

A Quality Aware Sample-to-Sample Comparison for Face Recognition

Mohammad Saeed Ebrahimi Saadabadi, Sahar Rahimi Malakshan,
Ali Zafari, Moktari Mostofa, and Nasser M. Nasrabadi

me00018, sr00033, az00004, mm0251@mix.wvu.edu, nasser.nasrabadi@mail.wvu.edu

Abstract

Currently available face datasets mainly consist of a large number of high-quality and a small number of low-quality samples. As a result, a Face Recognition (FR) network fails to learn the distribution of low-quality samples since they are less frequent during training (underrepresented). Moreover, current state-of-the-art FR training paradigms are based on the sample-to-center comparison (i.e., Softmax-based classifier), which results in a lack of uniformity between train and test metrics. This work integrates a quality-aware learning process at the sample level into the classification training paradigm (QAFace). In this regard, Softmax centers are adaptively guided to pay more attention to low-quality samples by using a quality-aware function. Accordingly, QAFace adds a quality-based adjustment to the updating procedure of the Softmax-based classifier to improve the performance on the underrepresented low-quality samples. Our method adaptively finds and assigns more attention to the recognizable low-quality samples in the training datasets. In addition, QAFace ignores the unrecognizable low-quality samples using the feature magnitude as a proxy for quality. As a result, QAFace prevents class centers from getting distracted from the optimal direction. The proposed method is superior to the state-of-the-art algorithms in extensive experimental results on the CFP-FP, LFW, CPLFW, CALFW, AgeDB, IJB-B, and IJB-C datasets.

1. Introduction

Recent advances in FR performance can be credited to introduction of novel network architectures, large-scale datasets, and new loss functions [23]. Regarding the architecture, ResNet and its variants are mostly used as the backbone for extracting features from the face images [11]. In terms of datasets, large-scale publicly available training data leads to unprecedented improvement in FR performance [13]. Recent attempts on FR are mainly focused on manipulating the training criteria [25, 40, 11, 32, 39, 42]. In this manner, Softmax with a cross-entropy loss, i.e.,

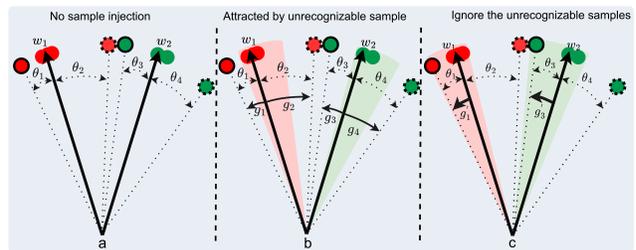


Figure 1. A binary classification example illustrates that unrecognizable samples can misguide the Softmax centers, w_1 and w_2 , from their optimal direction. Circles with solid black borders are recognizable low-quality samples, and black dashed borders are unrecognizable samples. g_i shows the direction that the centers are being pushed. a) without injection, there is no g . b) equally injecting samples results in stronger g from samples with more angular disparity, and c) injecting with emphasis on recognizable low-quality samples and ignoring unrecognizable (there is no g toward unrecognizable). Shaded areas are the direction in which feature injection causes the centers to move. Note that $g_1 < g_1'$ and $g_2 < g_2'$.

sample-to-center comparison, is the most popular criterion for FR, (i.e., classification) [25]. In the classification framework, the weights connecting the penultimate layer output (i.e., feature) to the classification layer represent the centers of Softmax classes [2].

Since FR is an open-set problem, during testing, sample-to-center (dis)similarity of the Softmax is irrelevant and sample-to-sample (dis)similarity matters. In order to unify the train and test similarity metrics, pioneering works [32, 43], devised metric-based loss functions based on sample-to-sample comparison [37, 43]. These metric-based losses try to directly minimize a distance metric when two samples come from the same identity (positive pair); otherwise (negative pair) impose a margin [32]. However, the sample-wise comparison highly depends on the pair selection strategy and requires a sophisticated mining method [22]. Besides, in large-scale datasets [11] with thousands of identities and millions of samples, there is a combinatorial explosion in the number of possible pairs, leading to costly pair selection, unstable training, and slow convergence [36].

Several studies show that projecting features and class

centers to the unit-hypersphere improves the discriminative power of representation learned by Softmax [40, 39, 25]. In this manner, Softmax classifies images using the angular distance between the feature representations and Softmax centers [39]. An angular margin is then integrated to the Softmax loss to further enforce intra-class compactness and inter-class separability [25, 40, 23, 46]. The angular-based Softmax loss functions are more stable than metric-based, i.e., no pair selection, and the number of centers is much fewer than the number of samples [24]. As a result, angular-based Softmax losses have become the state-of-the-art method for training FR frameworks [11, 28].

In sample-to-center training paradigms, every identity is represented as a deterministic point in high-dimensional latent space [6], i.e., centers. As a result, their performance degrades when there is a large disparity between training and testing data [34]. Although augmentation may narrow the gap between training and evaluation data distribution, it increases the occurrence of unrecognizable samples and overfitting [26], as illustrated in Fig. 2. To alleviate this problem, authors in [28] propose to use feature magnitude as a proxy to measure the image quality. Also, there are methods to estimate the distribution of each class instead of presenting them as a single deterministic point [6, 34]. Despite the performance improvement, these approaches do not propose a solution to the overfitting problem nor guarantee (dis)similarity between (negative)positive samples.

Usually, there are three different types of samples in FR datasets [28]. First, samples that are easy to learn by the model. These easy samples usually have good quality-related factors such as high resolution [28]. Second, images that have low-quality, but recognizable (hard-samples) [7]. Third, unrecognizable samples that even humans can not correctly recognize their identity [28, 7]. As shown in Fig. 1, unrecognizable samples have larger angular disparity, θ_2 and θ_4 , compared to low-quality instances, $\theta_2 > \theta_1$ and $\theta_4 > \theta_3$. The primary idea in recent works is to increase the margin constraints as the angular disparity between sample and the Softmax center increases [46, 23, 18]. However, the recognizability of samples and sample-wise (dis)similarity are not considered by any of the mentioned methods. Therefore, the model tries to reduce the training loss by overfitting on unrecognizable samples, which harms model generalization [13].

Recently, authors in [12] integrate sample-to-sample comparison to the Softmax via injecting sample representations to the centers. Although VPL [12] brings sample-to-sample comparison to the Softmax framework, the prior assumption is that the unrecognizable samples do not distract the learning. Due to the larger angular disparity, injecting without considering the recognizability of samples puts more emphasis on unrecognizable samples during injection, see Fig. 1. Adding variations toward the unrec-

ognizable samples harms the model learning paradigm and distract the Softmax centers from the optimal direction. The injection process directly changes centers. Therefore, it is important to push the centers to a valid direction.

Sample selection strategy is indispensable in every sample-wise FR training paradigm [32, 23]. In this work, we try to weigh samples based on their recognizability and quality. In this manner, the sample-wise part of the proposed method (QAFace), injection, benefits from recognizable low-quality (hard) samples. During training, QAFace effectively ignores the unrecognizable samples and prevents the class centers from being distracted from the optimal direction. At the same time, QAFace emphasizes on low-quality samples considering them as hard samples. Moreover, since high-quality samples are being well explored by sample-to-center part of the training, the proposed method puts less attention on high-quality samples during their injection. Compared to [12], our method adds no extra memory consumption and sampling strategy to the training. The presented model queues sample representation using MOCO [16] to maintain both samples and centers on the same embedding space. Contributions of this work can be summarized as follows:

- we use informative hard samples (low-quality instances) to introduce sample-wise comparison to the angular-margin Softmax loss.
- we propose a new quality-based weighting function that can effectively de-emphasize unrecognizable samples based on the magnitude of their feature representation as a proxy of image quality.
- we leverage hard samples to add uncertainty to the Softmax centers toward the direction of hard samples.

2. Related Works

2.1. FR Loss Functions

Most of the previous FR methods were established on a metric-learning loss function, such as triplet [32] or contrastive loss [8, 30]. These loss functions were based on sample-to-sample comparison in Euclidean space. Then [43] enhanced the intra-class similarity via proposing a new loss to directly minimize intra-class distance while doing classification. In this manner, the main challenges toward general FR were the necessity of sample mining, lack of generalization, and feature collapsing problem [40, 32, 15]. More recently, studies showed that applying Softmax to the angular space enhances the discriminability of features [25, 40, 11, 43]. Consequently, pioneering works of [39, 40, 11], introduced intuitive loss functions by applying three different types of margin to the angular space of Softmax: 1) multiplicative angular margin, 2) cosine margin,



Figure 2. Same Augmentation, i.e., down-sampling and random cropping, results in different recognizability among samples. Green border shows the recognizable samples, and red shows unrecognizable samples.

and 3) additive angular margin, resulting to state-of-the-art performance.

2.2. Angular Margin Variations

Recent studies explore the effects of adaptive angular margin on the learning paradigm of the network [46, 23, 23]. Liu *et al.* [23] propose to adaptively tune the margin value to put more constraint on the tail classes. The role of negative samples in obtaining more discriminative features is investigated in [18]. Authors of [14], add new term to the angular-margin loss function to supervise the uniformly spreading of class centers on the unit hyper-sphere. MagFace [28] establishes the norm of features as a proxy of sample recognizability. Image recognizability increases as the feature norm increase [28]. MagFace assigns high angular margin on the high-norm feature in the premises of pushing those samples to be closer to their class center. The drawback is that it fails to put emphasis on the valuable hard samples. Furthermore, none of the mentioned methods guarantee the samples-wise similarity. Also, representing each identity with a single deterministic point, i.e., center, in high-dimensional space results in the performance drop when testing data has a large disparity with training samples [34].

2.3. Probabilistic Face Modeling

Probabilistic face modeling is well-established in face template/video matching [5, 3]. In these works, a series of samples is used as the input rather than a single face image. Shi *et al.* for the first time integrate uncertainty into a single image FR [34]. PFE [34] represents each image as a Gaussian distribution. The mean and variance of the Gaussian reflect the “most likely latent feature”, and “the uncertainty in the feature values”, respectively [34]. The goal is to add uncertainty to the model to boost performance for unseen data with large disparity [6]. Instead of adding uncertainty to each image representation, VPL [12] assigns a distribution to each class within the classification framework. Specifically, VPL injects the class instances to the corresponding classifier to bring more uncertainty to centers and, at the same time, integrate sample-to-sample comparison to the classification paradigm. However, it fails to consider the image recognizability measure. Considering the

whole memorizing process in [12], projecting all the representations to the unit hyper-sphere results in an equal contribution of different instances in the memory. Therefore, because of large angular disparity with centers, unrecognizable samples distract centers from their optimal direction, see Fig. 1 (b).

3. Proposed Method

In this section, we begin by analyzing the Softmax-based loss functions. Then, we further explain the integration of Softmax-based classifier with the sample-to-sample comparison. We devise a new injection function to integrate a quality-aware sample-to-sample comparison to the classification framework (QAFace). Finally, we investigate the capability of our method to ignore the unrecognizable samples and the complementary role of our quality-aware injection to the learning signal of Softmax-based loss function.

3.1. Preliminaries

Most of the deep visual recognition modules, including FR, can be regarded as the stack of non-linear feature extractor layers (backbone), together with a classifier which is usually a Softmax layer [4]. Both the backbone and classifier will be trained end-to-end using a back-propagation algorithm. The Softmax training criterion can be formulated as follows [1, 20]:

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{e^{W_{y_i}^T x_i + b_{y_i}} + \sum_{j=1, j \neq y_i}^C e^{W_j^T x_i + b_j}}, \quad (1)$$

where $W_j \in \mathbb{R}^d$ is j -th classifier (center), d is the feature dimension, and b_j is the bias for j -th Softmax output. x_i is the learned representation of i -th sample, and y_i is its corresponding ground truth. N and C represent the mini-batch size and the total number of classes, respectively.

The angular distribution of representations learned via the Softmax loss, x_i , suggests using cosine distance as the metric rather than Euclidean distance [39]. Hence, a modified Softmax loss was defined by projecting both centers and representations to the unit-hypersphere [39, 25], $\|W_j\| = \|x_i\| = 1$ and $b_j = 0$.

$$L' = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i}))}}{e^{s(\cos(\theta_{y_i}))} + \sum_{j=1, j \neq y_i}^C e^{s(\cos(\theta_j))}}, \quad (2)$$

where $\cos(\theta_{y_i})$ reflects the cosine similarity between x_i and w_{y_i} and $\cos(\theta_j)$ denote similarity between x_i and w_j (negative centers). s is introduced as the scaling hyper-parameter which affects the curves of the output [46], see Fig. 6. In common FR practice, biases are removed from Eq. 1 because they are learned for close-set recognition and cannot be generalized to open-set testing.

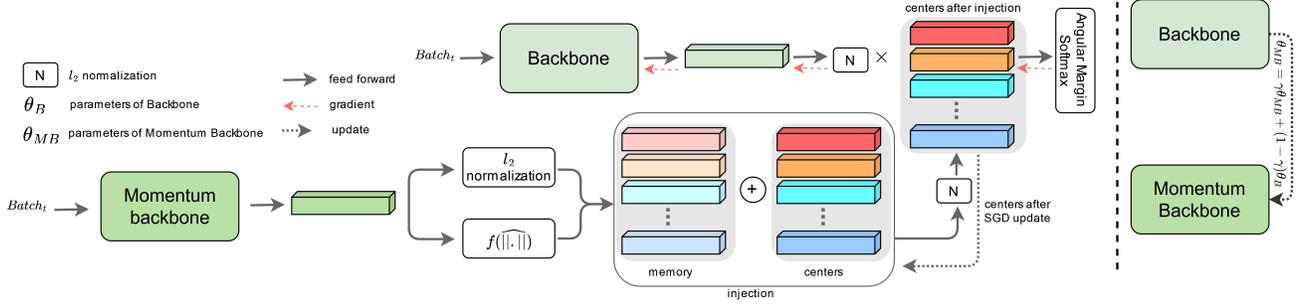


Figure 3. Left: the general architecture of the proposed method. In each iteration, the class centers represent the centers and accumulated features of the hard samples of previous iterations. The representations for injecting to the centers are obtained from the momentum backbone. Right: shows the updating of the Momentum Backbone parameters.

To enhance the intra-class compactness and inter-class separability, authors in [39, 40, 11] developed intuitive loss functions by applying three different types of margins to Eq. 2.

$$L'' = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(m_S \theta_{y_i} + m_A) - m_C)}}{e^{s(\cos(m_S \theta_{y_i} + m_A) - m_C)} + \sum_{\substack{j=1 \\ j \neq y_i}}^C e^{s \cos(\theta_j)}}, \quad (3)$$

SphereFace [25] introduced the multiplicative angular margin to modify decision boundaries from $\cos(\theta_1) = \cos(\theta_2)$ to $\cos(m_S \theta_1) = \cos(\theta_2)$. Where $(\theta_1) \theta_2$ represent angle between x_i and $(W_{y_i}) W_j$. Their modification to Softmax does improve the result; however, the proposed loss function is computed through a series of approximations, which results in unstable training [40]. CosFace modified the decision boundary to $\cos(\theta_1) + m_C = \cos(\theta_2)$ and ArcFace changed it to $\cos(\theta_1 + m_A) = \cos(\theta_2)$. Eq. 3 represents all mentioned modifications. Where m_S , m_C , and m_A are margins introduced by SphereFace [25], CosFace [40], and ArcFace [11], respectively. Despite the remarkable improvement, the sample-wise (dis)similarity is not considered in any of these modifications.

Also, presenting each identity with a single deterministic point in embedding space results in performance degradation when there is a significant disparity between training and testing data [34]. For better illustration, we experiment by manually degrading five high-quality testing datasets. Comparing the results of Arcface and VPL in Table 1, less performance gap in VPL shows that adding uncertainty to the Softmax centers results in better handling of quality disparity between train and test datasets. Comparing the results of QAface with VPL, it is shown that our proposed method can further reduce the gap between representation of high and low quality samples by putting more emphasise on the low-quality samples during the injection.

3.2. Classification-Based Gradient

We can divide a Softmax-based FR method into its backbone and classifier components. Hence, here we study the updating of the backbone and center of Softmax separately.

For the backbone, we show the gradient with regard to its output (feature), i.e., $\frac{\partial L}{\partial x_i}$. By omitting bias in the Eq. 1 the derivatives to j -th class center and i -th sample's feature are:

$$\frac{\partial L}{\partial x_i} = ((p_{i, y_i} - 1) W_{y_i}) + \sum_{\substack{j=1 \\ j \neq y_i}}^C p_{i, j} W_j, \quad (4)$$

$$\frac{\partial L}{\partial W_j} = \sum_{\substack{i=1 \\ y_i = j}}^N ((p_{i, y_i} - 1) x_i) + \sum_{\substack{i=1 \\ y_i \neq j}}^N p_{i, j} x_i, \quad (5)$$

where $p_{i, j} = \frac{e^{W_j^T x_i}}{\sum_{j=1}^C e^{W_j^T x_i}}$. Eq. 4 shows that from the backbone perspective, the network is being updated toward increasing the similarity between features and the positive class centers while decreasing similarity with negative centers. Moreover, Eq. 5 demonstrates that centers update toward being more similar to their corresponding class instances and away from samples of other classes. Hence, both backbone and centers are moving toward each other, and sample-wise (dis)similarity is being supervised indirectly.

3.3. Sample-Wise Similarity with Softmax

In order to directly supervise sample-wise (dis)similarity, [12] injects samples feature to their corresponding class center. To this end, a memory, M , is constructed, which memorizes the positive features of each class. The memory has the same shape as the Softmax centers: $W \in \mathbb{R}^{C \times d}$, and $M \in \mathbb{R}^{C \times d}$. Considering the injection process as: $\tilde{W}_{y_i} = W_{y_i} + \lambda M_{y_i}$, the derivative with regard to features changes to:

$$\frac{\partial L}{\partial x_i} = ((p_{i, y_i} - 1)(W_{y_i} + \lambda M_{y_i})) + \sum_{\substack{j=1 \\ j \neq y_i}}^C p_{i, j} (W_j + \lambda M_j), \quad (6)$$

here, the memorized features, M , affects the gradient that is updating the backbone. Therefore, sample-to-sample (dis)similarity is being directly supervised. λ is a hyper-parameter to adjust the amount of the injection and should

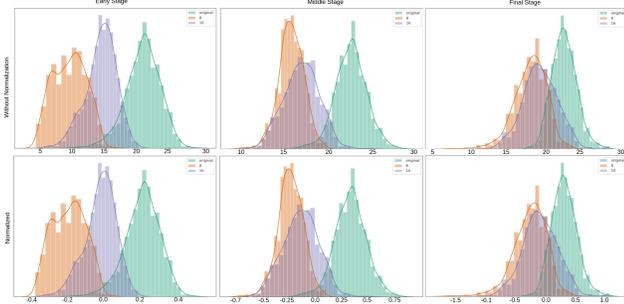


Figure 4. Histogram of the magnitude of features obtained from 10k randomly selected training images and their down-sampled version. Early stage: mean over histogram of epochs one to four. Middle stage: mean over histogram of epochs 10 to 15. Final stage: mean over histogram of epochs 20 to 24. Top (before) and bottom (after) applying Eq. 10.

be set manually. In this injection manner, all the representations are projected to the unit hyper-sphere first. Therefore, unrecognizable samples which have considerable angular disparity with Softmax centers have more influence on the centers than other samples. Therefore, centers will be distracted from the optimal direction. On the other hand, high-quality samples have high similarity with class centers and do not contribute to adding beneficiary variation to the centers.

3.4. Quality Aware Sample Injection

To address the mentioned shortcomings of the classification framework, prevent Softmax centers from being distracted, and explore recognizable low-quality samples, we propose QAFace, a quality-aware injection procedure. Proposed method ignores unrecognizable samples and, at the same time, uses recognizable low-quality samples to add more valid uncertainty to the centers. The injection process is as follows:

$$W_{y_i} = W_{y_i} + f(\widehat{\|x_i\|}) * \frac{x_i}{\|x_i\|}, \quad (7)$$

where $\widehat{\|x_i\|}$ is the normalized version of feature magnitude. We normalize the feature magnitude via batch statistics: μ and σ . To relax μ and σ from the batch size, we calculate them in an exponential moving average over the training iterations. $f(\widehat{\|x_i\|})$ projects $\widehat{\|x_i\|}$ to have zero value for the features that have $\widehat{\|x_i\|}$ lower than a threshold, $-\tau$, otherwise positive.

$$\sigma_t = \alpha\sigma_t + (1 - \alpha)\sigma_{t-1}, \quad (8)$$

$$\mu_t = \alpha\mu_t + (1 - \alpha)\mu_{t-1}, \quad (9)$$

$$\widehat{\|x_i\|} = \frac{\|x_i\| - (\mu)}{\sigma}, \quad (10)$$



Figure 5. Illustration of three types of samples with regard to the feature norm at the final stage of training. Left: samples that model ignored. Middle: Samples that are being emphasised. Right: Samples with high feature norm.

Table 1. Performance (%) of Arcface, VPL, and our method on the different down-sampled versions of LFW, CFP-FP, CALFW, CPLFW, AgeDB. 1:1 verification accuracy is reported.

	Resolution	LFW	CFP-FP	CPLFW	CALFW	AgeDB
ArcFace	8×8	71.86	56.92	56.43	57.56	54.48
	16×16	96.60	84.21	82.76	84.30	78.20
	original	99.83	98.27	92.08	95.45	98.28
VPL	8×8	71.96	60.75	57.78	59.56	52.45
	16×16	97.30	85.98	83.53	84.61	79.06
	original	99.83	99.11	93.45	96.12	98.60
QAFace	8×8	72.76	59.62	57.65	59.93	54.16
	16×16	98.26	89.57	86.75	88.20	83.56
	original	99.85	99.21	94.41	96.11	97.91

$$f(\widehat{\|x_i\|}) = \begin{cases} e^{-\widehat{\|x_i\|}} & \text{if } \widehat{\|x_i\|} \geq -\tau, \\ 0 & \text{else.} \end{cases} \quad (11)$$

Eqs. 10 and 11 together work in a way that 1) samples with $\widehat{\|x_i\|}$ lower than $-\tau$ would not affect the centers, 2) recognizable but low-quality samples will be emphasized during training, and 3) high-quality samples will receive less attention in comparison to recognizable low-quality samples. Using informative samples in metric-based FR training paradigm has been well-established [32]. Therefore, employing informative samples to add sample-wise comparison to the classification framework is of most importance. Our proposed algorithm adaptively: 1) assigns more weight to the recognizable low-quality (hard) samples, 2) ignores unrecognizable samples, and 3) puts less attention on easy high-quality samples. Hence, our approach can be regarded as a kind of hard-sample mining, but without adding any computational burden on hard sample selection. It is worth mentioning that, $f(\widehat{\|x_i\|}) * \frac{x_i}{\|x_i\|}$ is happening during the memorizing the representations in the memory. Consequently, we can re-write the Eq. 10 as:

$$W_{y_i} = W_{y_i} + M(\widehat{\|x_i\|}, x_i). \quad (12)$$

3.5. Distractor Samples

The major advantage of the QAFace over [12] is the ability to identify the unrecognizable from recognizable samples and emphasize the recognizable low-quality samples

Table 2. Performance (%) comparison of our method with other recent algorithms. 1:1 verification accuracy is reported on LFW, CFP-FP, CPLFW, AgeDB.

Method	Venue	Verification accuracy					TAR@FAR=1e-4	
		LFW	CFP-FP	CPLFW	CALFW	AgeDB	IJB-B	IJB-C
Wang <i>et al.</i> [40]	CVPR18	99.81	98.12	92.28	95.76	98.11	94.80	96.37
Deng <i>et al.</i> [11]	CVPR19	99.83	98.27	92.08	95.45	98.28	94.25	96.03
Sun <i>et al.</i> [38]	CVPR20	99.73	96.02	-	-	-	-	93.95
Deng <i>et al.</i> [9]	ECCV20	99.80	98.80	-	-	98.31	94.94	96.28
Wang <i>et al.</i> [41]	AAAI20	99.80	98.28	92.83	97.95	96.10	93.6	95.2
Huang <i>et al.</i> [18]	CVPR20	99.80	98.37	93.13	96.20	98.32	94.8	96.1
Kim <i>et al.</i> [22]	ECCV20	99.85	98.63	93.17	96.20	98.38	94.97	96.38
Shi <i>et al.</i> [35]	CVPR20	99.78	98.64	-	-	-	-	96.6
Kim <i>et al.</i> [21]	CVPR20	99.85	98.63	93.17	96.20	98.28	94.93	96.26
Chang <i>et al.</i> [6]	CVPR20	99.83	98.78	-	-	-	-	94.61
Meng <i>et al.</i> [28]	CVPR21	99.83	98.46	92.87	96.15	98.17	94.51	95.97
Deng <i>et al.</i> [12]	CVPR21	99.83	99.11	93.45	96.12	98.60	95.56	96.76
QAface		99.85	99.21	94.41	96.11	97.91	95.67	97.20

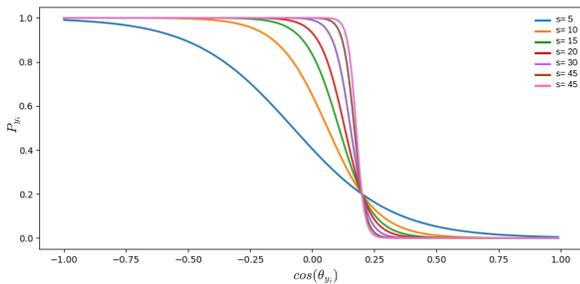


Figure 6. Curve of $p_{i,j} = \frac{e^{W_j^T x_i}}{\sum_{j=1}^C e^{W_j^T x_i}}$ when $\cos(\theta_j)$ is fixed and $\cos(\theta_{y_i})$ changes from -1.0 to 1.0.

in the injection process. To this end, we employ the magnitude of the feature vector as a proxy for the input sample recognizability [28]. We perform an experiment to demonstrate how the feature magnitude is affected by recognizability and how our method can use hard samples to reduce the gap between the representation of low and high-quality samples. We randomly select a subset of 10K images of the training data and down-sampled them to the different levels (8×8 , and 16×16). Then, at the end of every epoch, we save the magnitude of the representations obtained from original and down-sampled images, see Fig. 4. From the results shown in Fig. 4 (top) we can observe that as the recognizability increases, the magnitude of representations increases as well, i.e., *green* > *blue* > *orange*.

Complete overlapping of the distribution of down-sampled with the original instances is an ideal scenario, which means that the model became quality-agnostic. The model progressively learns to increase the lower bond of feature magnitude to narrow the gap between original and low-quality samples. At the early stages of the training, the full range of feature magnitude is around 25 (from 5 to 30). Then in the final stages, the range narrows to 15 (from 15

to 30). Also, the mean of the original image distribution (green) is always around 23, which shows that our proposed method can effectively involve low-quality samples in training without reducing the performance on the high-quality samples.

Another observation from Fig. 4 (bottom) is the necessity of normalizing the magnitude. Without applying Eq. 10 on the feature magnitude, the threshold for ignoring unrecognizable samples changes as the training progresses. Eq. 10 omits the bias from the distribution of features magnitude caused by the training stage. Therefore, we can choose a fixed τ for ignoring the unrecognizable and emphasizing the hard samples.

3.6. Complement to Angular-Margin Gradient

Unlike triplet and contrastive losses, softmax-based losses are not subject to explicit easy/hard sample mining [42, 31]. In this section, using a simple toy example, we show that the Softmax-based losses implicitly benefit easy/hard sample mining by their gradient. Additionally, we elaborate on the ability of the proposed $f(\widehat{\|x_i\|})$ to complement the Softmax learning signal (gradient). Consider a four-identity classification. For a given sample x_i with ground truth identity $y_i = 4$, the logits are $\cos(\theta_1), \cos(\theta_2), \cos(\theta_3), \cos(\theta_4)$. In Fig. 7, we plot the loss value for a fixed $\cos(\theta_j), j \neq 4$ while $\cos(\theta_{y_i=4})$ changes from -1 to 1. The first observation from Fig. 7 (right) is that the scaling parameter s is tuning the sensitivity of the loss function. As the scaling value increases, the slope of the loss function (gradient) increases [46].

Moreover, s directly influences the point that samples would be recognized as easy. Easy samples would barely experience change, i.e., low slope, while hard samples receive high gradient value, i.e., high slope. Also, in Fig. 7 (left), we demonstrate that the drawback here is the mono-

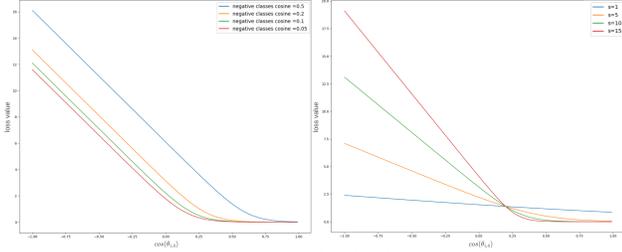


Figure 7. Curves of $p_{y_i} = \frac{e^{s(\cos(\theta_{y_i}))}}{\sum_{j=1}^C e^{s(\cos(\theta_j))}}$, when $y_i = 4$ and $\cos(\theta_{y_i})$ changes from -1 to 1.

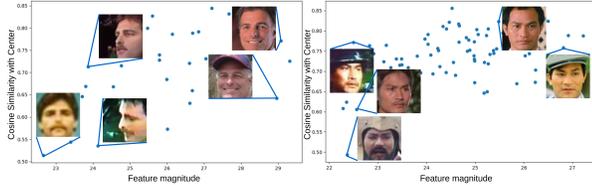


Figure 8. Illustration of sample-to-center cosine similarity for two randomly selected subjects. High-norm samples are very similar to the class center. Low-norm samples have lower similarity with Softmax center.

tonicity of the gradient among the hard samples. In other words, some samples are unrecognizable; however, their gradient is equivalent to those of informative but hard samples. Consequently, the model tries to overfit the unrecognizable samples because there is no identity information on those instances [35]. Our proposed method tries to compensate for this effect by ignoring the unrecognizable samples during the injection. It does not further involve the unrecognizable samples in the injection process and ignores them using the proposed feature weighting paradigm. As a result, our method justifies the classifier direction to tolerate more variation toward hard and informative samples and plays a complementary role to the Softmax-based learning signal.

4. Experiments

4.1. Datasets

We employ Webface4M [49] as our training data, which contain 4 million samples of around 200,000 identities, Table 2. For evaluation of our method, we use CFP-FP [33], CPLFW [47], CALFW [48], LFW [17], AgeDB [29], IJB-B [44], and IJB-C [27]. Based on the datasets evaluation protocols, we report 1:1 verification accuracy for CFP-FP, LFW, CPLFW, CALFW, and AgeDB datasets. For IJB-B [44] and IJB-C [27], we report the True Acceptance Rate (TAR) over the False Acceptance Rate of $1e-4$.

4.2. Training Settings

We use [10] to detect five landmarks in each image. Then images are aligned and rescaled to 112×112 , following the setting in [11]. We adopt ResNet [11] for the backbone. The model is trained for 24 epochs with Arcface loss. The

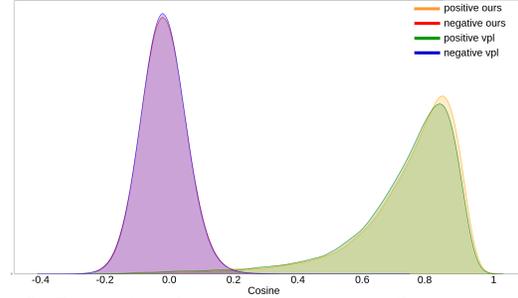


Figure 9. Comparison between pair-wise similarity score on the IJB-C dataset obtained from VPL and QAface.

Table 3. Ablation of Δt . The metrics are the same as Table 2.

Δt	Verification Accuracy			TAR@FAR: $1e-4$	
	LFW	CFP-FP	CPLFW	IJB-B	IJB-C
0	99.71	98.40	92.01	95.26	96.4
500	99.80	98.81	92.84	95.46	96.75
1000	99.85	99.21	94.41	95.67	97.20
1500	99.78	99.01	93.12	95.45	96.87
2000	99.69	98.33	92.95	95.14	96.35

optimizer is SGD, with the learning rate starting from 0.1, which is decreased by a factor of 10 at epochs $\{10, 16, 22\}$. The optimizer weight-decay is set to 0.0001, and the momentum is 0.9. During training, the mini-batch size on each GPU is 512, and the model is trained using two Quadro RTX 8000. Following [16], γ in Fig. 3 is 0.99. In calculating the μ and σ , α in Eq. 8 and 9, is 0.99. Given a pair of images, the cosine distance between the representations is the metric during inference.

4.3. Ablation Study

4.3.1 Impact of the Memory Length

In [16], the memory is a dynamic queue of representations. The whole memory has a length of $|M|$, and queuing new samples results in de-queuing the oldest samples. Here the length of memory is equal to the number of classes. Therefore, we should memorize the last iteration in that every instance in the memory was updated. In this way, we can prevent from employing outdated representations in the injection. For instance, if the training is on the iteration I and a specific instance in memory was updated on $I - \Delta t$, if Δt is larger than a threshold, that specific memory instance cannot be used during the injection. It is shown that in the early epochs, the changes in the feature space are drastic; after that, it is negligible [12]. Therefore, we start the injection after the fourth epoch of training. Table 3 shows the ablation experiments on Δt . In these experiments, we fixed τ to 2. We increase Δt with the interval of 500 iterations. As illustrated in Table 3, the performance constantly improves from $\Delta t = 0$ to $\Delta t = 1000$, and after that, it starts to degrade.

Table 4. Ablation of τ . The metrics are the same as Table 2.

τ	Verification Accuracy			TAR@FAR:1e-5	
	LFW	CFP-FP	CPLFW	IJB-B	IJB-C
0	99.86	99.23	93.20	95.03	96.12
1	99.83	99.11	93.14	95.41	96.91
2	99.85	99.21	94.41	95.67	97.20
3	99.83	99.02	93.02	95.51	96.84
4	98.76	98.45	92.32	95.02	96.20

Table 5. Ablation of augmentation probability, TAR@FAR=1e-4.

probability	cropping	down-sampling	IJB-B	IJB-C
0.0	-	-	95.34	96.60
0.1	-	✓	95.56	96.89
0.1	✓	-	95.41	96.65
0.1	✓	✓	95.57	96.95
0.2	-	✓	95.63	96.91
0.2	✓	-	95.45	96.67
0.2	✓	✓	95.67	97.20

4.3.2 Impact of Threshold (τ)

We fix the memory length to 1000 iterations. Then we investigate different values for the threshold (τ) in Eq. 10. As illustrated in Table 4, the performance on IJB-B and IJB-C constantly increases with changing τ from zero to 2. At $\tau = 0$, only samples with $\|x\|$ above the zero are involved in the injection. Consequently, the model performance decreases when the input comes from datasets like IJB-B and IJB-C, which contain low-quality samples [35]. On the other hand, the results in clean datasets like CFP, CPLFW, and LFW are reasonably good [35].

4.4. Impact of Augmentation

For data augmentations, we used random cropping and down-sampling [19, 45]. On-the-fly data augmentation provides more diverse training data. However, as illustrated in Fig. 2, it increases the occurrence of unrecognizable samples. We perform experiments on our method with and without the presence of data augmentation. Accordingly, we can show that our method can effectively ignore unrecognizable samples and, at the same time, benefits from more training instances, see Fig. 5. As shown in Table 5, the model gains performance on the IJB-B and IJB-C datasets by increasing the probability of augmentations. As these datasets contain low-quality samples, down-sampling leads to more performance improvement than random cropping.

4.5. Comparison with state-of-the-art

Table 2 shows the proposed method’s performance compared to the state-of-the-art algorithms. For better clarification, we explain our observation in two parts. For the results on LFW, CPLFW, CALFW, CFP-FP, and AgeDB, it is essential to mention that QAFace is built upon putting more emphasis on the low-quality samples and making

these samples’ representation more similar to the high-quality samples’ features. Consequently, the performance gain in these datasets is marginal, as they contain almost high-quality samples [35]. Although the performance is saturated in most of these datasets, our method strives to increase the 1:1 verification accuracy for the CFP-FP and CPLFW datasets. The IJB-B and IJB-C datasets are more challenging and have images/frames with diverse quality. Results on the IJB-B and IJB-C datasets show the superiority of our approach in more general face recognition. As these datasets contain low-quality and high-quality images, the performance gain in these datasets is more evident. In IJB-B, compared to VPL, QAFace improves the TAR at FAR=1e-4.

5. Conclusion

This work argues the importance of integrating sample-wise similarity to the Softmax framework. Also, we showed that existing angular-margin-based loss functions could be distracted by the unrecognizable samples in the dataset. Inspired by the well-established idea of hard sample mining in the sample-to-sample comparison framework, we proposed a weighting scenario to ignore unrecognizable samples and emphasize recognizable low-quality samples during the injection. We empirically showed the effect of ignoring unrecognizable samples by improving the similarity score between positive samples in the IJB-C dataset. Also, We analyzed the proposed function for weighting. Our proposed approach is based on the simple idea of using the norm of features as the proxy for the recognizability of face images. Furthermore, we empirically showed the effect of the quality of face images on the magnitude of features. We demonstrated that there is a direct proportion between the face image quality and the magnitude of its representation. Our approach could successfully outperform all of its competitors in five out of seven evaluation benchmarks, including the IJB-B and IJB-C datasets.

References

- [1] Poorya Aghdaie, Baaria Chaudhary, Sobhan Soleymani, Jeremy Dawson, and Nasser M Nasrabadi. Morph detection enhanced by structured group sparsity. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 311–320, 2022.
- [2] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4051, 2022.
- [3] Ognjen Arandjelovic, Gregory Shakhnarovich, John Fisher, Roberto Cipolla, and Trevor Darrell. Face recognition with image sets using manifold density divergence. In *2005*

- IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 581–588. IEEE, 2005.
- [4] Fadi Boutros, Naser Damer, Jan Niklas Kolf, Kiran Raja, Florian Kirchbuchner, Raghavendra Ramachandra, Arjan Kuijper, Pengcheng Fang, Chao Zhang, Fei Wang, et al. Mfr 2021: Masked face recognition competition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2021.
- [5] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2567–2573. IEEE, 2010.
- [6] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020.
- [7] Kai Chen, Taihe Yi, and Qi Lv. Lightqnet: Lightweight deep face quality assessment for risk-controlled face recognition. *IEEE Signal Processing Letters*, 28:1878–1882, 2021.
- [8] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.
- [9] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer, 2020.
- [10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotzia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.
- [11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [12] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11906–11915, 2021.
- [13] Hang Du, Hailin Shi, Yuchi Liu, Jun Wang, Zhen Lei, Dan Zeng, and Tao Mei. Semi-siamese training for shallow face learning. In *European Conference on Computer Vision*, pages 36–53. Springer, 2020.
- [14] Yueqi Duan, Jiwen Lu, and Jie Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2019.
- [15] Atiye Sadat Hashemi, Andreas Bär, Saeed Mozaffari, and Tim Fingscheidt. Improving transferability of generated universal adversarial perturbations for image classification and segmentation. In *Deep Neural Networks and Data for Automated Driving*, pages 171–196. Springer, Cham, 2022.
- [16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008.
- [18] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.
- [19] Amol S Joshi, Ali Dabouei, Jeremy Dawson, and Nasser M Nasrabadi. Fdeblur-gan: Fingerprint deblurring using generative adversarial network. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.
- [20] Hossein Kashiani, Shoaib Meraj Sami, Sobhan Soleymani, and Nasser M Nasrabadi. Robust ensemble morph detection with domain generalization. *arXiv preprint arXiv:2209.08130*, 2022.
- [21] Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, and Jongju Shin. Groupface: Learning latent groups and constructing group-based representations for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5621–5630, 2020.
- [22] Yonghyun Kim, Wonpyo Park, and Jongju Shin. Broadface: Looking at tens of thousands of people at once for face recognition. In *European Conference on Computer Vision*, pages 536–552. Springer, 2020.
- [23] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11947–11956, 2019.
- [24] Jiaheng Liu, Haoyu Qin, Yichao Wu, and Ding Liang. Anchorface: Boosting tar@ far for practical face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
- [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] Jiang-Jing Lv, Xiao-Hu Shao, Jia-Shui Huang, Xiang-Dong Zhou, and Xi Zhou. Data augmentation for face recognition. *Neurocomputing*, 230:184–196, 2017.
- [27] 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.
- [28] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and

- quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.
- [29] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.
- [30] Muktari Mostofa, Mohammad Saeed Ebrahimi Saadabadi, Sahar Rahimi Malakshan, and Nasser M Nasrabadi. Pose attention-guided profile-to-frontal face recognition. *arXiv preprint arXiv:2209.07001*, 2022.
- [31] Mehdi Nourelahi, Lars Kotthoff, Peijie Chen, and Anh Nguyen. How explainable are adversarially-robust cnns? *arXiv preprint arXiv:2205.13042*, 2022.
- [32] 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.
- [33] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.
- [34] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019.
- [35] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6817–6826, 2020.
- [36] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–821, 2018.
- [37] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, 27, 2014.
- [38] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020.
- [39] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.
- [40] 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.
- [41] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12241–12248, 2020.
- [42] Yandong Wen, Weiyang Liu, Adrian Weller, Bhiksha Raj, and Rita Singh. Sphereface2: Binary classification is all you need for deep face recognition. *arXiv preprint arXiv:2108.01513*, 2021.
- [43] 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.
- [44] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.
- [45] Ali Zafari, Atefeh Khoshkhahtinat, Piyush M Mehta, Nasser M Nasrabadi, Barbara J Thompson, Daniel da Silva, and Michael SF Kirk. Attention-based generative neural image compression on solar dynamics observatory. *arXiv preprint arXiv:2210.06478*, 2022.
- [46] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019.
- [47] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5:7, 2018.
- [48] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.
- [49] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.