in comparison to 8 bit mixed-quantized or FP32 models. For example, the verification accuracy of the FP32 ResNet18 model on CFP-FP benchmark is 94.47%. This verification accuracy in an extremely low bit scenario (2 bit) decreases to 89.46%. However, the required memory footprint, in this case, is significantly reduced from 96.10 MB to 6.01 MB. Similar drops in performance can be seen on other benchmarks, however, it comes with a great reduction in the required memory footprint. The same is true for ResNet50, when comparing the 2 bit quantized model to the FP32 ResNet50 model. However, the reduction in the model size drops by 16 times.

A similar observation can be made when the quantized model is MobileFaceNet. The large error of the 2 bit quantized model compared to the 2 bit ResNet18 and ResNet50 models may be attributed to the fact that MobileFaceNet has fewer parameters. The 2 bit quantization reduces the representativeness of the individual weights and layers, which could be compensated for with more parameters, as with ResNet18 and ResNet50. However, MobileFaceNet appears to lack the necessary amount of parameters to make up for the loss of information.

When comparing our mixed-precision quantized MobileFaceNet to recent efficient FR models (Table 2) evaluated within the Efficient FR Competition held at IJCB 2023 (Kolf et al., 2023a), it is evident that the low bit quantized models achieve a significantly smaller memory footprint when compared to participants of the competition while maintaining competitive verification accuracies on the evaluated benchmarks. The 6 bit quantized MobileFaceNet is outperforming the ShuffleNetv2x0.5 model, which has a comparable memory size to the MobileFaceNet. This further highlights the success of mixed-precision quantization to small-bit widths.

In Bunda et al. (2022), the authors utilized a fixed-precision quantization approach to quantize MobileFaceNet model (QMobileFaceNet) to 8, 4, and 2 bit, achieving stable convergence even on low bit width. When comparing our proposed mixed-precision quantization approach to QMobileFaceNet, as shown in Table 3, it is evident that the proposed mixed-quantization approach maintains competitive FR performance even when quantized to low bit width, e.g. 2 bit, in comparison to QMobileFaceNet, whichs performance degrades when quantized to 2 bit.

Periocular Recognition: The results for verification (Table 4) and identification (Table 5) scenarios show that the 8 bit mixed-precision quantized models achieve a lower error rate when compared to their 8 bit fixed-precision quantized and FP32 counterparts. In the identification scenario, the 8 bit mixed-precision quantized models achieved

Table 4
Comparison of Fixed and Mixed-Precision Quantized ResNet18, ResnNet50 and MobileFaceNet Models at Different Bit Levels on the Verification Protocol of the UFPR Dataset. The Metrics are Given in Percent. Models Quantized To 4 Bit Using Fixed-Precision Quantization are Significantly Loosing in Performance when Compared to the FP32 models. Unlike With Fixed-Precision, Applying Mixed-Precision Quantization Results in Quantized Models with Low Bit Widths That Maintain or Outperform the Performance of the Respective FP32 Models. Furthermore, the Mixed-Precision Quantized Models Outperform the Fixed-Precision Quantized Models at Similar Bit Widths.

Model	Precision	Avg. bits	Size (MB)	Verification (1:1)						
				AUC (↑)	EER (↓)	FNMR@ FMR=10^{-2} (↓)	FNMR@ FMR=10^{-3} (↓)	FNMR@ FMR=10^{-4} (↓)	FNMR@ FMR=10^{-5} (↓)	FNMR@ FMR=10^{-6} (↓)
ResNet18	Fixed	32.00	250.24	**98.51**	**5.76**	**13.22**	**25.45**	**38.86**	**52.38**	**65.92**
		8.00	62.56	98.40	5.99	13.83	26.10	40.02	55.15	72.33
		6.00	46.92	98.34	6.16	14.23	26.94	40.88	55.01	74.42
		4.00	31.28	75.69	29.18	95.05	98.90	99.79	99.99	100.00
	Mixed	8.00	62.56	98.70	5.28	11.61	22.85	35.73	50.43	66.50
		6.00	46.92	99.47	3.01	5.06	11.27	19.72	32.02	50.86
		4.58	35.82	**99.60**	**2.63**	**4.21**	**9.54**	**17.16**	**29.74**	52.51
		3.64	28.46	99.47	3.05	5.32	11.79	21.10	33.77	49.80
		3.06	23.93	99.39	3.31	5.92	12.96	22.57	35.18	50.90
		2.69	21.04	99.37	3.41	6.15	13.40	23.03	36.42	52.40
		2.45	19.16	99.36	3.43	6.24	13.42	23.25	35.49	52.33
		2.29	17.91	99.43	3.19	5.58	12.36	21.15	34.36	**47.83**
		2.19	17.13	99.41	3.22	5.58	12.18	21.42	36.04	54.37
		2.12	16.58	99.42	3.17	5.56	11.99	21.24	35.18	53.82
		2.04	15.95	99.42	3.23	5.67	12.39	21.74	33.52	49.76
		2.00	15.64	99.38	3.26	5.71	12.52	22.11	34.63	53.06
ResNet50	Fixed	32.00	328.50	**98.47**	**5.88**	**13.21**	**25.19**	**38.26**	**51.80**	**70.16**
		8.00	82.13	98.39	6.00	13.66	25.81	39.26	53.84	72.82
		6.00	61.59	98.29	6.28	14.47	27.09	40.70	56.20	73.92
		4.00	41.06	66.01	37.78	95.62	99.13	99.78	100.00	100.00
	Mixed	8.00	82.13	98.80	5.05	10.86	21.48	34.32	47.98	62.09
		6.00	61.59	99.54	2.76	4.50	10.11	18.24	30.05	47.51
		4.62	47.43	**99.61**	**2.59**	**4.10**	**9.23**	**16.79**	**28.72**	**45.88**
		3.72	38.19	99.55	2.79	4.63	10.27	18.07	30.84	51.12
		3.14	32.23	99.58	2.69	4.34	9.67	17.15	29.25	48.41
		2.74	28.13	99.59	2.69	4.32	9.53	17.42	31.15	48.08
		2.48	25.46	99.56	2.74	4.52	9.99	17.64	29.65	49.81
		2.30	23.61	99.57	2.74	4.44	10.08	18.16	32.02	48.49
		2.19	22.48	99.54	2.85	4.65	10.23	18.07	31.80	49.42
		2.12	21.76	99.56	2.76	4.51	10.03	18.13	30.55	50.17
		2.05	21.04	99.53	2.83	4.66	10.43	18.78	31.95	48.68
		2.00	20.53	99.56	2.73	4.38	9.96	18.29	31.92	51.49
MobileFaceNet	Fixed	32.00	5.10	**99.23**	**3.86**	**7.52**	**15.87**	**26.21**	**38.66**	**53.23**
		8.00	1.28	99.18	4.02	7.87	16.74	27.50	40.10	53.45
		6.00	0.96	99.11	4.22	8.72	18.13	29.34	40.96	54.39
		4.00	0.64	55.01	50.50	99.99	99.99	99.99	99.99	100.00
	Mixed	8.00	1.28	99.29	3.69	6.97	14.88	25.01	36.93	48.51
		6.00	0.96	**99.55**	**2.80**	**4.75**	**10.95**	**19.68**	**30.08**	47.17
		4.55	0.73	99.54	2.90	5.04	11.50	20.51	32.57	51.58
		3.61	0.58	99.54	2.88	4.99	11.49	20.56	32.87	48.21
		3.01	0.48	99.53	2.90	4.99	11.56	20.67	31.58	49.50
		2.63	0.42	99.51	2.96	5.17	11.97	21.06	32.48	48.66
		2.40	0.38	99.50	3.04	5.26	12.02	20.90	33.22	50.79
		2.25	0.36	99.48	3.09	5.46	12.58	21.74	33.09	47.69
		2.16	0.34	99.49	3.08	5.46	12.11	21.00	30.99	**46.85**
		2.10	0.33	99.49	3.07	5.49	12.54	22.09	34.07	48.91
		2.04	0.33	99.45	3.18	5.75	12.92	22.02	32.78	49.27
		2.00	0.32	99.44	3.20	5.74	12.98	22.48	35.33	51.05

a higher (ResNet18, ResNet50) or equivalent (MobileFaceNet) rank 1 and rank 5 scores compared to the FP32 models.

Mixed-precision quantized models with 6 bit precision achieved similar or even lower error rates compared to the FP32 models. E.g., the verification EER of FP32 ResNet18, 5.76%, drops to 3.01% when the model is quantized to 6 bit using mixed-precision.

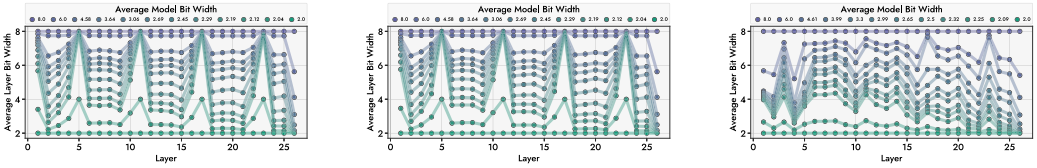
The 2 bit quantized models achieved lower EER than the respective FP32 model, both in the identification and verification scenarios. In low bit widths, the mixed-precision quantized models outperform the respective FP32, 8, and 6 bit fixed-precision quantized models.

A 4 bit fixed-precision quantization does not achieve a stable convergence. It is not able to achieve a competitive performance similar to the FP32 model. However, if mixed-precision quantization to 4 bit or less is used, an error rate similar to that of the FP32 model is obtained.

Comparing our mixed-precision quantized models to the PR identification (Table 7) and verification (Table 6) models published in Zanlorensi et al. (2022), it can be clearly noticed that our models are superior in terms of both the identification and verification metrics, as well as model size. For instance, the pairwise approach (Zanlorensi et al., 2022) is the smallest of the proposed methods (20MB model size) and achieved 98.62% verification accuracy on the open world set, while our 2 bit mixed-precision quantized MobileFaceNet (Chen et al., 2018) model (0.32MB model size) achieved 99.44% verification accuracy. Similar observations can be made for mixed-precision quantized ResNet18 (He et al., 2016b) and ResNet50 (He et al., 2016b). Both models can keep their identification and verification performance when quantized to low bit width, surpassing proposed methods in Zanlorensi et al. (2022). Moreover, the mixed-precision techniques reduce

Table 5
Comparison of Fixed and Mixed-Precision Quantized ResNet18, ResnNet50 and MobileFaceNet Models at Different Bit Levels on the Identification Protocol of the UFPR Dataset. The Metrics are Given in Percent. Models Quantized To 4 Bit Using Fixed-Precision Quantization are Significantly Loosing in Performance when Compared to the FP32 models. Unlike With Fixed-Precision, Applying Mixed-Precision Quantization Results in Quantized Models with Low Bit Widths That Maintain or Outperform the Performance of the Respective FP32 Models. Furthermore, the Mixed-Precision Quantized Models Outperform the Fixed-Precision Quantized Models at Similar Bit Widths.

Model	Precision	Avg. bits	Size (MB)	Identification (1:N)		Verification(1:1)						
				Rank 1 (↑)	Rank 5 (↑)	AUC (↑)	EER (↓)	FNMR@ FMR=10^{-2} (↓)	FNMR@ FMR=10^{-3} (↓)	FNMR@ FMR=10^{-4}()	FNMR@ FMR=10^{-5} (↓)	FNMR@ FMR=10^{-6} (↓)
ResNet18	Fixed	32.00	250.24	**99.61**	**99.88**	**99.77**	**1.75**	**2.19**	**4.72**	**8.65**	**14.32**	**27.87**
		8.00	62.56	99.54	99.85	99.76	**1.75**	2.22	4.78	8.88	15.96	35.72
		6.00	46.92	99.48	99.79	99.73	1.92	2.57	5.35	9.69	16.74	42.11
		4.00	31.28	50.96	62.01	86.18	19.17	92.08	98.73	99.67	99.93	100.00
	Mixed	8.00	62.56	99.66	99.90	99.83	1.43	1.67	3.76	6.98	12.57	26.97
		6.00	46.92	99.90	99.97	99.93	0.67	0.57	1.27	2.59	4.64	10.97
		4.58	35.82	**99.94**	**99.98**	**99.95**	**0.64**	**0.53**	**1.19**	**2.28**	**3.89**	**9.39**
		3.64	28.46	99.90	99.96	99.93	0.72	0.64	1.48	2.84	4.90	14.21
		3.06	23.93	99.90	**99.98**	99.93	0.79	0.72	1.64	3.24	5.74	14.73
		2.69	21.04	99.90	99.97	99.92	0.84	0.79	1.80	3.49	6.30	16.17
		2.45	19.16	99.92	**99.98**	99.93	0.73	0.64	1.54	3.05	5.37	12.13
		2.29	17.91	99.87	99.96	99.93	0.73	0.63	1.50	2.87	5.16	12.16
		2.19	17.13	99.92	99.97	99.93	0.72	0.64	1.52	2.93	5.31	11.81
		2.12	16.58	99.91	99.96	99.93	0.76	0.66	1.53	2.99	5.25	12.04
		2.04	15.95	99.92	99.96	99.93	0.73	0.61	1.50	2.94	5.13	14.61
		2.00	15.64	99.92	99.96	99.93	0.72	0.63	1.52	2.93	5.29	11.84
ResNet50	Fixed	32.00	328.50	**99.54**	99.81	**99.75**	1.76	2.25	**4.83**	**8.74**	**15.45**	**32.95**
		8.00	82.13	99.49	**99.82**	**99.75**	**1.74**	**2.24**	4.84	8.93	16.63	39.92
		6.00	61.59	99.34	99.71	99.73	1.95	2.59	5.62	10.13	17.87	51.17
		4.00	41.06	33.24	46.29	80.55	26.75	91.82	98.28	99.51	100.00	100.00
	Mixed	8.00	82.13	99.75	99.90	99.84	1.32	1.48	3.22	6.06	10.54	24.51
		6.00	61.59	99.91	99.97	99.95	0.58	0.47	1.08	2.15	3.85	9.34
		4.62	47.43	99.94	99.97	99.95	**0.52**	**0.40**	**0.96**	**1.92**	3.49	7.96
		3.72	38.19	99.92	99.97	99.95	0.56	0.43	1.05	2.13	3.72	9.80
		3.14	32.23	99.93	**99.98**	**99.96**	0.53	0.43	0.99	1.96	**3.44**	8.79
		2.74	28.13	**99.96**	99.97	99.95	0.57	0.45	0.99	1.95	3.52	**7.40**
		2.48	25.46	99.93	**99.98**	99.95	0.57	0.43	1.02	2.04	3.67	8.69
		2.30	23.61	99.94	**99.98**	99.95	0.59	0.48	1.12	2.21	3.97	10.45
		2.19	22.48	99.92	**99.98**	99.95	0.61	0.49	1.15	2.25	4.12	9.27
		2.12	21.76	99.93	**99.98**	99.95	0.58	0.47	1.09	2.18	4.01	11.86
		2.05	21.04	99.92	**99.98**	99.95	0.56	0.42	1.05	2.17	3.96	11.48
		2.00	20.53	99.90	99.97	99.94	0.63	0.51	1.19	2.35	4.14	11.30
MobileFaceNet	Fixed	32.00	5.10	**99.82**	**99.92**	**99.86**	1.26	1.41	2.98	5.27	8.69	14.63
		8.00	1.28	99.79	99.90	**99.86**	**1.17**	**1.25**	**2.73**	**4.97**	**8.14**	**13.62**
		6.00	0.96	99.80	99.90	**99.86**	1.22	1.34	2.89	5.21	8.81	15.85
		4.00	0.64	5.53	11.19	62.28	42.50	99.96	99.96	99.96	99.98	100.00
	Mixed	8.00	1.28	99.80	99.91	99.88	1.08	1.12	2.47	4.53	7.57	12.68
		6.00	0.96	99.85	99.94	**99.92**	**0.82**	0.76	**1.69**	**3.33**	**5.49**	**9.87**
		4.55	0.73	99.86	99.93	**99.92**	0.83	**0.74**	1.78	3.49	5.86	10.22
		3.61	0.58	**99.90**	**99.95**	**99.92**	0.86	0.81	1.87	3.64	6.27	10.60
		3.01	0.48	99.88	**99.95**	**99.92**	0.84	0.77	1.93	3.72	6.55	11.45
		2.63	0.42	99.87	99.94	**99.92**	0.88	0.83	2.08	4.06	6.88	11.37
		2.40	0.38	99.85	99.93	99.91	0.94	0.91	2.19	4.08	6.64	11.15
		2.25	0.36	99.89	99.94	99.91	0.94	0.91	2.17	4.13	7.09	12.13
		2.16	0.34	99.88	**99.95**	99.91	0.94	0.90	2.18	4.13	7.10	12.10
		2.10	0.33	99.88	99.94	**99.92**	0.95	0.92	2.25	4.40	7.46	13.05
		2.04	0.33	99.86	99.93	99.89	1.00	1.00	2.27	4.44	7.50	12.10
		2.00	0.32	99.87	99.94	99.91	0.98	0.98	2.32	4.38	7.46	12.47



(a) ResNet18, Periocular Verification. (b) ResNet18, Periocular Identification. (c) ResNet18, Face Verification.

Fig. 2. The average bit width per layer for the mixed-precision quantized ResNet18 backbone is shown for PR (UFPR verification (a), UFPR identification (b)) and FR (c). Only convolutional and fully-connected layers are plotted. In each iteration of the mixed-precision quantization scheme, 50% of weights > 2 bit have their bit width halfed.

the computational overhead, providing an additional benefit over the models compared to.

Analysis of Bit Width Reduction per Layer: The average bit width per layer of the convolutional and fully-connected layers of the ResNet18 backbone, which are quantized using mixed-precision quantization, is shown in Fig. 2. It is clear that there is no significant difference between the quantized models trained on the verification scenario of the UFPR dataset (Fig. 2(a)) and on the identification scenario (Fig. 2(b)).

It is also noticeable that the bit width of certain layers (e.g. layer 5) is on average much higher than the bit width of other layers. This indicates that the importance of these layers is much greater than the importance of layers whose bit width is reduced in nearly every iteration (e.g. layer 2). These layers are the first convolutional layers of each ResNet18 block.

When comparing the average layer bit width of PR models to the ResNet18 FR model (Fig. 2(c)), one can notice that the quantization and bit width behavior of the models differ between the FR and PR tasks. In FR, the importance of layers, based on the average layer bit width, is more evenly distributed throughout the network for average bit widths of the model above 4 bits. Convolutional layers (e.g. layers 6–9) have a similar bit width behavior over each quantization iteration as downsampling layers, which is different in the PR case. When the average bits of the ResNet18 backbone are reduced further and below an average of 4 bit, earlier layers in the FR model have a higher average bit width than deeper layers inside the network. This suggests that extracting features in the early stage requires a higher

Table 6
Comparison between Mixed-Precision Quantized ResNet18, ResnNet50 and MobileFaceNet Models and Recent Models Reported in Zanlorensi et al. (2022) at Different Bit Levels on the Verification Protocol of the UFPR Dataset. The Metrics are Given in Percent. Our Models Quantized to Low Bit Width using Mixed-Precision Quantization are Either Outperform the Recent Models or Achieve Competitive Results on This Benchmark. Especially MobileFaceNet Outperforms Other Models in Terms Of Performance and Model Size.

Model	Avg. bits	Size (MB)	Verification (1:1)	
			AUC (%)	EER (%)
VGG16 (Simonyan and Zisserman, 2015; Zanlorensi et al., 2022)	32.0	1088	97.38	8.52
VGG16-Face (Parkhi et al., 2015; Zanlorensi et al., 2022)	32.0	1088	97.70	7.78
InceptionResNet (Szegedy et al., 2017; Zanlorensi et al., 2022)	32.0	445	99.10	4.61
ResNet50V2 (He et al., 2016b; Zanlorensi et al., 2022)	32.0	400	98.73	5.69
ResNet50 (He et al., 2016a; Zanlorensi et al., 2022)	32.0	198	98.60	5.98
ResNet50-Face (Cao et al., 2018; Zanlorensi et al., 2022)	32.0	198	99.18	4.38
Xception (Chollet, 2017; Zanlorensi et al., 2022)	32.0	176	98.93	5.23
DenseNet121 (Huang et al., 2017; Zanlorensi et al., 2022)	32.0	64	99.51	3.39
Multi-task (Zanlorensi et al., 2022)	32.0	37	99.67	2.81
MobileNetV2 (Sandler et al., 2018; Zanlorensi et al., 2022)	32.0	26	99.56	3.17
Siamese (Zanlorensi et al., 2022)	32.0	21	97.27	8.10
Pairwise (Zanlorensi et al., 2022)	32.0	20	98.62	5.77
ResNet18 Mixed-Precision	8.00	62.56	98.70	5.28
	6.00	46.92	99.47	3.01
	3.06	23.93	99.39	3.31
	2.19	17.13	99.41	3.22
	2.00	15.64	99.38	3.26
ResNet50 Mixed-Precision	8.00	82.13	98.80	5.05
	6.00	61.59	99.54	2.76
	3.14	32.23	99.58	2.69
	2.19	22.48	99.54	2.85
	2.00	20.53	99.56	2.73
MobileFaceNet Mixed-Precision	8.00	1.28	99.29	3.69
	6.00	0.96	99.55	2.80
	3.01	0.48	99.53	2.90
	2.16	0.34	99.49	3.08
	2.00	0.32	99.44	3.20

representativeness than feature extraction in later stages. This behavior is not as pronounced in the PR scenario and might be attributed to the circumstance that mixed-precision quantization of PR models achieve better performance than FR models when compared to the respective FP32 models.

This further highlights the advantages of mixed-precision quantization, as the early feature extraction layers can use a larger bit width and thus extract more meaningful features that can be used by layers quantized to lower bits deeper inside the network.

5.2. Fixed vs. mixed-precision quantization

Fig. 3 illustrates the performance difference between fixed-precision and mixed-precision quantization for different bit widths on the respective FR and PR test benchmarks. As demonstrated in the verification scenario for FR (Fig. 3(a)) as well as in the identification scenario for PR (Fig. 3(b)), and verification scenario for PR (Fig. 3(c)), mixed-precision quantization outperforms fixed-precision quantization in low bit widths. Using fixed precision quantization, the model accuracy is significantly dropped when the model is quantized to a bit width lower

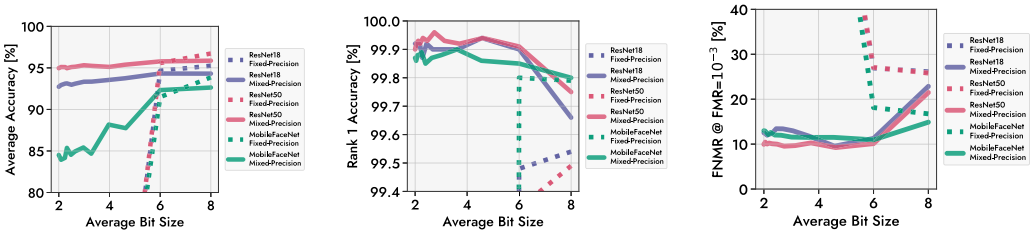
Table 7
Comparison between Mixed-Precision Quantized ResNet18, ResnNet50 and MobileFaceNet Models and Recent Models Reported in Zanlorensi et al. (2022) at Different Bit Levels on the Identification Protocol of the UFPR Dataset. The Metrics are Given in Percent. Our Models Quantized to Low Bit Width using Mixed-Precision Quantization are Either Outperform the Recent Models or Achieve Competitive Results on This Benchmark. Especially MobileFaceNet Outperforms Other Models in Terms Of Performance and Model Size.

Model	Avg. bits	Size (MB)	Identification (1:N)		Verification (1:1)	
			Rank 1 (%)	Rank 5 (%)	AUC (%)	EER (%)
VGG16 (Simonyan and Zisserman, 2015; Zanlorensi et al., 2022)	32.0	1088	50.56	68.73	99.41	3.59
VGG16-Face (Parkhi et al., 2015; Zanlorensi et al., 2022)	32.0	1088	56.29	73.84	99.43	3.44
InceptionResNet (Szegedy et al., 2017; Zanlorensi et al., 2022)	32.0	445	65.16	81.53	99.78	1.85
ResNet50V2 (He et al., 2016b; Zanlorensi et al., 2022)	32.0	400	63.18	77.79	99.74	2.24
ResNet50 (He et al., 2016a; Zanlorensi et al., 2022)	32.0	198	71.06	85.22	99.89	1.41
ResNet50-Face (Cao et al., 2018; Zanlorensi et al., 2022)	32.0	198	73.76	86.86	99.83	1.74
Xception (Chollet, 2017; Zanlorensi et al., 2022)	32.0	176	57.43	75.88	99.77	2.19
DenseNet121(Huang et al., 2017; Zanlorensi et al., 2022)	32.0	64	75.54	88.53	99.93	1.11
Multi-task (Zanlorensi et al., 2022)	32.0	37	84.32	94.55	99.96	0.81
MobileNetV2 (Sandler et al., 2018; Zanlorensi et al., 2022)	32.0	26	77.98	90.19	99.93	1.13
Siamese (Zanlorensi et al., 2022)	32.0	21	–	–	98.94	4.86
Pairwise (Zanlorensi et al., 2022)	32.0	20	–	–	99.44	3.06
ResNet18 Mixed-Precision	8.00	62.56	99.66	99.90	99.83	1.43
	6.00	46.92	99.90	99.97	99.93	0.67
	3.06	23.77	99.90	99.98	99.93	0.79
	2.19	17.05	99.92	99.97	99.93	0.72
	2.00	15.64	99.92	99.96	99.93	0.72
ResNet50 Mixed-Precision	8.00	82.13	99.75	99.90	99.84	1.32
	6.00	61.59	99.91	99.97	99.95	0.58
	3.14	32.13	99.93	99.98	99.96	0.53
	2.19	22.48	99.92	99.98	99.95	0.61
	2.00	20.53	99.90	99.97	99.94	0.63
MobileFaceNet Mixed-Precision	8.00	1.28	99.80	99.91	99.88	1.08
	6.00	0.96	99.85	99.94	99.92	0.82
	3.01	0.48	99.88	99.95	99.92	0.84
	2.16	0.34	99.88	99.95	99.91	0.94
	2.00	0.32	99.87	99.94	99.91	0.98

than 6. On the other hand, the mixed precision quantization was able to maintain, to a very large degree, the model accuracy, even in extremely low bit width of 2 bit. The advantage of mixed-precision quantization is that it allows a much better trade-off between model performance and model size, which has not been used in FR and PR before. This offers a significant step for the development of efficient FR and PR models.

6. Conclusion

Motivated by the need for lightweight biometric recognition models, this work extends to the direct idea of model quantization. Model quantization aims at representing the model parameters at a lower precision level than the conventional full-precision FP32, thus requiring a lower operational memory footprint. However, it was proven that quantizing such models has their limits in terms of the precision level, if substantial performance needs to be maintained. Knowing that the high precision of only part of the model parameters is needed to maintain the performance, while it is sufficient to have very low precision for other parameters, we aim at pushing the limits of the model compression while maintaining recognition performance by utilizing mixed-precision quantization. Here, we applied quantization on different levels of precision over different network parts, based on their effect on the model operation. By applying this concept on multiple backbones on both FR and PR, our results showed that the models can be compressed up to 16x (i.e. precision level of 2 bit) while maintaining



(a) Average Face Verification Accuracy over all Benchmarks (Higher is Better).

(b) UFPR Identification Performance (Higher is Better).

(c) UFPR Verification Performance (Lower is Better).

Fig. 3. Test performance of fixed-precision and mixed-precision quantized backbones ResNet18, ResNet50 and MobileFaceNet. The backbones are evaluated for FR with the average accuracy over all utilized benchmarks (a) and for PR on the UFPR dataset. For the identification scenario (b) the rank 1 accuracy is given, for verification scenario (c) the FNMR at FMR=10^{-3} is shown. It is evident that at low bit widths the fine-tuning process of fixed-precision quantized models is not converging. In contrast, the mixed-precision quantized models are able to maintain their performance to a large degree.

substantial performance and thus successfully overcoming the limits of fixed-precision biometric model quantization.

CRediT authorship contribution statement

Jan Niklas Kolf: Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Conceptualization. **Jurek Elliesen:** Software. **Naser Damer:** Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Fadi Boutros:** Writing – original draft, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

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

References

Ahuja, K., Bose, A., Nagar, S., Dey, K., Barbhuiya, F., 2016. ISURE: User authentication in mobile devices using ocular biometrics in visible spectrum. In: 2016 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 335–339.

Alansari, M., Hay, O.A., Javed, S., Shoufan, A., Zweiri, Y., Werghi, N., 2023. GhostFaceNets: Lightweight face recognition model from cheap operations. IEEE Access 11, 35429–35446. http://dx.doi.org/10.1109/ACCESS.2023.3266068.

Almadan, A., Rattani, A., 2021. Compact CNN models for on-device ocular-based user recognition in mobile devices. In: 2021 IEEE Symposium Series on Computational Intelligence. SSCI, IEEE, pp. 1–7.

Alonso-Fernandez, F., Bigun, J., 2014. Best regions for periocular recognition with nir and visible images. In: 2014 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 4987–4991.

Alonso-Fernandez, F., Raja, K.B., Raghavendra, R., Busch, C., Bigun, J., Vera-Rodriguez, R., Fierrez, J., 2022. Cross-sensor periocular biometrics in a global pandemic: Comparative benchmark and novel multialgorithmic approach. Inf. Fusion 83, 110–130.

Anon, 2022. IEEE International Joint Conference on Biometrics. IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022, IEEE, ISBN: 978-1-6654-6394-2, http://dx.doi.org/10.1109/IJCB54206.2022.

Banner, R., Nahshan, Y., Soudry, D., 2019. Post training 4-bit quantization of convolutional networks for rapid-deployment. Adv. Neural Inf. Process. Syst. 32.

Bharadwaj, S., Bhatt, H.S., Vatsa, M., Singh, R., 2010. Periocular biometrics: When iris recognition fails. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems. BTAS, IEEE, pp. 1–6.

Boutros, F., 2022. Efficient and High Performing Biometrics: Towards Enabling Recognition in Embedded Domains (Ph.D. thesis). TU Darmstadt, Germany, http://dx.doi.org/10.26083/TUPRINTS-00021571, URL http://tuprints.ulb.tu-darmstadt.de/21571/.

Boutros, F., Damer, N., Fang, M., Kirchbuchner, F., Kuijper, A., 2021a. Mixfacenets: Extremely efficient face recognition networks. In: 2021 IEEE International Joint Conference on Biometrics. IJCB, IEEE, pp. 1–8.

Boutros, F., Damer, N., Fang, M., Raja, K., Kirchbuchner, F., Kuijper, A., 2020. Compact models for periocular verification through knowledge distillation. In: 2020 International Conference of the Biometrics Special Interest Group. BIOSIG, IEEE, pp. 1–5.

Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A., 2022a. Elasticface: Elastic margin loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022. Computer Vision Foundation / IEEE.

Boutros, F., Damer, N., Kolf, J.N., Raja, K.B., Kirchbuchner, F., Ramachandra, R., Kuijper, A., Fang, P., Zhang, C., Wang, F., Montero, D., Aginako, N., Sierra, B., Nieto, M., Erakin, M.E., Demir, U., Ekenel, H.K., Kataoka, A., Ichikawa, K., Kubo, S., Zhang, J., He, M., Han, D., Shan, S., Grm, K., Struc, V., Seneviratne, S., Kasthuriarachchi, N., Rasnayaka, S., Neto, P.C., Sequeira, A.F., Pinto, J.R., Saffari, M., Cardoso, J.S., 2021b. MFR 2021: Masked face recognition competition. In: International IEEE Joint Conference on Biometrics. IJCB 2021, Shenzhen, China, August 4-7, 2021, IEEE, pp. 1–10. http://dx.doi.org/10.1109/IJCB52358.2021.9484337.

Boutros, F., Damer, N., Kuijper, A., 2022b. QuantFace: Towards lightweight face recognition by synthetic data low-bit quantization. In: 26th International Conference on Pattern Recognition. ICPR 2022, Montreal, Quebec, August 21-25, 2021, IEEE.

Boutros, F., Damer, N., Raja, K.B., Kirchbuchner, F., Kuijper, A., 2022c. Template-driven knowledge distillation for compact and accurate periocular biometrics deep-learning models. Sensors 22 (5), 1921. http://dx.doi.org/10.3390/s22051921.

Boutros, F., Kähm, O., Fang, M., Kirchbuchner, F., Damer, N., Kuijper, A., 2022d. Low-resolution iris recognition via knowledge transfer. In: Brömme, A., Damer, N., Gomez-Barrero, M., Raja, K.B., Rathgeb, C., Sequeira, A.F., Todisco, M., Uhl, A. (Eds.), Proceedings of the 21st International Conference of the Biometrics Special Interest Group. BIOSIG 2022, Darmstadt, Germany, September 14-16, 2022, In: LNI, vol. P-329, IEEE / Gesellschaft für Informatik e.V., pp. 293–300. http://dx.doi.org/10.1109/BIOSIG55365.2022.9896959, URL https://dl.gi.de/handle/20.500.12116/39709.

Boutros, F., Siebke, P., Klemt, M., Damer, N., Kirchbuchner, F., Kuijper, A., 2022e. PocketNet: Extreme lightweight face recognition network using neural architecture search and multi-step knowledge distillation. IEEE Access 1. http://dx.doi.org/10.1109/ACCESS.2022.3170561.

Bunda, S., Spreeuwers, L.J., Zeinstra, C.G., 2022. Sub-byte quantization of mobile face recognition convolutional neural networks. In: Brömme, A., Damer, N., Gomez-Barrero, M., Raja, K.B., Rathgeb, C., Sequeira, A.F., Todisco, M., Uhl, A. (Eds.), Proceedings of the 21st International Conference of the Biometrics Special Interest Group. BIOSIG 2022, Darmstadt, Germany, September 14-16, 2022, In: LNI, P-329, IEEE / Gesellschaft für Informatik e.V., pp. 229–236. http://dx.doi.org/10.1109/BIOSIG55365.2022.9897025, URL https://dl.gi.de/handle/20.500.12116/39700.

Cai, Y., Yao, Z., Dong, Z., Gholami, A., Mahoney, M.W., Keutzer, K., 2020. ZeroQ: A novel zero shot quantization framework. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, pp. 13166–13175. http://dx.doi.org/10.1109/CVPR42600.2020.01318.

Caldeira, E., Neto, P.C., Huber, M., Damer, N., Sequeira, A.F., 2024. Model compression techniques in biometrics applications: A survey. http://dx.doi.org/10.48550/ARXIV.2401.10139, CoRR abs/2401.10139. arXiv:2401.10139.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 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.

Chen, S., Liu, Y., Gao, X., Han, Z., 2018. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition. Springer, pp. 428–438.

Cheng, H., Zhang, M., Shi, J.Q., 2023. A survey on deep neural network pruning-taxonomy, comparison, analysis, and recommendations. http://dx.doi.org/10.48550/ARXIV.2308.06767, CoRR abs/2308.06767. arXiv:2308.06767.

Choi, J., Wang, Z., Venkataramani, S., Chuang, P.I., Srinivasan, V., Gopalakrishnan, K., 2018. PACT: parameterized clipping activation for quantized neural networks. CoRR abs/1805.06085. arXiv:1805.06085.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019a. 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.

Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S., 2019b. Lightweight face recognition challenge. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.

Fei, W., Dai, W., Li, C., Zou, J., Xiong, H., 2021. General bitwidth assignment for efficient deep convolutional neural network quantization. IEEE Trans. Neural Netw. Learn. Syst..

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W., Keutzer, K., 2021. A survey of quantization methods for efficient neural network inference. arXiv preprint arXiv:2103.13630.

Gong, J., Shen, H., Zhang, G., Liu, X., Li, S., Jin, G., Maheshwari, N., Fomenko, E., Segal, E., 2018. Highly efficient 8-bit low precision inference of convolutional neural networks with intelcaffe. In: Proceedings of the 1st on Reproducible Quality-Efficient Systems Tournament on Co-Designing Pareto-Efficient Deep Learning. p. 1.

Guo, Y., 2018. A survey on methods and theories of quantized neural networks. arXiv preprint arXiv:1808.04752.

Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J., 2016. MS-celeb-1M: A dataset and benchmark for large-scale face recognition. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part III. In: Lecture Notes in Computer Science, vol. 9907, Springer, pp. 87–102. http://dx.doi.org/10.1007/978-3-319-46487-9_6.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C., 2020. GhostNet: More features from cheap operations. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, pp. 1577–1586. http://dx.doi.org/10.1109/CVPR42600.2020.00165, URL https://openaccess.thecvf.com/content_CVPR_2020/html/Han_GhostNet_More_Features_From_Cheap_Operations_CVPR_2020_paper.html.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer, pp. 630–645.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Tech. Rep. 07-49, University of Massachusetts, Amherst.

Huang, Q., Wang, D., Dong, Z., Gao, Y., Cai, Y., Li, T., Wu, B., Keutzer, K., Wawrzynek, J., 2021. Codenet: Efficient deployment of input-adaptive object detection on embedded fpgas. In: The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. pp. 206–216.

Iandola, F.N., Moskewicz, M.W., Ashraf, K., Han, S., Dally, W.J., Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1mb model size. CoRR abs/1602.07360. arXiv:1602.07360.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D., 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2704–2713.

Jain, A.K., Ross, A., Prabhakar, S., 2004. An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. 14 (1), 4–20.

Jin, Q., Ren, J., Zhuang, R., Hanumante, S., Li, Z., Chen, Z., Wang, Y., Yang, K., Tulyakov, S., 2022. F8Net: Fixed-point 8-bit only multiplication for network quantization. arXiv preprint arXiv:2202.05239.

Kolf, J.N., Boutros, F., Elliesen, J., Theuerkauf, M., Damer, N., Alansari, M., Hay, O.A., Alansari, S., Javed, S., Werghi, N., Grm, K., Struc, V., Alonso-Fernandez, F., Hernandez-Diaz, K., Bigün, J., George, A., Ecabert, C., Shahreza, H.O., Kotwal, K., Marcel, S., Medvedev, I., Jin, B., Nunes, D., Hassanpour, A., Khatiwada, P., Toor, A.A., Yang, B., 2023a. EFaR 2023: Efficient face recognition competition. http://dx.doi.org/10.48550/arXiv.2308.04168, CoRR abs/2308.04168. arXiv:2308.04168.

Kolf, J.N., Boutros, F., Kirchbuchner, F., Damer, N., 2022. Lightweight periocular recognition through low-bit quantization. In: IEEE International Joint Conference on Biometrics. IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022, IEEE, pp. 1–12. http://dx.doi.org/10.1109/IJCB54206.2022.10007980.

Kolf, J.N., Elliesen, J., Boutros, F., Proença, H., Damer, N., 2023b. SyPer: Synthetic periocular data for quantized light-weight recognition in the NIR and visible domains. Image Vis. Comput. 135, 104692. http://dx.doi.org/10.1016/j.imavis.2023.104692.

Krishnamoorthi, R., 2018. Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a Meeting Held December 3-6, 2012, Lake Tahoe, Nevada, United States. pp. 1106–1114, URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Kumari, P., Seeja, K., 2019. Periocular biometrics: A survey. J. King Saud Univ.-Comput. Inf. Sci..

Li, H.-C., Deng, Z.-Y., Chiang, H.-H., 2020. Lightweight and resource-constrained learning network for face recognition with performance optimization. Sensors 20 (21), 6114.

Li, Z., Gong, Y., Zhang, Z., Xue, X., Chen, T., Liang, Y., Yuan, B., Wang, Z., 2023. Accelerable lottery tickets with the mixed-precision quantization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023, IEEE, pp. 4604–4612. http://dx.doi.org/10.1109/CVPRW59228.2023.00485.

Li, X., Wang, F., Hu, Q., Leng, C., 2019. Airface: Lightweight and efficient model for face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.

Liu, J., Qin, H., Wu, Y., Guo, J., Liang, D., Xu, K., 2022. CoupleFace: Relation matters for face recognition distillation. arXiv preprint arXiv:2204.05502.

Liu, B., Zhang, Z., He, P., Wang, Z., Xiao, Y., Ye, R., Zhou, Y., Ku, W., Hui, B., 2024. A survey of lottery ticket hypothesis. http://dx.doi.org/10.48550/ARXIV.2403.04861, CoRR abs/2403.04861. arXiv:2403.04861.

Ma, Y., Jin, T., Zheng, X., Wang, Y., Li, H., Wu, Y., Jiang, G., Zhang, W., Ji, R., 2023. OMPQ: Orthogonal mixed precision quantization. In: Williams, B., Chen, Y., Neville, J. (Eds.), Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence. EAAI 2023, Washington, DC, USA, February 7-14, 2023, AAAI Press, pp. 9029–9037. http://dx.doi.org/10.1609/aaai.v37i7.26084.

Ma, N., Zhang, X., Zheng, H.-T., Sun, J., 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 116–131.

Mansfield, A., 2006. Information technology–biometric performance testing and reporting–part 1: Principles and framework. ISO/IEC 1–55.

Martindez-Diaz, Y., Luevano, L.S., Mendez-Vazquez, H., Nicolas-Diaz, M., Chang, L., Gonzalez-Mendoza, M., 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.

Martínez-Díaz, Y., Méndez-Vázquez, H., Luevano, L.S., Chang, L., Gonzalez-Mendoza, M., 2021. Lightweight low-resolution face recognition for surveillance applications. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 5421–5428.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S., 2017. Agedb: The first manually collected, in-the-wild age database. In: 2017 IEEE CVPRW, CVPR Workshops 2017. Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, pp. 1997–2005. http://dx.doi.org/10.1109/CVPRW.2017.250.

Nie, L., Kumar, A., Zhan, S., 2014. Periocular recognition using unsupervised convolutional RBM feature learning. In: 2014 22nd International Conference on Pattern Recognition. IEEE, pp. 399–404.

Park, U., Ross, A., Jain, A.K., 2009. Periocular biometrics in the visible spectrum: A feasibility study. In: 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems. IEEE, pp. 1–6.

Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition. In: Xie, X., Jones, M.W., Tam, G.K.L. (Eds.), Proceedings of the British Machine Vision Conference 2015. BMVC 2015, Swansea, UK, September 7-10, 2015, BMVA Press, pp. 41.1–41.12. http://dx.doi.org/10.5244/C.29.41.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Proença, H., Neves, J.C., 2017. Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks. IEEE Trans. Inf. Forensics Secur. 13 (4), 888–896.

Raja, K.B., Damer, N., Ramachandra, R., Boutros, F., Busch, C., 2019. Cross-spectral periocular recognition by cascaded spectral image transformation. In: 2019 IEEE International Conference on Imaging Systems and Techniques. IST 2019, Abu Dhabi, United Arab Emirates, December 9-10, 2019, IEEE, pp. 1–7. http://dx.doi.org/10.1109/IST48021.2019.9010520.

Raja, K.B., Raghavendra, R., Busch, C., 2016. Collaborative representation of deep sparse filtered features for robust verification of smartphone periocular images. In: 2016 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 330–334.

Raja, K.B., Raghavendra, R., Stokkenes, M., Busch, C., 2014. Smartphone authentication system using periocular biometrics. In: 2014 International Conference of the Biometrics Special Interest Group. BIOSIG, IEEE, pp. 1–8.

Rakka, M., Fouda, M.E., Khargonekar, P., Kurdahi, F.J., 2022. Mixed-precision neural networks: A survey. http://dx.doi.org/10.48550/arXiv.2208.06064, CoRR abs/2208.06064. arXiv:2208.06064.

Rattani, A., Derakhshani, R., Saripalle, S.K., Gottemukkula, V., 2016. ICIP 2016 competition on mobile ocular biometric recognition. In: 2016 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 320–324.

Reddy, N., Rattani, A., Derakhshani, R., 2018a. Comparison of deep learning models for biometric-based mobile user authentication. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems. BTAS, IEEE, pp. 1–6.

Reddy, N., Rattani, A., Derakhshani, R., 2018b. Ocularnet: deep patch-based ocular biometric recognition. In: 2018 IEEE International Symposium on Technologies for Homeland Security. HST, IEEE, pp. 1–6.

Reddy, N., Rattani, A., Derakhshani, R., 2019. Robust subject-invariant feature learning for ocular biometrics in visible spectrum. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems. BTAS, IEEE, pp. 1–9.

Ross, A., Jillela, R., Smereka, J.M., Boddeti, V.N., Kumar, B.V., Barnard, R., Hu, X., Pauca, P., Plemmons, R., 2012. Matching highly non-ideal ocular images: An information fusion approach. In: 2012 5th IAPR International Conference on Biometrics. ICB, IEEE, pp. 446–453.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520.

Savchenko, A.V., Savchenko, L.V., Makarov, I., 2023. Fast search of face recognition model for a mobile device based on neural architecture comparator. IEEE Access 11, 65977–65990. http://dx.doi.org/10.1109/ACCESS.2023.3290902.

Sengupta, S., Chen, J., Castillo, C.D., Patel, V.M., Chellappa, R., Jacobs, D.W., 2016. Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision. WACV 2016, Lake Placid, NY, USA, March 7-10, 2016, IEEE Computer Society, pp. 1–9. http://dx.doi.org/10.1109/WACV.2016.7477558.

Sharma, R., Ross, A., 2022. Periocular biometrics and its relevance to partially masked faces: A survey. arXiv preprint arXiv:2203.15203.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL http://arxiv.org/abs/1409.1556.

Spolaor, R., Li, Q., Monaro, M., Conti, M., Gamberini, L., Sartori, G., 2016. Biometric authentication methods on smartphones: A survey. PsychNology J. 14 (2).

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence.

Tan, C.-W., Kumar, A., 2012. Human identification from at-a-distance images by simultaneously exploiting iris and periocular features. In: Proceedings of the 21st International Conference on Pattern Recognition. ICPR2012, IEEE, pp. 553–556.

Tan, M., Le, Q.V., 2019. MixConv: Mixed depthwise convolutional kernels. In: British Machine Vision Conference. URL https://api.semanticscholar.org/CorpusID:199064614.

Uzair, M., Mahmood, A., Mian, A., McDonald, C., 2013. Periocular biometric recognition using image sets. In: 2013 IEEE Workshop on Applications of Computer Vision. WACV, IEEE, pp. 246–251.

Van Der Walt, S., Colbert, S.C., Varoquaux, G., 2011. The numpy array: a structure for efficient numerical computation. Comput. Sci. Eng. 13 (2), 22–30.

Wang, M., Deng, W., 2021. Deep face recognition: A survey. Neurocomputing 429, 215–244. http://dx.doi.org/10.1016/j.neucom.2020.10.081.

Wang, K., Liu, Z., Lin, Y., Lin, J., Han, S., 2019. HAQ: Hardware-aware automated quantization with mixed precision. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, pp. 8612–8620.

Weng, O., 2021. Neural network quantization for efficient inference: A survey. arXiv preprint arXiv:2112.06126.

Wu, J., Leng, C., Wang, Y., Hu, Q., Cheng, J., 2016. Quantized convolutional neural networks for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4820–4828.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S., 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (Eds.), International Conference on Machine Learning. ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, In: Proceedings of Machine Learning Research, vol. 202, PMLR, pp. 38087–38099, URL https://proceedings.mlr.press/v202/xiao23c.html.

Xu, J., Cha, M., Heyman, J.L., Venugopalan, S., Abiantun, R., Savvides, M., 2010. Robust local binary pattern feature sets for periocular biometric identification. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems. BTAS, IEEE, pp. 1–8.

Yan, M., Zhao, M., Xu, Z., Zhang, Q., Wang, G., Su, Z., 2019. VarGFaceNet: An efficient variable group convolutional neural network for lightweight face recognition. In: 2019 IEEE/CVF International Conference on Computer Vision Workshops. ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019, IEEE, pp. 2647–2654. http://dx.doi.org/10.1109/ICCVW.2019.00323.

Yao, Z., Dong, Z., Zheng, Z., Gholami, A., Yu, J., Tan, E., Wang, L., Huang, Q., Wang, Y., Mahoney, M., et al., 2021. Hawq-v3: Dyadic neural network quantization. In: International Conference on Machine Learning. PMLR, pp. 11875–11886.

Yu, W., Li, Y., 2019. Fast face recognition model without pruning. In: IEEE Symposium Series on Computational Intelligence. SSCI 2019, Xiamen, China, December 6-9, 2019, IEEE, pp. 2563–2570. http://dx.doi.org/10.1109/SSCI44817.2019.9002922.

Zanlorensi, L.A., Laroca, R., Lucio, D.R., Santos, L.R., Britto, Jr., A.S., Menotti, D., 2020a. UFPR-periocular: a periocular dataset collected by mobile devices in unconstrained scenarios. arXiv preprint arXiv:2011.12427.

Zanlorensi, L.A., Laroca, R., Lucio, D.R., Santos, L.R., Britto, Jr., A.S., Menotti, D., 2022. A new periocular dataset collected by mobile devices in unconstrained scenarios. Sci. Rep. 12 (1), 17989.

Zanlorensi, L.A., Proença, H., Menotti, D., 2020b. Unconstrained periocular recognition: Using generative deep learning frameworks for attribute normalization. In: 2020 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 1361–1365.

Zhang, Q., Li, H., Sun, Z., Tan, T., 2018a. Deep feature fusion for iris and periocular biometrics on mobile devices. IEEE Trans. Inf. Forensics Secur. 13 (11), 2897–2912.

Zhang, Q., Li, J., Yao, M., Song, L., Zhou, H., Li, Z., Meng, W., Zhang, X., Wang, G., 2019. VarGNet: Variable group convolutional neural network for efficient embedded computing. CoRR abs/1907.05653. arXiv:1907.05653.

Zhang, D., Yang, J., Ye, D., Hua, G., 2018b. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 365–382.

Zhang, P., Zhao, F., Liu, P., Li, M., 2022. Efficient lightweight attention network for face recognition. IEEE Access 10, 31740–31750.

Zheng, T., Deng, W., 2018. Cross-pose LFW: A Database for Studying Cross-Pose Face Recognition in Unconstrained Environments. Tech. Rep. 18–01, Beijing University of Posts and Telecommunications.

Zheng, T., Deng, W., Hu, J., 2017. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. CoRR abs/1708.08197. arXiv:1708.08197.

Zhou, S., Ni, Z., Zhou, X., Wen, H., Wu, Y., Zou, Y., 2016. DoReFa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. CoRR abs/1606.06160. arXiv:1606.06160.

Zhuang, B., Shen, C., Tan, M., Liu, L., Reid, I., 2018. Towards effective low-bitwidth convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7920–7928.

Zuras, D., Cowlishaw, M., Aiken, A., Applegate, M., Bailey, D., Bass, S., Bhandarkar, D., Bhat, M., Bindel, D., Boldo, S., et al., 2008. IEEE standard for floating-point arithmetic. IEEE Std 754 (2008), 1–70.