EFaR 2023: Efficient Face Recognition Competition

Jan Niklas Kolf1,2,∗, Fadi Boutros1,∗, Jurek Elliesen1,2,∗, Markus Theuerkauf1,2,∗, Naser Damer1,2,∗
Mohamad Alansari3,+, Oussama Abdul Hay3,+, Sara Alansari3,+, Sajid Javed5,+, Naoufel Werghi3,+, Klemen Grm4,+
Vitomir Struc4,+, Fernando Alonso-Fernandez5,+, Kevin Hernandez Diaz5,+, Josep Bigun5,+, Anjith George6,+
Christophe Ecabert6,+, Hatem Otroshi Shahreza6,7,+, Ketan Kotwal6,+, Sébastien Marcel6,8,+, Iurii Medvedev9,+
Bo Jin9,+, Diogo Nunes9,+, Ahmad Hassanpour10,+, Pankaj Khatiwada10,+, Aafan Ahmad Toor10,+, Bian Yang10,+
1Fraunhofer Institute for Computer Graphics Research IGD, Germany 2TU Darmstadt, Germany
3Department of Electrical and Computer Engineering, Khalifa University, Abu Dhabi, United Arab Emirates
4Laboratory for Machine Intelligence, Faculty of Electrical Engineering, University of Ljubljana, Slovenia 5Halmstad University,
Sweden 6Idiap Research Institute, Martigny, Switzerland 7École Polytechnique Fédérale de Lausanne (EPFL), Lausanne,
Switzerland 8Université de Lausanne (UNIL), Lausanne, Switzerland 9Institute Of Systems and Robotics, University of
Coimbra, Coimbra, Portugal 10eHealth and Welfare Security Group, Department of Information Security and Communication
Technology, Norwegian University of Science and Technology, Norway
∗Competition organizer +Competition participant
Email: jan.niklas.kolf@igd.fraunhofer.de

Abstract

This paper presents the summary of the Efficient Face Recognition Competition (EFaR) held within the 2023 International Joint Conference on Biometrics (IJCB 2023). The competition received 17 submissions from 6 different teams. To drive further development of efficient face recognition models, the submitted solutions are ranked based on a weighted score of the achieved verification accuracies on a diverse set of benchmarks, as well as the deployability given by the number of floating-point operations and model size. The evaluation of submissions is extended to bias, cross-quality, and large-scale recognition benchmarks. Overall, the paper gives an overview of the achieved performance values of the submitted solutions as well as a diverse set of baselines, the methodologies used to achieve lightweight and efficient face recognition solutions, and outlooks on possible techniques that are underrepresented in current solutions.

1. Introduction

Biometrics use behavioral and physical characteristics of a person for recognition [29]. These systems are increasingly used in everyday life. Modern smartphones are equipped with front cameras that enable the use of the face as a biometric modality [48]. High-performing face recognition (FR) systems were specifically enabled with the breakthrough of deep neural networks (DNN) and deep convolutional neural networks (CNN) for computer vision tasks [6, 14, 35, 53]. Consequently, the performance of State-Of-The-Art (SOTA) FR models has been greatly improved, even under challenging conditions [5, 14]. Likewise, SOTA FR models surpass human performance on various benchmarks [53].

However, the increased performance of DNNs is associated with high computational complexity, which makes it difficult to apply the FR models in resource-restricted domains such as embedded devices or smartphones [8, 15, 42]. To utilize DNN on embedded devices, the computational complexity, measured as the number of floating-point operations (FLOPs), and the memory footprint (in MB) need to meet certain criteria. An indication of the targeted memory footprint of a DNN that can be deployed on embedded devices is the ICCV 2019 Workshop challenge on Eye Tracking organized by Facebook [1]. There, the targeted model footprint is set to 1 MB. A targeted upper bound of FLOPs for lightweight FR models is given by the ICCV 2019 Lightweight Face Recognition Challenge [15]. Its lightweight FR model track defined models with at most 1G FLOPs as lightweight models. To develop FR models that meet these criteria and address the challenges, different approaches are used.

One approach is to design lightweight DNN models [4]. A small and efficient network architecture is developed that reduces the computational effort in comparison to larger and more complex DNN. Recent advancements of lightweight DNN are MobileNets [46], SqueezeNet [28], VarGNet [60], MixNets [49], ShuffleNet [38] and GhostNet [23]. For FR, these architectures developed for common computer vision tasks are increasingly used and adapted [11, 41]. One of the earliest works to use lightweight DNN for FR was MobileFaceNet [11] that utilized MobileNetV2 [46], with 1M trainable parameters and 439M FLOPs. SqueezeFaceNet [57], VarGFaceNet [56], MixFaceNets [4], ShuffleFaceNet
The aim of this competition is to evaluate the latest and state-of-the-art approaches for efficient and lightweight face recognition and to motivate the development of novel techniques. Unlike in previous competitions, the result of this competition will be a ranking of the submitted methods not only based on the achieved recognition performance but also with a focus on the size of the models according to the computational complexity or model size. Furthermore, the recent advancements in lightweight face recognition are not covered by this competition.

The IARPA Janus Benchmark–C (IJB-C) [43] and GhostFaceNets [2] adapt SqueezeNet, VarGNet, MixNets and ShuffleNet [38] as their FR model architecture, respectively. The lightweight architectures find a trade-off between computational complexity and accuracy while keeping their computational complexity below 1G FLOPs.

To reduce the memory footprint together with the computational complexity, model quantization is also used [8, 20]. This is achieved by replacing the floating point parameters of the DNN, which are often represented with 32 or 16 bit, with low bit parameter representations, e.g. 8 or 4 bit signed integer [20]. The lower number of bits required per parameter reduces the memory footprint of the model and reduces the computational cost, as integer arithmetic is faster than floating-point [34].

In order to evaluate lightweight and efficient DNN for FR and to promote further developments, various competitions were held. The Masked Face Recognition Competition at IJCB 2021 [7] has taken up the challenge of masked face recognition raised by the COVID-19 pandemic. In order to meet the emerging need for efficient FR models that can handle masked faces, a dedicated dataset was created for evaluation. The submitted solutions are ranked by their recognition performance and by the computational complexity of the proposed solution. A similar goal was conducted by the Masked Face Recognition Challenge at ICCV 2021 [13], with large-scale, private test datasets. An extensive competition dedicated to lightweight face recognition was conducted by Deng et al. [15] that further indicated a great need for efficient FR models. A dedicated lightweight face recognition model track evaluated FR models with up to 1G FLOPs. While the track introduced a limit on FLOPs, the submissions were only evaluated based on their achieved recognition performance with no further focus on the computational complexity or model size. Furthermore, the recent advancements in lightweight face recognition are not covered by this competition.

The aim of this competition is to evaluate the latest and state-of-the-art approaches for efficient and lightweight face recognition and to motivate the development of novel techniques. Unlike in previous competitions, the result of this competition will be a ranking of the submitted methods not only based on the achieved recognition performance but also with a focus on the size of the models according to the number of parameters as well as the computational effort (FLOPs) associated with the inference. The evaluation protocol published for the competition and the metrics used can be utilized in future publications as a benchmark and comparison, so that the comparability of individual methodologies can be further enhanced, being one significant impact of this competition. The competition is organized with two different tracks. The first track includes models with up to 2M parameters, and the second track includes models with 2-5M parameters. The competition received 17 valid submissions from 6 different teams. The first track received 8, and the second track with larger models received 9 submissions, respectively. All submissions are evaluated on common FR benchmarks introduced in Section 2.1 and ranked based on their achieved recognition accuracy, number of parameters, and FLOPs. This paper summarizes the results of the efficient face recognition competition with a detailed presentation of the achieved recognition rates in combination with the model size and the computational complexity.

In the next section we introduce the evaluation datasets, evaluation criteria and the participating teams. Section 3 gives a short overview of the submitted solutions per team. Section 4 covers the achieved results. In Section 5 the by the participants used and unused methodologies for efficient DNN are discusses and a possible outlook for other approaches is given. Section 6 ends the paper with a final general conclusion.

2. Evaluation Setup and Participants

2.1. Evaluation datasets

AgeDB-30 [45]: AgeDB is an in-the-wild dataset for age-invariant face verification evaluation, containing 16,488 images of 568 identities. We report the performance for AgeDB-30 (30 years age gap) as it is the most reported and challenging subset of AgeDB.

LFW [27]: Labeled Faces in the Wild (LFW) is an unconstrained face verification dataset. The LFW contains 13,233 images of 5749 identities collected from the web. The result on LFW is reported as verification accuracy (as defined in [27]) following the “unrestricted with labeled outside data” protocol using the standard 6000 comparison pairs, 3000 genuine and 3000 imposter pairs, defined in [27].

CFP-FP [47]: Celebrities in Frontal-Profile in the Wild (CFP-FP) [47] dataset addresses the comparison between frontal and profile faces. CFP-FP dataset contains 7,000 images across 500 identities, where 10 frontal and 4 profile image per identity. CFP-FP evaluation protocol contains 3500 genuine pairs and 3500 imposter pairs.

CalfW [62]: The Cross-age LFW (CalfW) dataset [62] is based on LFW with a focus on comparison pairs with the age gap, however not as large as AgeDB-30. Age gap distribution of the CalfW is provided in [62]. It contains 3000 genuine comparisons, and the negative pairs are selected of the same gender and race to reduce the effect of attributes.

CPLFW [61]: The Cross-Pose LFW (CPLFW) dataset [61] is based on LFW with a focus on comparison pairs with pose differences. CPLFW contains 3000 genuine comparisons, while the negative pairs are selected of the same gender and race.

IJB-C [43]: The IARPA Janus Benchmark–C (IJB-C)
The evaluation protocol of RFW contains 3000 genuine and
namely Caucasian, Asian, Indian, and African. Each test-

RFW [54]: The Racial Faces in the Wild (RFW) is com-
monly used to evaluate racial bias in face recognition al-

XQLFW [33]: The Cross-Quality LFW (XQLFW) is
derived from the LFW dataset. The XQLFW maximizes
the quality difference, which contains only more realistic
synthetically degraded images when necessary and is used
to investigate the influence of image quality. XQLFW eval-
uation protocol contains 3000 genuine pairs and 3000 im-
postor pairs.

The Cross-Quality LFW (XQLFW) is
derived from the LFW dataset. The XQLFW maximizes
the quality difference, which contains only more realistic
synthetically degraded images when necessary and is used
to investigate the influence of image quality. XQLFW eval-
uation protocol contains 3000 genuine pairs and 3000 im-
postor pairs.

RFW [54]: The Racial Faces in the Wild (RFW) is com-
monly used to evaluate racial bias in face recognition al-

2.3. Submission and Evaluation Process

Each of the teams was asked to submit their solutions
as a Linux console application or Python code with the re-
quired packages listed. The applications should accept as a
parameter a text file with a list of images and an output path.
The image list contains paths to aligned images for which a
template is extracted and saved in the output path. The
competition organizers provide pre-aligned images follow-
ing [14]. If a participant uses a different alignment method,
a dedicated alignment script is submitted by the team. The
script is provided with image paths, bounding boxes, and
landmarks. The bounding boxes and landmarks are detected
using pre-trained Multitask Cascaded Convolutional Net-
works (MTCNN) [59]. The extracted features are used with the
respective protocol of the evaluation dataset. The partici-
ants are free to choose their training data. However, these
databases must be publicly accessible and the authors must
have a license to use these databases (when required by the
data creators). The participants takes full responsibility of
insuring the proper legal use of the data.

2.4. Competition Participants

The competition aimed at attracting teams from various
research institutes with high geographic and activity vari-
ation. The call for participation was shared on the Inter-
national Joint Conference on Biometrics (IJCB 2023) website,
on the dedicated competition website, on mailing lists, and
social media. The call has attracted 7 registered teams. Of
these, 6 teams submitted at least one solution. Each of the
teams is from a different country. Each team was allowed to
submit any number of submissions for both tracks. A total of
17 different submissions were submitted by the teams. For the 2-5 MP track 9 solutions were submitted, and 8 sol-
lutions for the < 2 MP track.
ShuffleNetv2 submitted three models based on ShuffleNetV2 architecture [39]. All submitted models, ShuffleNetV2x0.5, ShuffleNetV2x1.5, and ShuffleNetV2x2.0 utilized full-precision FP32. These architectures were varied in size by scaling the width of every convolutional layer of ShuffleNetV2 by a factor of 0.5, 1.5, and 2, respectively. The embeddings are derived by global average pooling over the response of the final convolutional layer. Furthermore, the dimensionality of the embedding layer was decreased to 128 by reducing the number of filters in the final convolutional layer. To train the model, a linear classifier was used that projects the embedding into an 8631-dimensional logit vector. As a training set, the VGGFace2 dataset is used [10]. All images are aligned using the code provided by the organizers. This results in a dataset of 2.76M 112px images of 8631 subjects. This dataset is used to train the model with a cross-entropy loss, using the AdamW optimizer [37] and a batch size of 64, and cosine learning rate annealing from 3e-4 towards 1e-6. At test time, the classifier and the activation function over the embedding were removed. After training, the pointwise mean and standard deviation vectors of the embeddings over the training dataset are recorded and used to normalize the embeddings at test time, as \( \mathbf{v}_{emb} = (\mathbf{v}_{emb} - \mathbf{\mu}) / \mathbf{\sigma} \).

3.3 SQ-HH

HH-MB2 and HH-SQ employ a MobileNetV2 [46] and SqueezeNet [28] backbone, respectively. ImageNet pre-trained networks used as initial weights SQ-HH has also added batch normalization between convolutions and ReLU layers of SqueezeNet, which are not included in their original implementation. Both CNNs have been modified to employ an input size of 113×113×3 by changing the stride of the first convolutional layer from 2 to 1. This allows to keep the rest of the network unchanged and to reuse ImageNet parameters as starting model. Then, the networks undergo a double fine-tuning, first over MS1M-RetinaFace cleaned set [22] (35k subjects/3.16M images, only subjects with more than 70 images), and then over VGGFace2 [10] (9k subjects/3.31M images). Although VGG2 contains fewer users, it has more intra-class diversity due to more images per subject. Due to this fact, the double fine-tuning strategy employed has shown increased performance compared to training the models only with one database, especially if it has few images per identity [3, 10]. The networks are trained for biometric identification using the soft-max function and ImageNet as initialization. The optimizer is SGDM (mini-batch=128, learning rate=0.01, 0.005, 0.001, and 0.0001, decreased when the validation loss plateaus).
Two percent of images per subject in the training set are set aside for validation. The proposed models are trained with Matlab r2022b and use the ImageNet pre-trained model that comes with such a release. After training, feature vectors for biometric verification are extracted from the layer adjacent to the classification layer (i.e., the Global Average Pooling). This corresponds to a descriptor of 1280 (HH-MB2) and 1000 (HH-SQ) elements.

### 3.4. Idiap EdgeFace

Idiap EdgeFace submitted four solutions EdgeFace-S, EdgeFace-XS-1, EdgeFace-XS-2, and EdgeFace-XXS. These models adapted from EgdeNeXt [40] were used as the lightweight face recognition model. Specifically, the S for EdgeFace-S, XS for EdgeFace-XS1 and EdgeFace-XS2, and XXS for EdgeFace-XXS variants of the architecture were adopted by introducing a 512-D classification head and introducing low-rank linear layers, where \( \gamma \) is the ratio of rank. The details of these models can be found in the publication [19]. EdgeFace-S and EdgeFace-XS1 utilized full-precision FP32; EdgeFace-XS2 utilized Integer8 (Int8) and EdgeFace-XXS utilized Integer8 (Int8). CosFace [51] is used as the loss function and trained the network from scratch using a polynomial decaying training schedule with restarts for 200 epochs along with the learning rate warm-up for 2 epochs. The models were trained with 4/8 Nvidia RTX 3090 (24GB) GPUs using a distributed training strategy. The submitted models were trained using PyTorch with AdamW optimizer [37] with a weight decay of 5e-2, and a learning rate of 3e-3. The models were trained on WebFace4M/WebFace12M dataset. The batch size used was 512 and the network was trained using a distributed strategy using 4xGeForce RTX 3090, 24GB GPUs. Data augmentations such as random horizontal flips, random blur, grayscale, and resizing were used to improve the robustness of the model.

### 3.5. Visteam

The team submitted two half-precision models EfficientNet\( b0 \)\_visteam and MobileNet\( v2 \)\_visteam based on EfficientNet\( b0 \) [26] and MobileNetV2 [46] architecture. Both models were trained from scratch using ArcFace [14] loss function. The deep embedding during the training process is supervised by a large in-house model. Training is performed on the WebFace42M dataset [63]. To reduce the number of classes in the original dataset, its subset with \( \text{img\_per\_class} \geq 50 \). The Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and weight decay of 1e-4. The learning rate linearly decays from 0.01 to 0.0001 for 15 epochs. The batch size in the training phase was 128. Conventional data augmentation techniques: horizontal flipping, random gamma adjustment, RGB shifting, and color jittering, occlusion were used to avoid overfitting. After training the full-precision network was quantized to half-precision floating point.

### 3.6. SAM-MFaceNet eHWS

The team submitted two models, SAM-MFaceNet eHWS and Modified-MobileFaceNet In SAM-MFaceNet eHWS, a pre-trained full-precision FP32 MobileFaceNet architecture was trained using the MagFace loss function on a portion of WebFace42M dataset including 100K identities and 4.2M images. In Modified-MobileFaceNet, MobileFaceNet architecture was modified by doubling the kernels of the first three CNN layers. It was trained from scratch using the MagFace [44] loss function on a portion of WebFace42M [63] dataset including 100K identities. For both models, the sharpness-aware minimization (SAM) [17] optimizer with \( \rho = 0.05 \) and an exponential learning scheduler with a gamma of 0.998 was applied in the training processes. The initial learning rate for training the ResNet models was set to 0.1. The batch size in the training phase was 256 and the training epoch was set to 100.

### 3.7. Baselines

Baseslines have been selected to compare the submitted approaches with state of the art FR model performance. A diverse selection of previously published approaches for efficient FR is given to allow a comparison with a range of existing methods (e.g. large ResNet models [25, 14, 6], winner of the ICCV 2019 Lightweight Face Recognition challenge VarGFaceNet [56], PocketNet [9], MobileFaceNet [11]).

### 4. Results and Analysis

This section first introduces and compares the results of the baselines and the proposed solutions on the benchmarks used to determine the final ranking of the submissions. This analysis is followed by an additional in-depth evaluation of bias, cross-quality, and a large-scale recognition benchmark.

#### 4.1. Ranking benchmarks

Table 3 shows an overview of the baseline models and the submitted solution on the employed benchmarks including rankings of the participants’ solutions. The ranking is performed based on the evaluation criteria described in Section 2.2.

The largest baseline models use ResNet-100 as a backbone architecture. The network is trained once with the ElasticFace-Cos+ and ArcFace loss functions. On average, both models achieve the highest accuracies on the selected benchmarks. VarGFaceNet, the winner of the ICCV 2019 Lightweight Face Recognition Challenge [15] has only slightly lower accuracies than the ResNet-100 model,
especially on the cross-pose dataset CPLFW, but achieves these comparable results with only 8% of the parameters of the large ResNet-100 model. And thus VarGFaceNets requires only 20MB of memory footprint instead of the 261.22MB of the ResNet-100 models. The MixFaceNets [4] models have in comparison to ResNet-100 models and VarGFaceNet a slightly lower verification accuracy on the AgeDB30, LFW, and IJB-C benchmark. MixFaceNet-XS has a 4 percentage point reduction in accuracy on IJB-C in comparison to VarGFaceNet. But it uses only 161.9 FLOPs instead of 1022 FLOPs of VarGFaceNet. ShuffleFaceNet 1.5x [41] achieves a similar performance to VarGFaceNet but requires only half of the number of FLOPs. It has 2.6M parameters in comparison to VarGFaceNets 5.0M, and therefore requires a memory footprint of 10.5MB. The PocketNet [9] models achieve a small increase in verification accuracies (around 1 percentage point) in comparison to MixFaceNet-XS, but the number of FLOPs of these models has 6-fold increase when compared to MixFaceNet-XS.

From the results of the experiment of the models with 2-5 million parameters, the following conclusions can be drawn:

- The overall top-ranked solution is GhostFaceNetV1-1 KU (rank 1), followed by GhostFaceNetV1-2 KU (rank 2) and Idiap EdgeFace-S(γ=0.5) (rank 3).
- Highest ranked solution in terms of verification accuracies over the evaluated benchmarks is Idiap EdgeFace-S(γ=0.5). It also achieved the highest TAR at FAR=10^{-4} on the large IJB-C benchmark for all submitted solutions in this category. The ROC plots for models with 2-5 million parameters are shown in Figure 1b.
- GhostFaceNetV1-2 KU has the lowest number of FLOPs in the model category, and its number of FLOPs is lower than any of the baseline models.
- The lowest model size in MB is achieved by Idiap EdgeFace-XS-Q. With 2.24M parameters that are quantized to 8-bit integers, the model results in 2.99MB of memory footprint. In comparison, the baseline with the lowest number of parameters (MixFaceNet-XS) has a model size of 4.16MB, as 32-bit floating-point numbers are used for each of the 1.04M parameters.
- No submission of the category achieves high ranks in both cross-pose and cross-age benchmarks simultaneously.
- The model size and therefore the number of parameters of the submitted solution is in a similar range to that of the baseline models.
- The highest ranked solution GhostFaceNetV1-1 KU has at most a 3.5 percentage-points difference to the best performing baseline (ResNet-100 ElasticFace-Cos+), but requires only 0.89% of the number of FLOPs to the ResNet-100 model. When computational complexity and verification accuracies are considered, it outperforms the respective baseline.

For the submitted solutions with < 2 million parameters can be observed the following:

- Idiap EdgeFace-XS(γ=0.6) is the overall best performing model, SAM-MFaceNet eHWS V1 is ranked 2, SAM-MFaceNet eHWS V2 is ranked 3, respectively.
- Idiap EdgeFace-XS(γ=0.6) also ranks first when considering the verification accuracy, including first rank on IJB-C. The ROC curve for all submissions with < 2 million parameters are shown in Figure 1a.
- The smallest and most efficient model is ShuffleNetv2x0.5 (rank 7), with 0.17M parameters resulting in 0.77MB of model size and 17.14 numbers of FLOPs.
- Idiap EdgeFace-XS(γ=0.6) has at most a 4.5 percentage-points difference to the best performing baseline (ResNet-100 ElasticFace-Cos+), but requires only 0.63% of the number of FLOPs to the ResNet-100 model.

Almost all of the submitted models, either with < 2M parameters or 2-5M parameters, follow [14] and [51] respectively with the alignment procedure, but also with the embedding size. All top-ranking models use a margin-penalty based loss function like ArcFace [14], CosFace [51] or MagFace [44]. The majority of submissions use either SGD or AdamW as an optimizer. Only three different base datasets have been selected for training, which differ in their respective version primarily in their number of training samples and different processing steps. The data sets used are different versions of MS-Celeb-1M [22], VGGFace2 [10] and WebFace260M [63]. The submissions use similar network architectures: GhostNet [23], ShuffleNet [39], EdgeNext [40] and MobileNetV2 [46] are the most used backbone architectures. These models are established architectures that focus on efficient execution.

4.2. Further Benchmarks

We extend our evaluation beyond the competition ranking by evaluating bias (RFW), cross-quality (XQLFW), low-resolution (TinyFace), and a large scale recognition (IJB-C) benchmarks evaluated at different operation thresholds. The results for the additional evaluation is shown in Table 4.

To evaluate the bias of each submitted solution, we follow [52] and calculate the verification accuracies for Asian, African, Caucasian, and Indian identities of the RFW dataset. The average and standard deviation of the accuracies, as well as the skewed error ratio (SER), are computed. A higher SER indicates that a model inhibits a higher bias than a model with a lower SER value. The results show that the top-ranking models (GhostFaceNetV1-1
The methodologies used to develop the submitted efficient models are primarily compact network architectures. The top-performing submissions are adapting the architectures GhostNet and EgdeNeXt. Two submissions, Idiap EdgeFace-XS-Q, and Idiap EdgeFace-XXS-Q, are using model quantization with 8-bit integer quantization. Techniques that were unused by the submissions include knowledge distillation, pruning and sparsity-inducing training, neural architecture search, and binary neural networks.

In knowledge distillation [21], the knowledge of a larger, high-performing teacher network is transferred into a smaller student network. Pruning [36] involves dropping neurons or convolutional filters, which are not critical to the high performance of the network. If computational components can be removed without significant performance loss, a much more efficient and memory-saving model can be created. A similar approach is sparsity-induced training [18]. There, as many weights of the network as possible are set to 0 during the training process. This creates sparse matrices that are simpler and more efficient to compute, depending on the architecture. Neural architecture search [16], as used in the baseline PocketNet [9], tries to learn by an optimization algorithm as compact and small network components as possible, which are adapted to the respective learning task and therefore achieve higher performance with less computational effort. Instead of using float-point or integer parameters, binary neural networks [58] use binary values for the computations. This reduces the calculation effort considerably since significantly fewer logical operations are required for the calculations. Furthermore, it reduces the memory footprint of the model, as only a single bit per parameter is stored. Due to the described different, currently unused techniques, there is considerable scope for application and development to make efficient FR models even more compact and powerful.

6. Conclusion

The aim of this Efficient Face Recognition Competition (EFaR) was to benchmark and evaluate current technologies for lightweight and efficient face recognition. The submissions were evaluated with respect to the achieved verification accuracies on a diverse benchmark suite, in combination with their deployability given by the number of floating-point operations and the required memory footprint. Furthermore, additional evaluation was performed on bias, cross-quality, and large scale face recognition benchmarks. The results show that, when deployability is considered as well, top-ranking efficient face recognition solutions achieve comparable results to large face recognition models. Further, it is observed, that the majority of top-ranking solutions inhibit racial bias and achieve similar performance on XQLFW and IJB-C benchmarks. While the submissions show high performing, efficient face recognition models, further, currently unused techniques for the design of lightweight face recognition models can be adopted.

Acknowledgment 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. This work has been partially funded by the German Federal Ministry of Education and Research (BMBF) through the Software Campus Project.

References

Table 3: Evaluation results of the submitted solutions on the benchmarks introduced in Section 2.1. Also, FLOPs, model size and number of parameters are given. For each dataset, the respective achieved rank for each submission is given. The Borda count and rank over all verification benchmarks is given in the Accuracy column. The respective ranking is given for FLOPs and model size. The combined, final ranking is a weighted Borda count of the achieved accuracy (70%), the FLOPs ranking (15%) and the model size (15%). Submissions with 2-5M parameters are labeled as "2-5 MP". Submissions with < 2M parameters are labeled as "2 MP".

Table 4: Evaluation results of the submitted solutions on the RFW, XQLFW and IJB-C benchmarks. For RFW, the verification accuracy is given for Asian, African, Caucasian and Indian subgroups. The average accuracy and the standard deviation is calculated from the accuracies of the four ethnicity datasets. The SER value indicates the bias of a model, with a higher value indicating a higher bias of the model. For XQLFW, the verification accuracy is given. For TinyFaces, rank 1 and rank 5 metrics are shown. For IJB-C the TAR at different FAR are shown. Additionally, baselines are given for all datasets. Submissions with 2-5M parameters are labeled as "2-5 MP". Submissions with < 2M parameters are labeled as "2 MP".
Figure 1: The ROC curve scored by models in the respective track on the IJB-C benchmark.

(a) < 2M parameters models  (b) 2-5M parameters models


