# Rotation Consistent Margin Loss for Efficient Low-bit Face Recognition

Yudong Wu[1], Yichao Wu[1], Ruihao Gong[2], Yuanhao Lv[1], Ken Chen[1],
Ding Liang[1], Xiaolin Hu[3], Xianglong Liu[2], Junjie Yan[1]
[1]SenseTime Group Limited [2]BeiHang Unversity [3]Tsinghua Unversity
{wuyudong,wuyichao,liangding,yanjunjie}@sensetime.com

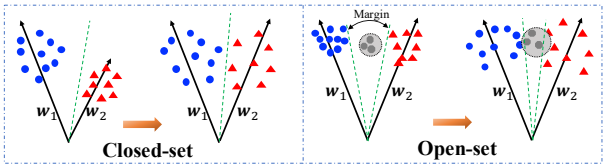
## Abstract

*In this paper, we consider the low-bit quantization problem of face recognition (FR) under the open-set protocol. Different from well explored low-bit quantization on closed-set image classification task, the open-set task is more sensitive to quantization errors (QEs). We redefine the QEs in angular space and disentangle it into class error and individual error. These two parts correspond to inter-class separability and intra-class compactness, respectively. Instead of eliminating the entire QEs, we propose the rotation consistent margin (RCM) loss to minimize the individual error, which is more essential to feature discriminative power. Extensive experiments on popular benchmark datasets such as MegaFace Challenge, Youtube Faces (YTF), Labeled Face in the Wild (LFW) and IJB-C show the superiority of proposed loss in low-bit (e.g., 4-, 3-bit) FR quantization tasks.*

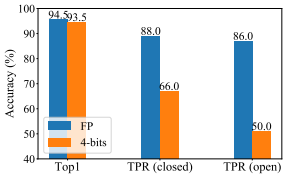
## 1. Introduction

The problem of face recognition (FR) has been well investigated [41, 42, 12, 4, 25, 8, 46, 54] in recent years. Among them, the deep FR technique, which leverages hierarchical and heavy-weight network architectures [43, 34, 38, 14, 16], has significantly improved the state-of-the-art performance and fostered wide-spread applications. Whereas excessive memory and computational consumption make it impractical to deploy massive networks on mobile or embedded devices.

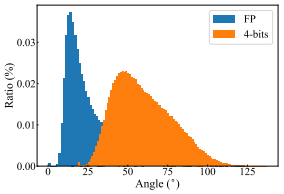
Quantization technique [6, 5, 31, 55, 56, 10] emerges as an elegant compression solution to address this problem. The core idea of this method is to reduce bit-width of weights and activations by mapping continuous values to discrete integers, which can not only reduce the memory footprint but also accelerate the inference directly. Despite the attractive benefits, low bit-width may degrade accuracy due to quantization errors (QEs). To minimize QEs, many methods have been proposed [53, 36, 2, 1], and have been demonstrated successful in the fields of closed-set computer vision problems[1], such as image classification and object detection [23]. However, for real-world FR, it is common that the testing identities (IDs) are disjoint from the training set, which makes open-set FR[2] problems more challenging and practical [25, 11]. In this study, we show that FR is much more sensitive to the QEs than closed-set classification tasks, thus novel sophisticated techniques are required to address this issue.




(a) Quantization processes of closed-set and open-set




(b) Comparison of accuracy




(c) Angles of positive pairs

Figure 1: Difference between closed-set and open-set quantization. (a) sketches the quantization processes of closed and open-set, respectively. $\boldsymbol{W}_i$ is the $i$-th column of the weight of the last fully connected layer ($i$-th class weight). Gray circle represents unknown face categories in test set, and the green dotted line denotes decision boundary. (b) compares the degradation of top1 accuracy and TPR after quantization. FP refers to full precision. (c) illustrates the increase of angles between positive pairs caused by quantization.

Essentially, FR is a typical scenario of metric learning where features of different IDs are expected to have discriminative large-margins rather than just be separable, as illustrated in Fig. 1a. To demonstrate the difficulty in open-set model quantization, we conduct a toy experiment. For

[1]For closed-set protocol, all testing classes are predefined in the training set [25].

[2]In this paper, "open-set FR" and "FR" can be used interchangeably.

closed-set protocol, we select 100 categories from CASIA-WebFace [52], and then choose 70% of all the data as training set and the rest as the test set. For open-set protocol, 10 additional categories are collected as open test set. We train a ResNet18 [14] and quantize it into 4-bit. As shown in Fig. 1b, the top1 accuracy of classification keeps almost unchanged after quantization, while the true positive ratio (TPR) of FR decreases more than 20%, especially on the open test set. Angles between positive pairs (two samples belong to the same id) also increase significantly after quantization (Fig. 1c), which indicates the reduced intra-class compactness.

Previous methods for closed-set classification quantization training attempt to completely eliminate the QEs by well-designed quantizers [53, 1, 2] or knowledge distillation (KD) [31]. In this paper, we argue that improving performance of low-bit FR models is not entirely equivalent to reducing the QEs, and propose a novel rotation consistent margin (RCM) loss function for efficient low-bit FR quantization. We define the QEs of face feature representation as the angle between its full precision (FP) feature and its quantized feature, named "A-QE". For a single sample, A-QE can be disentangled into two parts: *class error* and *individual error*. The *class error* refers to the overall rotation of a class caused by quantization. We observe that although it is enormous in low-bit models, it does not weaken the inter-class separability. On the other hand, *individual error* refers to the within-class deviation of each sample. It represents intra-class structural changes, which affect the performance of the quantized FR models essentially. Therefore, in this study, we attempt to minimize the individual error instead of the entire entangled A-QE by introducing it into cosine-based softmax loss function as an angular margin, which we call rotation consistent margin (RCM).

It is known that the embedding features of FR distribute on a fixed radius hypersphere [35, 8, 46, 45, 54]. Thus, QEs can be considered as rotation over the original distribution. Rotation consistent means that intra-class compactness remains unchanged regardless of class rotations, which can maintain same feature discrimination as the FP model. The proposed method rebuilds intra-class compactness efficiently by focusing on minimizing individual error. Meanwhile, rotation consistent is more realistic and easier to achieve because QEs are intrinsical and cannot be eliminated completely. Extensive experiments on MegaFace [21], LFW [17], YTF [51] and IJB-C [28] show that RCM loss function significantly improves the performance of low-bit FR models. Moreover, our method can be combined with other quantization methods and boost their performance.

To sum up, our contributions could be summarized into three parts:

- We redefine the QEs of FR in angular space, and disentangle QEs into class error and individual error. The former modifies inter-classes distribution, and the later determines the change of intra-class compactness.
- We rethink the essence of improving quantized FR models, and propose a novel loss function named as rotation consistent margin (RCM) loss for efficient low-bit FR model training by minimizing individual errors.
- To the best of our knowledge, we are the first to explore the quantization of FR. Extensive experiments on several accessible benchmark datasets demonstrate that our method effectively improves the performance of different low-bit FR models.

## 2. Related Work

### 2.1. Large-scale face recognition

Practical applications of FR are usually under the open-set protocol, where test categories are different from training categories. Therefore, FR is regarded as a typical metric learning task [33, 40, 47, 25, 46], whose objective is to increase the intra-class compactness and inter-class discrepancy. To this end, many loss functions have been proposed. Sun et al. [41, 42] and Wang et al. [50] combine softmax loss with contrast loss or center loss, respectively, to explicitly increase the margin between different classes or reduce the distance between positive sample pairs. FaceNet [38] adopts triplet loss to optimize the embedding features directly, and greatly boost the performance. Recent works [26, 25, 46, 8, 45] propose the cosine-based softmax loss function and incorporate with margin to enhance the feature discriminative power. In these methods, features and class weights are normalized to a fixed scale, and the hypersphere manifold distribution hypothesis arises correspondingly and is widely recognized due to the concise geometric interpretation and impressive performance achieved by those loss functions.

### 2.2. Network quantization

Network quantization is a technique of network compression that works as an analog-to-digital converter: quantizing FP weights and activations into low precision fixed-point integers. Through efficient bit operation or integer-only arithmetic, it can both reduce the storage overhead and accelerate the inference significantly. Common quantization types include binary/ternary [5, 6, 18, 22, 57], uniform [58, 31, 29, 55, 20, 48] and non-uniform [56, 44, 3, 53, 13] quantization. Both uniform and binary/ternary quantizers are hardware-friendly that can enjoy acceleration directly on off-the-shelf hardware [20, 19, 9, 30]. Post-training quantization and quantization-aware training are two typical quantization schemes. The former solves the value ranges without re-training. The later finetunes the optimized FP model in a simulated quantization scheme and usually yields higher accuracy.

## 3. Preliminaries

### 3.1. Cosine-based softmax loss function

In original deep FR methods [43, 41], models are trained using softmax cross-entropy loss function [3],

$$\mathcal{L}_{\text{softmax}} = -\log \frac{e^{z_{i,j}}}{\sum_{j=1}^{n} e^{z_{i,j}}}, \tag{1}$$

where $z_{i,j} = \boldsymbol{W}_j^T \boldsymbol{f}_i$ is the logit of $j$-th class. $\boldsymbol{W}_j \in \mathbb{R}^d$ is the $j$-th column of weights of last fully connected layer, and bias is omitted for simplicity, $\boldsymbol{f}_i \in \mathbb{R}^d$ refers to the feature of $i$-th sample.

Recently, it is argued that the vanilla softmax loss cannot force features to have higher discriminative power, and cosine-based softmax incorporated with margin is proposed [8, 54, 46], where $\boldsymbol{W}_j^T \boldsymbol{f}_i$ is reformulated as $s \cdot \cos\theta_{i,j}$, and $\theta_{i,j}$ is the angle between $\boldsymbol{W}_j$ and $\boldsymbol{f}_i$, and $s$ is a scale hyperparameter. ArcFace [8] uses additive angular margin $z_{i,j} = s \cdot \cos(\theta_{i,j} + \mathbb{1}\{j = y_i\} \cdot m)$, CosFace [46] uses additive cosine margin $z_{i,j} = s \cdot (\cos\theta_{i,j} - \mathbb{1}\{j = y_i\} \cdot m)$, and SphereFace [25] uses multiplicative angular margin $z_{i,j} = s \cdot \cos(\mathbb{1}\{j = y_i\} \cdot m \cdot \theta_{i,j})$. The indicator function $\mathbb{1}\{j = y_i\}$ returns 1 when $j = y_i$ and 0 otherwise. All of them achieve significant improvement.

### 3.2. Quantization process

Generally, weights and activations of deep models are represented in FP values with 32-bit. Network quantization represents them in fixed-point integers with lower bit-width (e.g., 8-, 4-bit etc.). Among popular quantizers, the binary and uniform quantizers are hardware-friendly, which enables us to accelerate the inference directly on off-the-shelf hardwares [20, 19, 9, 30]. Therefore, the following discussions are all under the uniform protocol.

For $n$-bit uniform quantization, the process can be defined as:

$$x_Q = \text{round}\left(\frac{\text{clamp}(x_{min}, x_{max}, x) - x_{min}}{\Delta}\right) \quad , \tag{2}$$

where $x_Q$ is the integer number in $n$-bit width, and $x_{min}$, $x_{max}$ is the lower and upper bound of FP values. For per-layer quantization scheme, an entire layer shares the same $(x_{min}, x_{max})$, and for per-channel scheme, each channel has different boundaries. $\Delta = \frac{x_{max} - x_{min}}{2^n - 1}$ is the interval length.

## 4. Proposed Approach

### 4.1. Angle based quantization errors in FR

Generally, when the bit-width goes down, the accuracy of quantized models degrades dramatically due to QEs. QEs refer to the rounding and truncation errors introduced by representing continuous values in $n$-bit fixed-point number. For a single value $x$, QE defines as follows,

$$\text{QE}(x) = x - \text{Q}(x) \tag{3}$$

where $\text{Q}(x)$ is de-quantization FP value of $x_Q$. For $n$-bit uniform quantization, the de-quantization operation is:

$$\text{Q}(x) = x_{min} + x_Q * \Delta. \tag{4}$$

In previous works [53, 1], QE of the $d$-dimension feature or tensor is defined as the average error of each dimension,

$$\text{QE}(\boldsymbol{f}_i) = \frac{1}{d}\sum_{l=1}^{d}(f_i^l - \text{Q}(f_i^l))^2 \quad , \tag{5}$$

where $f_i^l$ is the $l$-th dimension of feature $\boldsymbol{f}_i$. As for FR, features are angularly distributed, i.e., on the surface of a fixed radius hypersphere, and the angle or cosine similarity models the interrelation between samples. Reasonably, we redefine the QE of face feature as the angle between its quantized feature and its FP feature,

$$\text{A-QE}(\boldsymbol{f}_i) = \arccos\left(\left\langle \frac{\boldsymbol{f}_i}{||\boldsymbol{f}_i||_2}, \frac{\hat{\text{Q}}(\boldsymbol{f}_i)}{||\hat{\text{Q}}(\boldsymbol{f}_i)||_2} \right\rangle\right) \quad , \tag{6}$$

where $\hat{\text{Q}}(\boldsymbol{f}_i)$ is the feature of quantized model. Compared with the vanilla mean-square error definition, A-QE has a more clear geometric interpretation and is also more intuitive in FR. It intuitively reflects the rotation caused by quantization.

### 4.2. Disentangling A-QE

In this subsection, we investigate the effects of A-QE on FR by theoretical analysis and empirical experiments, and then propose to disentangle the A-QE of a single sample into *class error* and *individual error*.

Initially, we represent the class center as the mean of all sample features:

$$\boldsymbol{c}_{y_i} = \frac{1}{n}\sum_{i=0}^{n}\boldsymbol{f}_i, \qquad s.t. \boldsymbol{f}_i \in \mathcal{C}_{y_i}, \tag{7}$$

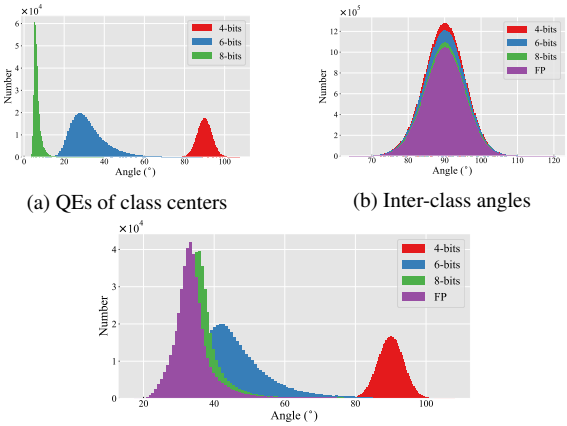
where $\mathcal{C}_{y_i}$ is the $y_i$-th class set. For $\boldsymbol{f}_i \in \mathcal{C}_{y_i}$, we denote the QE of $l$-dimension as $\delta^l$, then the center of the quantized class,

$$\begin{aligned} qc_{y_i}^l &= \frac{1}{n}\sum \hat{Q}(f_i^l) \\ &= \frac{1}{n}\sum (f_i^l + \delta_i^l) \\ &= \frac{1}{n}\sum f_i^l + \frac{1}{n}\sum \delta_i^l \quad , \end{aligned} \tag{8}$$

where $qc_{y_i}$ denotes the center of the quantized class. If we assume that $\delta^l$ is a Gaussian distribution, i.e. $\delta^l \sim \mathcal{N}(\mu_{y_i}, \sigma_{y_i}^2)$, then $qc_{y_i}^l = c_{y_i}^l + \mu_{y_i}^l$. The rotation angle of class center after quantization is

$$\theta_{c_{y_i}} = \arccos\left(\frac{\sum_{l=0}^{d} c_{y_i}^l \cdot (c_{y_i}^l + \mu_{y_i}^l)}{\|\boldsymbol{c}_{y_i}\|_2 \cdot \|\boldsymbol{qc}_{y_i}\|_2}\right) \quad . \tag{9}$$

[3] We denote it as "softmax loss" for short in the following sections.




(a) QEs of class centers (b) Inter-class angles

(c) Intra-class angles

Figure 2: QEs analysis of FR model. We train a ResNet18 [14] on CASIA-WebFace [52] dataset and quantize it into 8-, 6-, 4-bit, respectively. (a) plots the rotation angles of class centers of different bit-width models. (b) illustrates the angles between all class center pairs (inter-class angles). (c) demonstrates the within-class angles.

To reveal the actual rotation angles in FR, we conduct experiments on the CASIA-WebFace [52] and plot the distribution of A-QE of each class center in Fig. 2a. Specifically, we train an FP ResNet18 [14] model using Arcface [8] and quantize it into 8-, 6-, 4-bit, respectively. To avoid modifying model weights, we adopt the post-training quantization scheme.

**Observation #1.** *As shown in Fig. 2a, the rotation angles of class centers are significant and increase dramatically as the bit-width decreases.* We name this rotation caused by quantization, i.e., A-QE($\mathrm{c}_{\mathrm{y_i}}$), as *class error*.

**Analysis.** The significant class error indicates that the A-QEs of samples in the same category are not completely random; samples are rotated in a common direction. Accordingly, the entire class is rotated. Repeat experiments are performed on MobileNetV2 [37], VGG [39] to exclude the effect of network architectures, and the same phenomena are observed.

The distribution of classes mainly determines the inter-class separability. To further investigate, we use the angles between class center pairs (inter-class angles) to demonstrate the overall inter-class separability of different bit-width models. The distributions are shown in Fig. 2b.

**Observation #2.** *Compared with the original FP model, there is no obvious change of the inter-class angles after quantization: most remain around* $\frac{\pi}{2}$. This observation holds when the bit-width is reduced.

**Analysis.** The stationary inter-class angle means the inter-class discrepancy keeps stable after quantization. Based on the above two observations, we can briefly summarize that although QEs rotate classes sorely, the discriminative power between classes has not been weakened.

On the other hand, we also investigate the change of intra-class compactness after quantization. The original definition of within-class scatter is based on the distance between samples and class centers [4]. We replace the Euclidean distance by angle and show within-class angles in Fig. 2c.

**Observation #3.** *The within-class angles increase significantly after quantization, which means that the intra-class compactness is noticeably reduced, especially in low-bit models.* The within-class angles are about 30° in the FP model, whereas they surge to 90° when quantized to 4-bit.

**Analysis.** We can empirically conclude that QEs weaken the intra-class compactness, and as the bit width decreases, the degree of weakening increases.

Based the on above investigations, we propose to decompose A-QE of a sample into *class error* and *individual error* two parts,

$$\text{A-QE}(\boldsymbol{f}_i) = \text{A-QE}(\boldsymbol{c}_{\mathrm{y_i}}) + \mathcal{I}(\boldsymbol{f}_i) \ , \tag{10}$$

where $\mathcal{I}(\boldsymbol{f}_i)$ refers to the individual error. From the perspective of a single sample, its rotation caused by quantization can be considered to follow the class center and then deflect within the class. The within-class deflection degrades the original stable intra-class compactness and leads to inferior performance. The dissection is sketched in Fig.3.

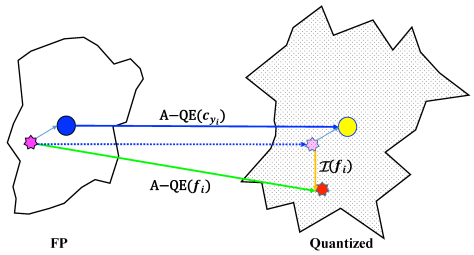



Figure 3: Dissection of A-QE. The left showcases a class of FP model, and the right illustrates that class after quantization. Circles and stars refer to the class centers and individual samples, respectively. The green solid line represents the entire A-QE of the sample, blue represents class error and orange represents the individual error. For a clear illustration, we use euclidean distances instead of vector angles to denote errors.

### 4.3. Rotation consistent margin

Initially, quantization methods [20, 49, 18] ignore QEs by taking quantizers as general operators and directly use task loss to tune models in quantization aware training. Afterward, several methods attempt to minimize the entire QEs by well-designed quantizers [53, 1, 2] or knowledge distillation [31], and usually yield higher performance. In spite of the improvement brought by reducing the entire QEs, a question worth thinking about is: **Is improving the**

[4]The within-class scatter matrix is defined as $\mathbf{S}_w = \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_{y_i})(\mathbf{x}_i - \boldsymbol{\mu}_{y_i})^T$, where $\boldsymbol{\mu}_{y_i}$ is the mean of all samples in class $y_i$.

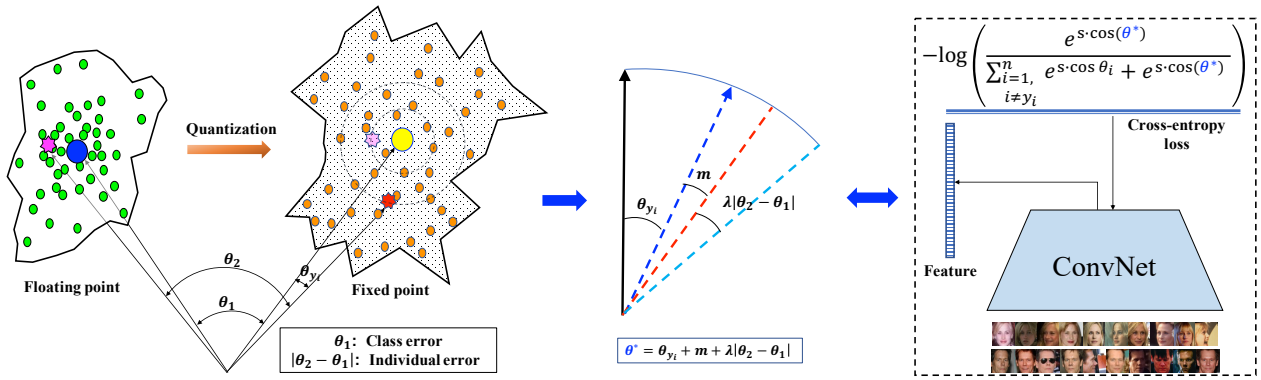



Figure 4: Training process for low-bit FR models supervised by RCM loss function $\mathcal{L}_{\mathrm{A}}$. Firstly, $\theta_{y_i}$ in quantized class is calculated. Then we employ two class centers (before and after quantization, represented in big circles) as anchors to obtain the class error ($\theta_1$). The entire A-QE of samples (represented in stars) is also calculated, i.e., $\theta_2 = \arccos\left(\left\langle \frac{\boldsymbol{f}_i}{||\boldsymbol{f}_i||_2}, \frac{\hat{\mathrm{Q}}(\boldsymbol{f}_i)}{||\hat{\mathrm{Q}}(\boldsymbol{f}_i)||_2} \right\rangle\right)$, and individual error is $|\theta_1 - \theta_2|$. We use the $\theta^* = \theta_{y_i} + \lambda|\theta_1 - \theta_2| + m$ as the final angle to calculate logits. The logits then go through the softmax function and contribute to the cross-entropy loss. The FP feature $\boldsymbol{f}_i$ is extracted off-line before training.

**accuracy of quantized models exactly equivalent to reducing all the quantization errors?**

Our answer is "not for FR", the essence of improving low-bit FR models is to rebuild the intra-class compactness and the inter-class separability. By dissecting A-QE, we disentangle the QE of a sample into class error and individual error. The class error is the common part of all class samples (class-wise) which rotates class as a whole. Although it is considerable, it dose not impair inter-class discrepancy. Actually, due to the sparsity of high-dimensional class weights, the inter-class angles keep around $\frac{\pi}{2}$ during the whole training process, and this phenomenon has also been validated in other works [58, 15]. We validate that this phenomenon also exists in the low-bit scenario. It indicates that during training, the objective of inter-class separability maintains as a regularization other than pushing class weights further apart in whether FP or low-bit models. In contrast, the individual error is a unique part of each sample (sample-wise), and it changes the intra-class structure. As shown in Fig. 2c, the within-class scatter increases significantly after quantization, which implies the individual error cracks the intra-class compactness. Therefore, we argue that improving the accuracy of quantized FR models is not completely equivalent to reducing all the quantization error. If we directly minimize the entire A-QE, supervision of class error would drive the model to pull the class towards the FP position instead of rebuilding the impaired intra-class compactness.

Alternatively, we propose to only minimize the individual error. We introduce the individual error into the popular cosine-based softmax loss function as an additive angular margin, named *rotation consistent margin* (RCM),

$$\mathcal{L}_{\mathrm{A}} = -\log \frac{e^{s\cdot\cos\left(\theta_{i,j}+\mathbb{1}\cdot m+\mathbb{1}\cdot\lambda\theta_Q\right)}}{\sum e^{s\cdot\cos\left(\theta_{i,j}+\mathbb{1}\cdot m+\mathbb{1}\cdot\lambda\theta_Q\right)}} \quad , \tag{11}$$

where $\lambda$ is the scaling parameter, and we write $\mathbb{1}\{j = y_i\}$ as $\mathbb{1}$ for the clear demonstration. As the original cosine-based softmax loss function is not the focus of our discussion, for simplicity and fairness, we choose currently the competitive loss function, ArcFace [8], as our baseline. We keep the original additive margin $m$. The individual error is calculated as follows,

$$\theta_Q = |\text{A-QE}(\boldsymbol{f}_i) - \text{A-QE}(\boldsymbol{c}_{y_i})| \quad . \tag{12}$$

The complete training pipeline supervised by $L_A$ is illustrated in Fig. 4.

Rotation consistent means that intra-class compactness stays stable regardless of the huge rotation of the classes, and the performance is improved even though the class errors are still huge, which is validated in Sec 5.2. By minimizing the individual error, the proposed loss can enhance the impaired compactness efficiently. Meanwhile, inter-class distribution is tuned appropriately via classification loss rather than being pulled towards the original position.

### 4.4. Discussions

**Why angle margin.** Besides incorporating individual error to cosine-based softmax as an angular margin, combining by weighted sum is also an intuitive way to reduce individual error,

$$\mathcal{L}_{\mathrm{sum}} = \mathcal{L}_{\cos} + \lambda\theta_Q \quad , \tag{13}$$

where $\mathcal{L}_{\cos}$ refers to the cosine-based softmax loss function. The reasons for adopting the angular margin are two-folds. On the one hand, angular margin brings more clear geometric interpretation and directly links to the discrimination on the hypersphere manifold. On the other hand, incorporating as an angular margin has stronger supervision on most medium hard samples, and produces superior optimization by weakening the influence of too simple and hard samples.

The gradient of $\mathcal{L}_{\mathrm{A}}$ respect to $\theta_Q$ is,

$$\frac{d\mathcal{L}_{\mathrm{A}}}{d\theta_Q} = \lambda s^2 \frac{\sum_{j=1,j\neq i}^{C} e^{z_{i,j}}}{\sum_{j=1,j\neq i}^{C} e^{z_{i,j}} + e^{s\cdot\cos(\theta^*)}} \cdot \sin(\theta^*) \ , \quad (14)$$

where $\theta^* = \theta + m + \lambda\theta_Q$ and $C$ is the class number. For non-corresponding classes $j \neq i$, $\theta_{i,j}$ always stays around $\frac{\pi}{2}$ during training, thus we assume $e^{z_{i,j}} \approx 1$ as in [54]. We plot $\frac{d\mathcal{L}_{\mathrm{A}}}{d\theta_Q} = \lambda s^2 \frac{C}{C-1+e^{s\cdot\cos(\theta^*)}} \sin(\theta^*)$ with different $C$ in Fig. 5. At around $\theta^* = \frac{\pi}{2}$, the gradient has a maximum value. By appropriate $\lambda$, $\mathcal{L}_A$ can have stronger supervision at the medium hard samples instead of equal supervision for too simple or too hard samples. Too hard samples can be noise, which is common in large scale FR training datasets, and too simple samples are hard to be further optimized whereas would dominate the training and result in inferior models [24].

**Decision boundary.** Considering a binary-classes scenario, the decision boundaries of the proposed loss function is defined by,

$$\cos(\theta_1 + m + \lambda\theta_Q) = \cos(\theta_2) \ . \quad (15)$$

For $c_1$, it requires $\theta_1 < \theta_2 - m - \lambda\theta_Q$. In ArcFace, the margin $m$ is immutable and identical for all samples, and the decision boundary also stays the same during training (Fig. 6a). Whereas the rotation consistent margin of the proposed approach is sample-wise and dynamically shrinks during the training (Fig. 6b). In the beginning, significant individual errors generate strong supervision to rebuild intra-class compactness efficiently. At the later stages of training, the rotation consistent margin becomes steady, and optimization turns to learn customized model weights in discrete parameter space.

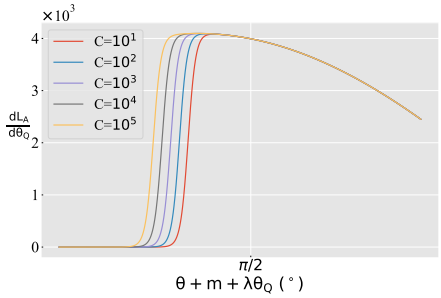
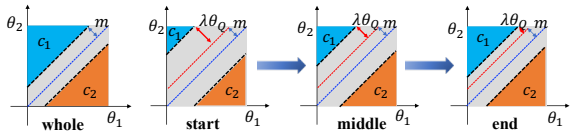



Figure 5: The gradient of $\mathcal{L}_{\mathrm{A}}$ respect to $\theta_Q$ of different class number $C$. $s$ is set to 64 as in ArcFace [8] and $\lambda$ is set to 1.

# 5. Experiment

## 5.1. Experiments setting

**Datasets.** For datasets, we separately employ publicly available CASIA-WebFace [52] and MS1MV1 [7] cleaned by deepglint as training datasets. The CASIA-WebFace dataset contains 0.49M images of 10,575 face identities.




(a) ArcFace (b) $L_A$

Figure 6: The comparison of decision boundaries. (a): Decision boundary of ArcFace, which is the same for all samples and immutable during training. (b): Decision boundary of $L_A$. It is sample-wise and dynamically changing during training.

Cleaned MS1MV1 is a large scale dataset that consists of 3.9M images from 87K face identities. We extensively evaluate the performance of our approach on several most widely used benchmark face datasets, including MegaFace [21], Labeled Face in the Wild (LFW) [17], Youtube Faces (YTF) [51] and IJB-C [28].

**Training.** We employ the widely used CNN architectures, ResNet18 [14] and MobileNetV2 [37]. Following [8], BN-Dropout-FC-BN structure is adopted to get the final $256-\mathrm{D}$ features. We use the SGD algorithm with a momentum of 0.9 and set weight decay 0.0005. Eight GPUs are used with a single batch size of 64. For training on CASIA-WebFace, the learning rate is initially 0.1 and divided by 10 at the 20K, 28K iterations, and the training finished at 32K iterations. For large scale dataset MS1MV1, the learning rate dropped at 100k, 160k iterations, and terminates at 180K.

**Quantization setting.** In this paper, we employ the asymmetric uniform quantizer, which is hardware-friendly. As per-channel quantization scheme usually yields higher accuracy, we adopt it for weights. As for activations, the per-layer scheme is used because per-channel would complicate the inner product computation at the core of conv and matmul operations [20]. All convolution and fully-connected layers except the first and last one are quantized. We adopt quantization-aware training and initialize quantized weights from FP models.

## 5.2. Ablation studies on RCM loss

**Selection of $\lambda$ and formulas.** There are three potential positions for margin in cosine-based softmax, like additive angle margin in ArcFace [8] (i.e.,$\mathcal{L}_{\mathrm{A}}$), additive cosine margin in CosFace [46] and multiplicative angular margin in SphereFace [25]. We denote another two feasible formulas as

$$\mathcal{L}_{\mathrm{B}} = -\log \frac{e^{s\cdot(\cos(\theta_{i,j}+\mathbb{1}\cdot m)-\mathbb{1}\cdot\lambda\theta_Q)}}{\sum e^{s\cdot(\cos(\theta_{i,j}+\mathbb{1}\cdot m)-\mathbb{1}\cdot\lambda\theta_Q)}}$$

and

$$\mathcal{L}_{\mathrm{C}} = -\log \frac{e^{s\cdot\cos((\mathbb{1}\cdot\lambda\theta_Q+1)\theta_{i,j}+\mathbb{1}\cdot m)}}{\sum e^{s\cdot\cos((\mathbb{1}\cdot\lambda\theta_Q+1)\theta_{i,j}+\mathbb{1}\cdot m)}} \ .$$

In this part, we explore the effect of the scaling parameter $\lambda$ and the performance of different formulas. To this end,

we train an FP ResNet18 on CASIA-WebFace. Then we quantize the resulting model into 4-bit using three different formulas with $\lambda$ varying from 1 to 7. Performance is evaluated on MegaFace and illustrated in Fig. 7a. We can see that the performance of $\mathcal{L}_{\mathrm{A}}$ outperforms the other two formulas slightly. For $\mathcal{L}_{\mathrm{A}}$, the performance gets saturated at $\lambda = 5.0$, thus we use $\mathcal{L}_{\mathrm{A}}$ with $\lambda = 5.0$ in the subsequent experiments of this study.

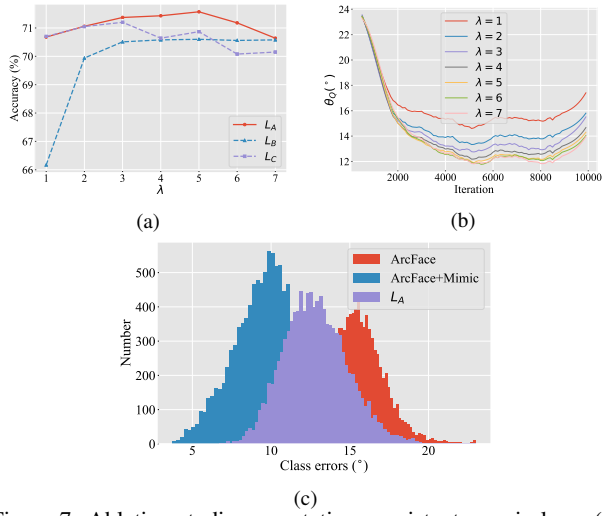



Figure 7: Ablation studies on rotation consistent margin loss. (a) accuracy(%) of 4-bit ResNet18 with different formulas and various $\lambda$ on MegaFace. (b) change of individual error of $\mathcal{L}_A$ with different $\lambda$ during the training. (c) class errors of 4-bit ResNet18 trained by ArcFace, ArcFace+Mimic and $\mathcal{L}_A$. The errors are in angle.

**Change of RCM during training.** In this part, we will show the change of RCM $\theta_Q$ of $\mathcal{L}_A$ loss during training. Fig. 7b illustrates $\theta_Q$ of different $\lambda$ settings. As we can see, the margin gradually decreases as training, and the larger $\lambda$ brings stronger supervision, which leads to a smaller $\theta_Q$. Whereas when the $\lambda$ increases to 5, the RCM becomes stable and hard to be further compressed. Meanwhile, the accuracy also reaches the inflection point, and the larger $\lambda$ would degrade the accuracy. It is due to that $\lambda$ balances the supervision of classification and minimizing individual error, excessively increasing $\lambda$ cannot consistently improve intra-class compactness but weakens the supervision of classification.

**Comparison of class errors.** To investigate the resulting class errors, we train 4-bit ResNet18 using ArcFace, ArcFace+Mimic and $L_A$, respectively. Directly supervised by ArcFace ignores the QEs and takes quantizers as general operators. Combining with mimic or knowledge distillation (KD) not only finetunes the models according to the classification loss but also pulls the features towards the original position by extra supervision from FP models. Empirical evidence [27] shows that mimicking the feature layer brings more improvements for FR than KD. Thus we use ArcFace+Mimic instead of ArcFace+KD. The performance on MegaFace and resulting class errors are illustrated in Tab. 1 and Fig. 7c. Compared with ArcFace, ArcFace+Mimic results in smaller class errors, meanwhile, improves performance. Whereas, our approach achieves higher accuracy than ArcFace+Mimic with larger class errors. The experimental phenomenon supports our hypothesis that improving the accuracy of quantized FR models is not precisely equivalent to reducing all the QEs, and minimizing only individual errors can bring more improvement.

| loss | Size of MegaFace Distractor | | | | | |
|---|---|---|---|---|---|---|
| | $10^1$ | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| FP | 98.13 | 95.75 | 92.10 | 86.96 | 80.72 | 73.35 |
| ArcFace | 97.73 | 94.52 | 90.28 | 84.29 | 76.37 | 67.75 |
| +Mimic | **97.93** | 94.88 | 90.65 | 84.78 | 77.39 | 69.04 |
| $\mathcal{L}_A$ | 97.85 | **95.32** | **91.46** | **86.33** | **79.37** | **71.57** |

Table 1: Identification accuracy (%) of rank-1 of 4-bit ResNet18 on the MegaFace dataset. "+Mimic" refers to ArcFace+Mimic.

## 5.3. Results on LFW and YTF

LFW [17] is a standard face verification testing dataset in unconstrained conditions, and all images are collected from the website. It contains 13,233 face images from 5,749 identities with a total of 6,000 ground-truth matches. Half of the matches are positive, while the other half are negative ones. YTF [51] consists of 3,425 videos from 1,595 different people. All the videos are collected from YouTube. In this paper, evaluation results are reported strictly following the standard protocol of *unrestricted with labeled outside data*.

We train ResNet18 on CASIA-WebFace dataset and report performance of 4-bit and 3-bit quantized models supervised by several accessible loss functions in Tab. 2. On LFW and YTF, the proposed RCM loss achieves 98.91% and 94.98% at 4-bit, and 98.73% and 94.56% at 3-bit. The results outperform all compared loss functions.

| Method | LFW | | YTF | |
|---|---|---|---|---|
| | 4-bit | 3-bit | 4-bit | 3-bit |
| FP | 98.93 | | 94.97 | |
| SoftMax | 98.60 | 98.26 | 94.28 | 93.49 |
| l2-Softmax [35] | 98.55 | 97.85 | 94.01 | 93.51 |
| CosFace [46] | 98.66 | 97.88 | 94.36 | 93.28 |
| CosFace+Mimic | 98.76 | 98.18 | 94.37 | 93.53 |
| ArcFace [8] | 98.63 | 98.55 | 94.62 | 93.50 |
| ArcFace+Mimic | 98.68 | 98.46 | 94.76 | 93.95 |
| $\mathcal{L}_{\mathrm{A}}$ | **98.91** | **98.73** | **94.95** | **94.56** |

Table 2: Verification accuracy (%) on the LFW and the YTF datasets. The model ResNet18 is trained on WebFace.

| Model | Method | Accuracy(%) 4-bit | 3-bit |
|---|---|---|---|
| ResNet18 | FP | 93.71 | |
| | Softmax | 82.43 | 27.22 |
| | CosFace [46] | 87.04 | 60.41 |
| | CosFace+Mimic | 87.73 | 57.68 |
| | ArcFace [8] | 87.31 | 67.83 |
| | ArcFace+Mimic | 87.74 | 72.84 |
| | $\mathcal{L}_{\mathrm{A}}$ | **88.56** | **74.34** |
| MobileNetV2 | FP | 91.31 | |
| | Softmax | 69.21 | 24.11 |
| | l2-Softmax [35] | 71.68 | 47.5 |
| | CosFace [46] | 74.99 | 57.00 |
| | CosFace+Mimic | 74.05 | 59.44 |
| | ArcFace [8] | 77.73 | 56.79 |
| | ArcFace+Mimic | 77.90 | 59.59 |
| | $\mathcal{L}_{\mathrm{A}}$ | **80.12** | **62.58** |

Table 3: Identification accuracy (%) of rank-1 on MegaFace dataset. The size of distractor is 1M.

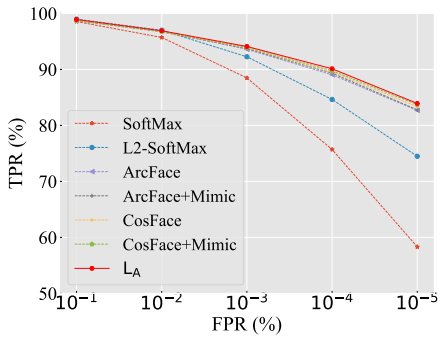



Figure 8: TPR on IJB-C benchmark with FPR varying from $10^{-1}$ to $10^{-5}$. The model is 4-bit ResNet18 and trained on MS1MV1.

### 5.4. Results on MegaFace

The MegaFace dataset [21] is a very challenging large-scale testing benchmark. It contains 1M images from 690K different individuals as the gallery set and 100K photos of 530 unique individuals from FaceScrub [32] as the probe set. We follow the testing protocol of ArcFace [8]. All the models are trained on the MS1MV1 dataset.

The rank-1 identification accuracies with 1M distractors are summarized in Tab. 3. Our RCM approach shows its superiority on both the ResNet18 and MobileNetV2 with a clear margin. For ResNet18, RCM achieves 88.56% and 74.34% on 4- and 3-bit and outperforms other losses. MobileNetV2 has excellent speed-accuracy trade-off and is hard to quantize, and RCM loss boosts the accuracy significantly. Whereas the gap with FP model still exists, which needs further investigations.

### 5.5. Results on IJB-C

The IJB-C dataset [28] contains about 3,500 identities, with a total of 31,334 still facial images and 117,542 unconstrained video frames. In the 1:1 verification, there are a total of 19,557 positive pairs and 15.6M negative pairs.

We employ the MS1MV1 dataset as training data and report TPR of 4-bit ResNet18 in Fig. 8. Compared with other popular loss functions, our proposed RCM achieves state-of-the-art performance at different FPR.

### 5.6. Compatibility with other quantization baselines

The proposed RCM loss improves the performance of the low-bit model from the perspective of the discriminative essence of open-set tasks. It can be combined with previous methods to boost their performance further. Here, we re-implement the recent two state-of-the-art quantization methods Dorefa-Net [55] and DSQ [10] as quantization baselines and combine them with different loss functions.

We report the evaluation results of 4-bit ResNet18 on MegaFace. As shown in Tab. 4, both Dorefa-Net and DSQ can improve the baseline of different loss functions, and combining with RCM achieves the best accuracy. It demonstrates our approach is compatible with different quantization methods and can further boost their performance.

| Basic | Method | Accuracy(%) |
|---|---|---|
| Dorefa-Net [55] | +CosFace | 87.11 |
| | +CosFace+Mimic | 90.59 |
| | +ArcFace | 88.27 |
| | +ArcFace+Mimic | 89.63 |
| | +$\mathcal{L}_{\mathrm{A}}$ | **91.55** |
| DSQ [10] | +CosFace | 87.07 |
| | +CosFace+Mimic | 88.98 |
| | +ArcFace | 87.79 |
| | +ArcFace+Mimic | 88.42 |
| | +$\mathcal{L}_{\mathrm{A}}$ | **89.46** |

Table 4: Identification accuracy (%) of rank-1 4-bit ResNet18 on the MegaFace dataset. The size of distractor is 1M and the FP models are trained on MS1MV1.

## 6. Conclusion

In this work, we investigate the effect of quantization errors on FR and propose rotation consistent margin loss for efficient low-bit FR training. Competitive results on several popular face benchmarks demonstrate the superiority and great potentials of our approach. It is hoped that our substantial explorations will inspire more researches on quantization problems for the open-set scenario.

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China under Grant Nos. U19B2034 and 61836014.

## References


[1] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Aciq: Analytical clipping for integer quantization of neural networks. *arXiv preprint arXiv:1810.05723*, 2018.

[2] R Banner, Y Nahshan, E Hoffer, and D Soudry. Post training 4-bit quantization of convolution networks for rapid-deployment. *CoRR, abs/1810.05723*, 1:2, 2018.

[3] Zhaowei Cai, Xiaodong He, Jian Sun, and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5918–5926, 2017.

[4] Ken Chen, Yichao Wu, Haoyu Qin, Ding Liang, Xuebo Liu, and Junjie Yan. R3 adversarial network for cross model face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9868–9876, 2019.

[5] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pages 3123–3131, 2015.

[6] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

[7] deepglint. Face feature test/trillion pairs. `http://trillionpairs.deepglint.com/overview`.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[9] Jiong Gong, Haihao Shen, Guoming Zhang, Xiaoli Liu, Shane Li, Ge Jin, Niharika Maheshwari, Evarist Fomenko, and Eden Segal. 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*, page 2. ACM, 2018.

[10] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4852–4861, 2019.

[11] Manuel Gunther, Steve Cruz, Ethan M Rudd, and Terrance E Boult. Toward open-set face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 71–80, 2017.

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[13] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Lanqing He, Zhongdao Wang, Yali Li, and Shengjin Wang. Softmax dissection: Towards understanding intra-and inter-clas objective for embedding learning. *arXiv preprint arXiv:1908.01281*, 2019.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[18] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in neural information processing systems*, pages 4107–4115, 2016.

[19] Benoit Jacob et al. gemmlowp: a small self-contained low-precision gemm library.(2017), 2017.

[20] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018.

[21] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[22] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

[23] Rundong Li, Feng Liang, Hongwei Qin, Yan Wang, Rui Fan, and Junjie Yan. Fully quantized network for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016.

[27] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang. Face model compression by distilling knowledge from neurons. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[28] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus

benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[29] Jeffrey L McKinstry, Steven K Esser, Rathinakumar Appuswamy, Deepika Bablani, John V Arthur, Izzet B Yildiz, and Dharmendra S Modha. Discovering low-precision networks close to full-precision networks for efficient embedded inference. *arXiv preprint arXiv:1809.04191*, 2018.

[30] Szymon Migacz. 8-bit inference with tensorrt. In *GPU technology conference*, volume 2, page 7, 2017.

[31] Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *arXiv preprint arXiv:1711.05852*, 2017.

[32] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347. IEEE, 2014.

[33] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016.

[34] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[35] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[36] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, pages 525–542. Springer, 2016.

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016.

[41] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[42] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.

[43] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[44] Frederick Tung and Greg Mori. Clip-q: Deep network compression learning by in-parallel pruning-quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7873–7882, 2018.

[45] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[46] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[47] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2593–2601, 2017.

[48] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8612–8620, 2019.

[49] Peisong Wang, Qinghao Hu, Yifan Zhang, Chunjie Zhang, Yang Liu, and Jian Cheng. Two-step quantization for low-bit neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4376–4384, 2018.

[50] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[51] Lior Wolf, Tal Hassner, and Itay Maoz. *Face recognition in unconstrained videos with matched background similarity*. IEEE, 2011.

[52] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[53] Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, and Gang Hua. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–382, 2018.

[54] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019.

[55] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.

[56] Shu-Chang Zhou, Yu-Zhi Wang, He Wen, Qin-Yao He, and Yu-Heng Zou. Balanced quantization: An effective and efficient approach to quantized neural networks. *Journal of Computer Science and Technology*, 32(4):667–682, 2017.

[57] Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.

[58] Xiaotian Zhu, Wengang Zhou, and Houqiang Li. Adaptive layerwise quantization for deep neural network compression. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.