UniFace: Unified Cross-Entropy Loss for Deep Face Recognition

Jiancan Zhou[1,2,3,†], Xi Jia[1,2,4,†], Qiufu Li[1,2,†], Linlin Shen[1,2,#], Jinming Duan[4,5]
[1]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[2]Computer Vision Institute, Shenzhen University
[3]Aqara, Lumi United Technology Co., Ltd.
[4]School of Computer Science, University of Birmingham, UK
[5]Alan Turing Institute, UK
zhoujiancan@foxmail.com;x.jia.1@cs.bham.ac.uk;{liqiufu,llshen}@szu.edu.cn;j.duan@bham.ac.uk

Abstract

As a widely used loss function in deep face recognition, the softmax loss cannot guarantee that the minimum positive sample-to-class similarity is larger than the maximum negative sample-to-class similarity. As a result, no unified threshold is available to separate positive sample-to-class pairs from negative sample-to-class pairs. To bridge this gap, we design a UCE (Unified Cross-Entropy) loss for face recognition model training, which is built on the vital constraint that all the positive sample-to-class similarities shall be larger than the negative ones. Our UCE loss can be integrated with margins for a further performance boost. The face recognition model trained with the proposed UCE loss, UniFace, was intensively evaluated using a number of popular public datasets like MFR, IJB-C, LFW, CFP-FP, AgeDB, and MegaFace. Experimental results show that our approach outperforms SOTA methods like SphereFace, CosFace, ArcFace, Partial FC, etc. Especially, till the submission of this work (Mar. 8, 2023), the proposed UniFace achieves the highest TAR@MR-All on the academic track of the MFR-ongoing challenge. Code is publicly available.

1. Introduction

Face recognition, from verification on mobile phones to identification on surveillance streams, plays an important role in our daily life. A general face recognition system contains three core steps: face detection, facial feature extraction, and recognition (including one-to-one verification and one-to-all identification). Discriminative facial feature learning is therefore crucial to face recognition systems. Specifically, a facial feature of a subject should be close to the features belonging to the same identity, while being far from the features of the other identities, i.e., the minimum feature similarity from positive pairs should ideally be larger than a threshold t and the maximum feature similarity from negative pairs should be smaller than this t. Inspired by the success of deep neural networks on natural image classification tasks, modern facial recognition approaches are mostly based on deep convolutional neural networks. Such approaches can be broadly split into two categories according to their learning objectives, i.e., 1) sample-to-sample distance and 2) sample-to-class similarity. Sample-to-sample distance-based methods [19, 17] map the face images into a high-compact Euclidean feature space where distances are used to measure the facial feature similarity. However, the training of such methods is difficult [17], as they require sophisticated sampling strategies for constructing efficient negative and positive pairs/tuples. Sample-to-class similarity-based methods [22, 21] usually adopt the softmax loss as the learning objective and tackle face recog-

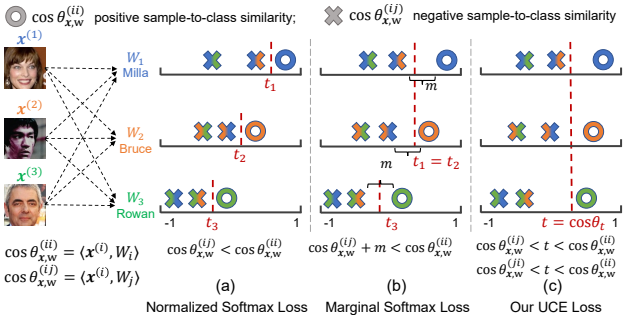

Figure 1. From (a), (b), and (c), we illustrate the sample-to-class similarities learned from normalized softmax loss, marginal softmax loss, and the proposed UCE loss, in which W_i is the class proxy, and $\boldsymbol{x}^{(i)}$ is a face image/feature with identity i. The classification of all three faces is correct with all three losses. However, both normalized and marginal softmax loss can not separate positive from negative sample-to-class pairs with a proper threshold, while with a unified threshold $t = \cos\theta_t$, our UCE loss can.

[†]Equal contribution; [#]Corresponding author. Parts of the work were done when J. Z and X. J were students at Shenzhen University.

nition as a multi-class classification problem. Some works proposed to combine the softmax loss with extra carefully designed losses to either increase the intra-class similarity [28] or decrease the inter-class similarity [11, 7, 30]. Such methods, however, introduce additional hyper-parameters which require elaborated hyper-parameter tuning. The other works (such as L-softmax loss [14], SphereFace[13], AM-softmax[23], CosFace [25], and ArcFace [6]) extended the original softmax loss by introducing marginal distance between classes to decrease the intra-class distance and increase the inter-class distance.

However, softmax loss only encourages the largest angular similarity between an individual training sample and its corresponding positive class proxy, it does not consider the similarities between this positive class proxy and other samples (see Sec. 3.1). In other words, it's difficult to select a unified threshold t to separate negative sample-to-class pairs from positive ones with both the softmax and marginal softmax loss, as illustrated in Fig. 1. To solve the problem, we propose a Unified Cross-Entropy (UCE) loss which explicitly encourages that all the positive sample-to-class similarities are larger than a threshold $t = \cos\theta_t$, while all negative sample-to-class similarities are smaller than this t. We elaborate on the proposed UCE loss (Sec. 3.2) and discuss its design principle linked with real applications (Sec. 3.3). We further improve the UCE loss by 1) introducing an enforced margin and 2) proposing two alternative ways to balance its training on a large number of face identities (Sec. 3.4). We name the face model trained with our UCE loss as UniFace and evaluate it on several public large-scale benchmarks (Sec. 4 and 5). The contributions of this work are summarized as follows:

- After investigating the softmax loss, we found that its learned minimum positive sample-to-class similarity is actually not guaranteed to be larger than its maximum negative sample-to-class similarity. To address this problem, we design the UCE loss by supposing a fixed threshold t to constrain the similarity of both positive and negative sample-to-class pairs.

- Though separating positive and negative pairs with a unified threshold t is a prestigious idea, to the best of our knowledge, this paper is the first work that incorporates the unified t as an automatic learnable parameter in a deep face recognition framework. Our UCE loss encourages that all the positive sample-to-class similarities are larger than the negative ones, which matches well with the expectation of real face recognition applications.

- Our UCE loss works well alone and can directly replace the softmax loss in existing deep face recognition models (Table 1 and 2). We additionally propose two extensions of UCE, i.e., marginal UCE and balanced UCE losses, to integrate with margins and balancing strategies to improve the performance of UCE loss. It is noticeable that the marginal UCE loss is more robust to hyper-parameters than the softmax loss (Fig. 3(a)).

- The face recognition model trained with the proposed UCE loss, UniFace, was intensively evaluated using a number of popular public datasets like MFR, IJB-C, LFW, CFP-FP, AgeDB, and MegaFace. Experimental results show that our approach outperforms SOTA methods such as SphereFace, CosFace, ArcFace, and Partial FC.

2. Related Works

2.1. Sample-to-Sample Distance based Methods

To reduce intra-subject variations while enlarging inter-subject differences, DeepID2 [19], apart from softmax loss, additionally uses a contrastive loss to encourage the features learned from the same identity to be close while that learned from different identity to be distant. FaceNet[17] proposes a triplet loss to map the images to high-compact Euclidean space where distances measure the face similarity. Using an enforced margin, Triplet loss minimizes the distance between an anchor and a positive sample and maximizes the distance between the anchor and a negative sample. However, the success of contrastive/triplet loss depends on a careful selection of pairs/triplets.

2.2. Sample-to-Class Distance based Methods

Different from the sample-to-sample-based contrastive loss, Wen et al. [28] propose the center loss to minimize the distances between a facial feature and its corresponding class center. SphereFace+[11], UniformFace[7], and RegularFace[30] propose additional regularization losses to maximize the distances between all the class proxies. Though these methods [28, 7, 30, 11] improve the feature learning of softmax loss, they all unavoidably introduce extra hyper-parameters to balance the softmax loss and the extra loss. The selection of these hyper-parameters requires intensive tuning, especially for methods (such as [28, 7]) that jointly use Euclidean-based loss and softmax loss at the same time. Liu et al. [14] propose a large margin softmax (L-softmax) loss by applying an angular margin to the angles between the training sample and its class proxy, such that the samples from the same identity are close to each other, while samples from different identities are apart. Subsequently, by first normalizing the weights and zeroing the biases, SphereFace [13] extends the L-softmax loss to angular softmax loss (A-softmax) to learn hyperspherical features. At the same time, NormFace [24] normalizes both the weights and the features, so that the loss only depends on the cosine similarities between the weights and the

features. CosFace [25] and ArcFace [6] adopt normalization and introduce an additive margin which is more stable than the multiplicative angular margin. Sphereface-R [12] extends SphereFace by re-implementing the multiplicative margin and proposes the Characteristic Gradient Detachment (CGD) strategy to further stabilize training.

Though these methods show improvement over the softmax loss, none of them explicitly constrains that all the positive sample-to-class similarities are larger than all negative sample-to-class similarities. GB-CosFace[3] also attempts to use a global threshold to align the training objective and the testing process. The threshold in [3] can only be calculated by a sophisticated hand-crafted statistical strategy involving extra hyper-parameters, the threshold t in our UCE loss, however, is automatically learned during training.

3. Methods

3.1. Revisiting Softmax Loss

Suppose $\mathcal{M}$ is a deep face model trained on a facial sample set $\mathcal{D}$ consisting of N subjects,

$$\mathcal{D} = \bigcup_{i=1}^{N} \mathcal{D}_i, \tag{1}$$

where $\mathcal{D}_i$ denotes the subset that contains the facial images of the same subject i. For any sample $\boldsymbol{X} \in \mathcal{D}$, let

$$\boldsymbol{x} = \mathcal{M}(\boldsymbol{X}) \in \mathbb{R}^{M\times 1} \tag{2}$$

denote the feature of $\boldsymbol{X}$, where M is the length of feature vector. Then, we can get a feature set $\mathcal{F}$,

$$\mathcal{F} = \bigcup_{i=1}^{N} \mathcal{F}_i = \bigcup_{i=1}^{N} \{\boldsymbol{x}^{(i)} = \mathcal{M}(\boldsymbol{X}^{(i)})\}_{\boldsymbol{X}^{(i)}\in\mathcal{D}_i}. \tag{3}$$

In the face models trained with the softmax loss, a full connection (FC) classifier, with weight matrix $\boldsymbol{W}$ and bias $\boldsymbol{b}$, is adopted to classify $\boldsymbol{X}$ based on its feature $\boldsymbol{x}$, where

$$\boldsymbol{W} = \left(W_1, W_2, \cdots, W_N\right) \in \mathbb{R}^{M\times N}, \tag{4}$$

$$\boldsymbol{b} = \left(b_1, b_2, \cdots, b_N\right)^T \in \mathbb{R}^{N\times 1}. \tag{5}$$

In Eq. (4), $W_i \in \mathbb{R}^{M\times 1}$ is the **class proxy** for subject i.

Following [13, 24], for convenience, we respectively normalize $\boldsymbol{W}$ and omit $\boldsymbol{b}$ as

$$\|W_i\| = 1, \quad b_i = 0, \quad 1 \le i \le N, \tag{6}$$

and, for all $\boldsymbol{x} \in \mathcal{F}$, we normalize the feature as $\|\boldsymbol{x}\| = s$.

Randomly take N samples $\{\boldsymbol{X}^{(i)}\}_{i=1}^{N} \subset \mathcal{D}$ with $\boldsymbol{X}^{(i)} \in \mathcal{D}_i$ for $\forall\, i$. Then, for a given sample $\boldsymbol{X}^{(i)}$, the typical **multi-class softmax loss** is

$$L_{\text{sl}}(\boldsymbol{X}^{(i)}) = -\log \frac{\mathrm{e}^{W_i^T\boldsymbol{x}^{(i)}+b_i}}{\mathrm{e}^{W_i^T\boldsymbol{x}^{(i)}+b_i} + \sum_{j\neq i}\mathrm{e}^{W_j^T\boldsymbol{x}^{(i)}+b_j}} \tag{7}$$

$$= -\log \frac{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}} + \sum_{j\neq i}\mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}}, \tag{8}$$

where

$$\theta^{(ij)}_{\boldsymbol{x},\text{w}} = \arccos\frac{\langle\boldsymbol{x}^{(i)}, W_j\rangle}{\|\boldsymbol{x}^{(i)}\|\|W_j\|} \tag{9}$$

$$= \arccos(\frac{1}{s}W_j^T\boldsymbol{x}^{(i)}) \in [0,\pi], \quad \forall\, i,j, \tag{10}$$

and $\langle\boldsymbol{x}^{(i)}, W_j\rangle$ denotes the inner product of the two vectors.

We term $\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}} = \frac{1}{s}W_i^T\boldsymbol{x}^{(i)}$ the **positive sample-to-class similarity** while $\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}} = \frac{1}{s}W_j^T\boldsymbol{x}^{(i)}$ the **negative sample-to-class similarity**, then we can get a sample-to-class similarity matrix $\mathcal{S}_{\text{sam-cla}}$,

$$\mathcal{S}_{\text{sam-cla}} = \begin{pmatrix} \cos\theta^{(11)}_{\boldsymbol{x},\text{w}} & \cos\theta^{(12)}_{\boldsymbol{x},\text{w}} & \cdots & \cos\theta^{(1N)}_{\boldsymbol{x},\text{w}} \\ \cos\theta^{(21)}_{\boldsymbol{x},\text{w}} & \cos\theta^{(22)}_{\boldsymbol{x},\text{w}} & \cdots & \cos\theta^{(2N)}_{\boldsymbol{x},\text{w}} \\ \vdots & \vdots & \ddots & \vdots \\ \cos\theta^{(N1)}_{\boldsymbol{x},\text{w}} & \cos\theta^{(N2)}_{\boldsymbol{x},\text{w}} & \cdots & \cos\theta^{(NN)}_{\boldsymbol{x},\text{w}} \end{pmatrix}. \tag{11}$$

To correctly classify $\boldsymbol{X}^{(i)}$, the softmax loss encourages larger positive sample-to-class similarity ($\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}$) than negative sample-to-class similarity ($\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}$), i.e.,

$$\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}} \le \cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}, \quad \forall\, i,j\neq i. \tag{12}$$

Then, $\exists\, t_i$, such that,

$$\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}} \le t_i \le \cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}, \quad \forall\, i. \tag{13}$$

However, the softmax loss does not consider the relationship between $\cos\theta^{(ji)}_{\boldsymbol{x},\text{w}}$ and $\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}$. In other words, there might exists a negative sample-to-class pair $(\boldsymbol{x}^{(j)}, W_i)$, whose similarity is even larger than that of the positive sample-to-class pair, $(\boldsymbol{x}^{(i)}, W_i)$, i.e., there might exists a sample $\boldsymbol{X}^{(j)} \in \mathcal{D}_j$ with $j \neq i$, whose feature $\boldsymbol{x}^{(j)}$ satisfies

$$t_i \le \cos\theta^{(ii)}_{\boldsymbol{x},\text{w}} < \cos\theta^{(ji)}_{\boldsymbol{x},\text{w}} \le t_j. \tag{14}$$

We expect that the similarities from positive pairs are larger than a threshold t and the similarities from negative ones are smaller than t. Though both the facial images $\boldsymbol{X}^{(i)}$ and $\boldsymbol{X}^{(j)}$ are correctly categorized into the right subjects in Eqs. (13) and (14). We notice that, contrary to our expectation, the positive sample-to-class pair, $(\boldsymbol{x}^{(i)}, W_i)$, has even

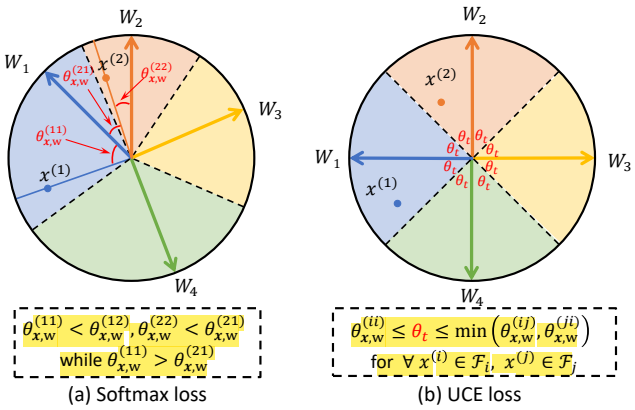

Figure 2. Geometric interpretations for four classes learned from (a) the normalized softmax loss, (b) our UCE loss. $\boldsymbol{x}^{(i)}$ is the feature belonging to the class W_i. The dark dashed lines denote the decision boundaries. Ideally, the sample features of the four different classes (W_i) are respectively contained in the four different sectors marked with four colors (i.e., blue, orange, yellow, and green). For a model well trained using softmax loss or UCE loss, its any sample feature $\boldsymbol{x}^{(i)}$ is closer to its class proxy W_i than to other class proxies $W_j, j \neq i$. However, for the softmax loss, the sample feature $\boldsymbol{x}^{(2)}$ is closer to W_1 than $\boldsymbol{x}^{(1)}$.

smaller similarity than the negative pair, $(\boldsymbol{x}^{(j)}, W_i)$. Since $t_i < t_j$, we can easily conclude that no **unified** similarity threshold $t = t_i = t_j$ is available to correctly separate the positive sample-to-class pairs ($(\boldsymbol{x}^{(i)}, W_i)$ and $(\boldsymbol{x}^{(j)}, W_j)$) from the negative pair ($(\boldsymbol{x}^{(j)}, W_i)$) at the same time.

We can further highlight this difficulty in the selection of a threshold with the sample-to-class similarity matrix $\mathcal{S}_{\text{sam-cla}}$ in Eq. (11), i.e., the softmax loss only encourages the diagonal element $\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}$ be dominant in the row i to achieve a good classification of sample $\boldsymbol{X}^{(i)}$, but neglects to encourage the domination of $\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}$ in the column i, which however is also important in face recognition.

3.2. Unified Cross-Entropy Loss

To avoid the problem in Eq. (14), and to encourage a similarity matrix $\mathcal{S}_{\text{sam-cla}}$ (see Eq. (11)) that is diagonally dominant in both its rows and columns. We expect a **unified threshold** t, such that

$$\begin{aligned} &\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}} \le t \le \cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}, \quad \text{and} \\ &\cos\theta^{(ji)}_{\boldsymbol{x},\text{w}} \le t \le \cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}, \quad \forall\, i,\ j,\ \text{with } j \neq i. \end{aligned} \tag{15}$$

If we define the maximum angle between the features and their positive class proxy as θ_{pos} and the minimum angle between the features and their negative class proxies as θ_{neg}, that is

$$\theta_{\text{pos}} = \max\left(\bigcup_{i=1}^{N}\left\{\theta^{(ii)}_{\boldsymbol{x},\text{w}} : \boldsymbol{x}^{(i)} \in \mathcal{F}_i\right\}\right), \tag{16}$$

$$\theta_{\text{neg}} = \min\left(\bigcup_{i=1}^{N}\bigcup_{\substack{j=1\\ j\neq i}}^{N}\left\{\theta^{(ij)}_{\boldsymbol{x},\text{w}} : \boldsymbol{x}^{(i)} \in \mathcal{F}_i\right\}\right), \tag{17}$$

then, there exists a threshold t satisfying Eq. (15) for any samples, if and only if $\theta_{\text{pos}} \le \theta_{\text{neg}}$, and the unified threshold $t = \cos\theta_t$ is valid for any

$$\theta_t \in [\theta_{\text{pos}}, \theta_{\text{neg}}]. \tag{18}$$

According to the analysis in Sec. 3.1, the softmax loss does not consider the constraint of the unified threshold t. For the first time, we incorporate the unified threshold t as an automatic learnable parameter in a loss function. The proposed Unified Cross-Entropy (UCE) loss is based on the assumption that a unified threshold $t = \cos\theta_t$ exists (i.e., $\theta_{\text{pos}} \le \theta_{\text{neg}}$). We elaborate on the derivations below, starting from the original softmax loss in Eq. (8)

$$\begin{aligned} &L_{\text{sl}}(\boldsymbol{X}^{(i)}) \\ &= -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}} + \sum_{j\neq i}\mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}} \end{aligned} \tag{19}$$

$$\begin{aligned} &= -\frac{1}{N}\Bigg(\log\frac{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}} + \sum_{j\neq i}\mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}} \\ &\quad + \sum_{\substack{k=1\\ k\neq i}}^{N}\log\frac{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}} + \mathrm{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\text{w}}} + \sum_{\substack{j\neq i\\ j\neq k}}\mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}}\Bigg). \end{aligned} \tag{20}$$

According to Eqs. (16) - (18), we can get

$$\begin{aligned} &L_{\text{sl}}(\boldsymbol{X}^{(i)}) \\ &\le -\frac{1}{N}\Bigg(\log\frac{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}} + \sum_{j\neq i}\mathrm{e}^{s\cos\theta_t}} \\ &\quad + \sum_{k\neq i}\log\frac{\mathrm{e}^{s\cos\theta_t}}{\mathrm{e}^{s\cos\theta_t} + \mathrm{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\text{w}}} + \sum_{\substack{j\neq i\\ j\neq k}}\mathrm{e}^{s\cos\theta_t}}\Bigg) \end{aligned} \tag{21}$$

$$\begin{aligned} &= \frac{1}{N}\Big[\log\left(1 + \mathrm{e}^{-s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}} + s\cos\theta_t + \log(N-1)}\right) \\ &\quad + \sum_{j\neq i}\log\left(1 + \mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}} - (s\cos\theta_t + \log(N-1))}\right) \\ &\quad + (N-1)\log(N-1)\Big]. \end{aligned} \tag{22}$$

Note that the detailed derivations of this inequality are described in Supplementary.

We define the **UCE** loss $L_{\text{uce}}(\boldsymbol{X}^{(i)})$ as

$$\begin{aligned} L_{\text{uce}}(\boldsymbol{X}^{(i)}) &= \log(1 + \mathrm{e}^{-s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}} + \tilde{b}}) \\ &\quad + \sum_{\substack{j\neq i\\ j=1}}^{N}\log(1 + \mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}} - \tilde{b}}), \end{aligned} \tag{23}$$

where $\tilde{b} = s\cos\theta_t + \log(N-1)$ is a constant to be learned.

UCE loss is based on the more vital constraint between positive and negative sample-to-class features than the softmax loss (via t in Eq. (15)). When a model is trained using UCE loss instead of softmax loss, it is expected that the final sample features are more discriminative. As depicted in Fig. 2, in the feature space partitioned by the proposed UCE loss, the similarity between $\boldsymbol{x}^{(1)}$ and W_1 is increased from the original softmax loss, while the similarity between $\boldsymbol{x}^{(2)}$ and W_1 is decreased.

Though the final formula of UCE loss (Eq. (23)) is similar to binary cross entropy (BCE) loss, there are several key differences between them. Firstly, UCE loss is designed from the objective of an explicit unified threshold t to constrain the similarity of both positive and negative sample-to-class pairs, while BCE loss and its variants [27] do not have such explicit constraints. Secondly, we derive the UCE loss from softmax loss, and we present the relationship between the unified threshold t and bias $\tilde{b} = s\cos\theta_t + \log(N-1)$ with a clear mathematical derivation, we then evaluate that the t is in line with the expectation of face verification with a qualitative illustration in Fig. 3 (c). Lastly, we systematically compare the UCE loss and BCE loss on a large benchmark dataset, where we compare (1) a standard BCE loss assigning respective biases for different classes (in Table 1), and (2) a simple modification of BCE loss excluding any biases, implying bias $b = 0$ (in Supplementary). The experimental results show continuous improvements by UCE loss over the two naive variants of BCE loss.

3.3. Rethinking UCE Loss for Face Verification

In real face verification systems, for any two facial image samples, $\boldsymbol{X}^{(i)} \in \mathcal{D}_i$, $\boldsymbol{X}^{(j)} \in \mathcal{D}_j$, a unified threshold t^* is chosen to verify whether they are taken from the same subject, by comparing their feature similarity $g(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ with the threshold t^*. This process implies a loss L_{v},

$$L_{\text{v}}(\boldsymbol{X}^{(i)}, \boldsymbol{X}^{(j)}) = \begin{cases} \alpha\max(0, t^* - g(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})) & i = j \\ \alpha\max(0, g(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) - t^*) & i \neq j \end{cases}, \tag{24}$$

where α is a re-weighting parameter.

Then, in the training, for a given $\boldsymbol{X}^{(i)}$, its loss is

$$\begin{aligned} L_{\text{v2}}(\boldsymbol{X}^{(i)}) &= \sum_{\boldsymbol{x}\in\mathcal{F}} L_{\text{v}}(\boldsymbol{x}, \boldsymbol{x}^{(i)}) \\ &= \sum_{\boldsymbol{x}\in\mathcal{F}_i} L_{\text{v}}(\boldsymbol{x}, \boldsymbol{x}^{(i)}) + \sum_{\substack{\boldsymbol{x}\in\mathcal{F}_j \\ j\neq i}} L_{\text{v}}(\boldsymbol{x}, \boldsymbol{x}^{(i)}). \end{aligned} \tag{25}$$

It will cost a large number of computations for every sample. A reasonable loss is designed using the class proxy W_i instead of the all features $\boldsymbol{x}^{(i)}$ in $\mathcal{F}_i$, for $\forall\, i$,

$$\begin{aligned} L_{\text{v3}}(\boldsymbol{X}^{(i)}) =& L_{\text{v}}(W_i, \boldsymbol{x}^{(i)}) + \sum_{\substack{j=1 \\ j\neq i}}^{N} L_{\text{v}}(W_j, \boldsymbol{x}^{(i)}) \qquad (26) \\ =& \alpha\max(0, t^* - g(W_i, \boldsymbol{x}^{(i)})) \\ &+ \sum_{j\neq i} \alpha\max(0, g(W_j, \boldsymbol{x}^{(i)}) - t^*). \qquad (27) \end{aligned}$$

In Eq. (27), the loss adopts the function $\text{ReLU}(x) = \max(0, x)$, which is not differentiable at $x = 0$. A proper substitution can be the softplus function,

$$\text{softplus}(x) = \frac{1}{\beta}\log(1 + \exp(\beta x)), \tag{28}$$

which tends to $\text{ReLU}(x)$ when β tends to $+\infty$.

Using the softplus function, then $L_{\text{v3}}(\boldsymbol{X}^{(i)})$ can be substituted as

$$\begin{aligned} L_{\text{v4}}(\boldsymbol{X}^{(i)}) =& \frac{\alpha}{\beta}\log(1 + \mathrm{e}^{\beta(t^* - g(W_i, \boldsymbol{x}^{(i)}))}) \\ &+ \sum_{j\neq i} \frac{\alpha}{\beta}\log(1 + \mathrm{e}^{\beta(g(W_j, \boldsymbol{x}^{(i)}) - t^*)}). \end{aligned} \tag{29}$$

When we set $\alpha = \beta = s$, and

$$g(W_j, \boldsymbol{x}^{(i)}) = \frac{1}{s} W_j^T \boldsymbol{x}^{(i)} = \frac{W_j^T \boldsymbol{x}^{(i)}}{\|W_j^T\|\|\boldsymbol{x}^{(i)}\|} = \cos\theta_{\boldsymbol{x},\text{w}}^{(ij)}, \tag{30}$$

the loss $L_{\text{v4}}(\boldsymbol{X}^{(i)})$ is our proposed UCE loss $L_{\text{uce}}(\boldsymbol{X}^{(i)})$ with $\tilde{b}^* = st^*$. To this end, we have linked the design of the proposed UCE loss with face verification.

3.4. Further Improvements

Marginal UCE Loss. Previous works [25] demonstrate that marginal softmax loss achieves better performance than the original one. We here extend the proposed UCE loss to marginal UCE loss by adding a cosine margin m,

$$\begin{aligned} &L_{\text{uce-m}}(\boldsymbol{X}^{(i)}) \\ &= \log(1 + \mathrm{e}^{-s(\cos\theta_{\boldsymbol{x},\text{w}}^{(ii)} - m) + \tilde{b}}) + \sum_{\substack{j\neq i \\ j=1}}^{N} \log(1 + \mathrm{e}^{s\cos\theta_{\boldsymbol{x},\text{w}}^{(ij)} - \tilde{b}}). \end{aligned} \tag{31}$$

Balanced UCE Loss. L_{uce} or $L_{\text{uce-m}}$ computes the similarity for only one positive sample-to-class pair $(\boldsymbol{x}^{(i)}, W_i)$, but $N-1$ negative pairs $(\boldsymbol{x}^{(i)}, W_j), \forall\, j,\ j \neq i$. This imbalance would result in unsatisfactory performance. Inspired by [2, 1] and [27], we introduce two parameters to balance

the number of positive and negative pairs,

$$L_{\text{uce-mb}}(\boldsymbol{X}^{(i)}) = \log(1+\mathrm{e}^{-s(\cos\theta^{(ii)}_{\boldsymbol{x},\mathrm{w}}-m)+\tilde{b}}) + \lambda \sum_{\substack{j=1 \\ j\neq i, p_j<r}}^{N} \log(1+\mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\mathrm{w}}-\tilde{b}}), \quad (32)$$

where p_j is a random number sampled from a uniform distribution (i.e., $U(0,1)$) for a negative sample-to-class pair $(\boldsymbol{x}^{(i)}, W_j)$. λ and r are respectively the re-weighting and sampling parameters for all negative sample-to-class pairs. The impacts of different λ and r are depicted in Fig. 3.

4. Experiments

4.1. Datasets and Evaluations

Datasets. For training, we use 4 publicly available datasets i.e., CASIA-WebFace [29] (0.5M images of 10K identities), Glint360K [2] (17.1M images of 360K identities), WebFace42M [31] (42.5M images of 2M identities), and WebFace4M, which is a subset of WebFace42M and has 4.2M images of 0.2M identities.

For face verification, we adopt the online testing of ICCV-2021 Masked Face Recognition Challenge (MFR Ongoing)[4], which contains not only previous popular testsets, like LFW [8], CFP-FP [18], AgeDB [16], and IJB-C [15], but also its own testsets such as the Mask set, Children set, and Multi-Racial set (MR-All, containing 4 different racial faces: African, Caucasian, S-Asian, and E-Asian).

For face identification, we employ the MegaFace Challenge 1[10] as the test set, which contains a gallery set with more than 1M images from 690K different identities, and a probe set with 3,530 images from 530 identities. Note that the MegaFace Challenge 1 has an additional verification track, which is also included in our experiments.

Evaluation. For MFR, we directly submit the trained models to the online MFR Ongoing Challenge server and report the performance. 1:1 verification accuracy is reported for LFW, CFP-FP, and AgeDB. True Accept Rate (TAR) @ False Accept Rate (FAR) = 1e-4 and TAR@FAR= 1e-5 are reported on the IJB-C. TARs@FAR= 1e-4 are reported for Mask and Children, and TARs@FAR=1e-6 are reported for the MR-All. For MegaFace Challenge 1, we used the official MegaFace devkit and dataset archived by InsightFace, reporting Rank1 accuracy for identification and TAR@FAR=1e-6 for verification.

4.2. Implementation Details

Preprocessing. We only use the standard face preprocessing. Each face image is first aligned using similarity transformation based on the five face landmarks detected by RetinaFace [5], and then cropped the center 112×112 patch. The intensity of all images is normalized to $[-1, 1]$ and is randomly horizontal-flipped for data augmentation.

Training. Following [6], we use the customized ResNets as our backbone. All models are implemented using Pytorch and trained with the SGD optimizer (5e-4 weight decay and 0.9 momentum). Following [25, 6], the feature norm s in our UCE is fixated at 64 in all experiments. We investigate the impact of different hyper-parameters in Sec. 4.4. *All training details are contained in the Supplementary.*

Testing. Given a face image, two 512-D features are extracted from the original and its horizontal-flipped image, such features are then added together as the final representation. The matching score is measured by cosine similarity.

4.3. Ablation Study

Effectiveness of the Proposed UCE Loss. In Table 1, the first two rows are the baseline results of a ResNet-50 model trained on the CASIA-WebFace dataset using the original normalized softmax loss L_{sl} and BCE loss L_{bce}, while the 3^{rd} row lists the results from our UCE loss L_{uce}, which is about 1% higher than both the original normalized softmax loss and BCE loss in terms of TAR@FAR=1e-6 on MR-All. The performance gains are more significant in terms of TAR@FAR=1e-4 on IJB-C: our UCE loss outperforms the original normalized softmax loss and BCE loss by 3.27% and 5.12%, respectively. When comparing to the marginal softmax loss $L_{\text{sl-m}}$ and marginal BCE loss $L_{\text{bce-m}}$, the performance of our marginal UCE loss $L_{\text{uce-m}}$ are respectively 5.65% and 42.48% higher than $L_{\text{sl-m}}$ in terms of TAR@FAR=1e-6 and TAR@FAR=1e-4 on MR-All and IJB-C, and are 2.1% and 4.77% higher than those of $L_{\text{bce-m}}$.

These improvements clearly suggest the effectiveness of our proposed UCE loss. In the last two rows of Table 1, though the balanced UCE loss introduces a slight performance fluctuation on the small LFW dataset, both re-weighting (λ) and sampling (r) improve the performance on much larger datasets, i.e., MR-All and IJB-C, which suggests the superiority of the balanced UCE loss.

Scalability. The UCE loss can directly replace the softmax-based loss in different frameworks. In Table 2, we integrate the UCE loss into three state-of-the-art face models i.e., SphereFace-R[12], ArcFace[6], and CosFace[25]. We then retrain these models and term them as Sphere-UniFace, Arc-UniFace, and Cos-UniFace. We can observe

Loss	UT	m	λ	r	MR-All	IJB-C	LFW
L_{sl}	✗	✗	✗	✗	18.52	71.53	98.30
L_{bce}	✗	✗	✗	✗	18.90	69.68	98.68
L_{uce}	✓	✗	✗	✗	19.59	74.80	98.45
$L_{\text{sl-m}}$	✗	✓	✗	✗	41.80	46.17	99.50
$L_{\text{bce-m}}$	✗	✓	✗	✗	45.35	83.88	99.46
$L_{\text{uce-m}}$	✓	✓	✗	✗	47.45	88.65	**99.56**
$L_{\text{uce-mb-}\lambda}$	✓	✓	✓	✗	48.54	**88.96**	99.55
$L_{\text{uce-mb-}r}$	✓	✓	✗	✓	**48.72**	88.94	99.30

Table 1. Ablation study on the proposed UCE loss. The "UT" marks whether an explicit threshold t is considered in losses.

Method	MR-All	IJB-C	LFW
SphereFace-R v1[12]	39.92	86.35	99.38
Sphere-UniFace	**41.00**	**88.42**	**99.43**
ArcFace[6]	45.59	60.31	99.40
Arc-UniFace	**47.97**	**88.70**	**99.43**
CosFace[25]	41.80	46.17	99.50
Cos-UniFace	**47.45**	**88.65**	**99.56**

Table 2. Performance of UCE loss with different frameworks.

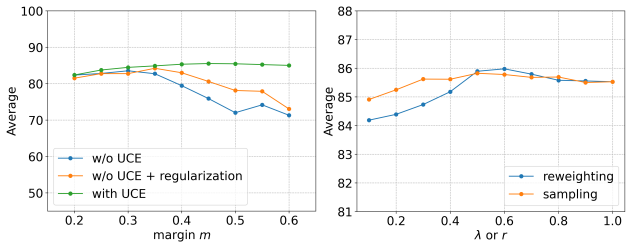

Figure 3. Left: impacts of different m of the compared losses, our marginal UCE loss stably improves the performance with larger m. Right: impacts of different λ and r of our balanced UCE loss on the average results of MFR ongoing testset.

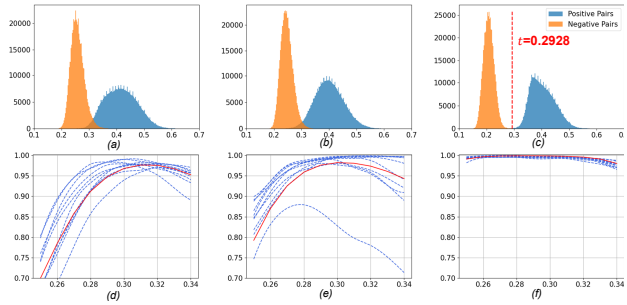

Figure 4. (a, d) normalized softmax loss, (b, e) BCE loss, and (c, f) our UCE loss. **First row**: Positive and negative sample-to-class similarity ($\cos\theta^{(ii)}_{\boldsymbol{x},\mathrm{w}}$ and $\cos\theta^{(ij)}_{\boldsymbol{x},\mathrm{w}}$) distributions of 490,623 samples in CASIA-WebFace. The positive and negative sample-to-class similarities are well separated for our UCE loss, while large overlapping exists for both normalized softmax loss and BCE loss. The red line indicates the final unified threshold $t \approx 0.2928$. **Second row**: The changes of verification accuracy with the threshold on 10 randomly selected identities (blue lines), while the red curves are averaged on all identities.

a consistent improvement brought by our UCE loss.

4.4. Parameter Study

We study the impact of different hyper-parameters of the two proposed marginal and balanced UCE losses below.

Qualitative Results. To better understand the proposed UCE loss, in Fig. 4, we plot the positive and negative sample-to-class similarity distributions of the normalized softmax loss, BCE loss, and UCE loss. From this figure, we can observe that an intersection between positive and negative similarities exists in the softmax and BCE losses, while they are well separated in our UCE loss. We further compute the unified threshold $t \approx 0.2928$ from the $\tilde{b}$ learned by the UCE loss in Eq. (23), which can well separate the positive and negative sample-to-class pairs and again proves that our UCE loss can solve the difficulty of softmax loss in the selection of a unified threshold. On the other hand, we randomly select 10 identities to plot their changes in sample-to-class verification accuracy regarding the threshold, we can clearly observe that all 10 identities achieve stable and similar accuracy around the unified threshold $t \approx 0.2928$ with our UCE loss, while there is large variance regarding the accuracy with the other two losses. These findings suggest the advantage of learning a unified threshold t.

Robustness Against Different Margins. We first investigate the impact of different margins (m in Eq. (31)) of the marginal UCE loss. In Table A of Supplementary, when increasing the m from 0.2 to 0.6 with an interval of 0.05, the average performance of the original marginal softmax loss and that of the Exclusive Regularization loss [30] first improves and then rapidly drops. For our marginal UCE loss, however, the performance is stably increased. To help clarify the differences between the three methods, we plot the changes in the average performance with increasing m in the left sub-figure of Fig. 3, which suggests that our UCE loss is more robust and less sensitive to larger margins, while both the original marginal loss and the Exclusive Regularization loss [30] are not.

Effects of Different Balance Strategies. We then study different hyper-parameters for the proposed balanced UCE loss. As per Eq. (32), we have two alternative ways to balance the proposed UCE loss, i.e., re-weight all the negative sample-to-class losses with λ or randomly sample the negative sample-to-class losses with a ratio of $r \times 100\%$. We examine different λ and r from 0.1 to 1.0 in Table B of Supplementary. It shows that proper adjustment of these parameters can improve the final performance, while a too-small value can lead to performance drops. To display the difference more clearly, we also plot the average results in 3 (the right sub-figure). It suggests that, with proper parameters, the balanced UCE loss can further improve the performance of the marginal UCE loss. Experimentally, the sampling strategy is better than the re-weighting strategy when r and λ are small, otherwise, the re-weighting strategy is better than the sampling strategy.

5. Comparison with the State-Of-The-Art

5.1. MFR Ongoing Benchmarks

We first compare the proposed UniFace with a number of state-of-the-art methods, such as SphereFace-R[12], GB-CosFace[3], SphereFace2[27], ArcFace[6], CosFace[25], and Regularization[30]. For a fair comparison, we re-implement these methods and employ the optimal hyper-parameters recommend in their original papers. All the compared models are trained with a ResNet-50 backbone

Method	Network + Dataset	MFR							IJB-C		Verification Accuracy		
		Mask	Children	African	Caucasian	S-Asian	E-Asian	**MR-All**	1e-4	1e-5	LFW	CFP-FP	AgeDB
CosFace[25]	R50 + CASIA	37.32	29.12	46.53	61.06	56.91	22.23	43.34	79.78	38.82	99.36	96.60	94.53
CosFace[25] + Regularization[30]		**39.70**	30.27	43.69	59.73	56.30	24.67	44.57	85.84	61.59	99.36	96.70	94.61
GB-CosFace[3]		29.25	22.97	34.63	52.60	44.92	16.06	32.92	85.97	76.40	99.35	97.18	93.81
ArcFace[6]		39.38	32.48	49.52	64.38	59.37	19.55	45.59	60.31	17.20	99.40	97.27	94.96
SphereFace-R v1[12]		32.80	28.09	40.24	57.24	50.38	22.30	39.92	86.35	75.81	99.38	96.95	94.48
SphereFace2[27]		35.40	30.55	46.65	62.69	56.23	26.65	44.20	88.41	**79.18**	99.46	97.42	94.96
UniFace, $L_{\text{uce-m}}$		38.75	31.67	49.25	66.39	60.44	28.58	47.45	88.65	78.42	**99.56**	97.24	94.71
UniFace, $L_{\text{uce-mb-}\lambda}$		37.86	32.74	49.81	64.82	60.55	29.70	48.54	**88.96**	78.40	99.55	**97.47**	**95.36**
UniFace, $L_{\text{uce-mb-}r}$		39.25	**33.11**	**50.79**	**66.46**	**61.47**	**29.71**	**48.72**	88.94	78.30	99.30	97.20	94.95
Partial FC [1]	R50+WF4M	72.28	-	84.86	91.57	88.57	67.52	86.85	-	-	-	-	-
UniFace, $L_{\text{uce-m}}$		75.48	69.00	86.62	92.99	90.26	68.81	88.30	96.92	**95.16**	99.80	99.10	**97.96**
UniFace, $L_{\text{uce-mb-}r}$		**75.56**	**69.49**	**87.30**	**93.38**	**90.59**	69.09	88.49	**96.97**	94.85	99.76	**99.14**	97.71
UniFace, $L_{\text{uce-mb-}\lambda}$		75.46	69.32	86.89	93.02	90.36	**69.46**	**88.55**	96.96	94.90	**99.80**	98.98	97.88
Partial FC [1]	R200 + WF42M	91.87	-	97.79	98.70	98.54	89.52	97.70	97.97	96.93	99.83	**99.51**	**98.70**
Partial FC [1]	ViT-L + WF42M	90.88	92.37	98.07	98.81	98.66	89.97	97.85	**98.00**	**97.23**	99.83	99.44	98.67
UniFace, $L_{\text{uce-mb-}r}$	R200 + WF42M	**92.43**	**93.11**	**98.14**	**98.98**	**98.84**	**90.01**	**97.92**	97.91	96.68	**99.83**	99.42	98.66
UniFace, $L_{\text{uce-mb-}\lambda}$		92.18	93.05	98.02	98.89	98.69	89.52	97.78	97.98	96.88	99.81	99.40	98.66

Table 3. Comparisons between different methods on MFR-Ongoing (Results of Partial FC are from the original paper[1]).

and the CASIA-WebFace dataset. As reported in Table 3, our UniFace achieves clear improvement over other methods. In order to further explore the capacity of the proposed UniFace, we also compare our method with Partial FC[1], which is the leading method on the MFR ongoing challenge. To guarantee a fair comparison, following Partial FC[1], we train the proposed UniFace on the WebFace4M and WebFace42M datasets using two different architectures, i.e., ResNet-50 and ResNet-200. We can observe that, on average, our performance is better than that of the Partial FC.

5.2. MegaFace Challenge 1

We compare the identification performance of UniFace with a number of state-of-the-art methods on MegaFace Challenge 1, such as SphereFace[13], CosFace[25], ArcFace[6], UniformFace[7], MV-AM-Softmax[26], Circle Loss[20], CurricularFace[9], and Partial FC[2]. For a fair comparison, as per the official protocols, UniFace trained on the CASIA-WebFace is compared with the models trained on 'Small' datasets while UniFace trained on Glint360K is compared with the models trained on 'Large' datasets in Table 4. Among the compared models trained on 'Small' datasets, UniFace achieves the highest accuracy, i.e., 77.83% identification and 93.64% verification accuracy. When the label refinement [6] is used, the accuracies can further be increased to 92.75% and 95.17% respectively. Among the compared models trained on 'Large' datasets, UniFace also achieves the highest accuracy on both identification and verification, whether the label refinement is used, or not. Both the verification and identification results prove the effectiveness of the proposed UniFace.

6. Conclusion

In this paper, by explicitly introducing a learnable threshold to constrain the similarity of both positive and negative sample-to-class pairs, we propose the UCE loss, which is the first work that incorporates the unified t as an automatic learnable parameter in a deep face recognition framework.

Method	Protocol	Refine	Iden.	Veri.
Softmax Loss [13]	Small	No	54.85	65.92
Triplet Loss [13, 17]	Small	No	64.79	78.32
Softmax + Contrastive Loss [13, 19]	Small	No	65.21	78.86
Softmax + Center Loss [13, 28]	Small	No	65.49	80.14
L-Softmax Loss [13, 14]	Small	No	67.12	80.42
SphereFace [13]	Small	No	72.72	85.56
SphereFace+ [11]	Small	No	73.03	-
CosFace [25]	Small	No	77.11	89.88
ArcFace, R50 [6]	Small	No	77.50	92.34
CurricularFace, R50 [9]	Small	No	77.65	92.91
UniFace, $L_{\text{uce-mb-}\lambda}$, R50 + CASIA	Small	No	**77.83**	**93.64**
ArcFace, R50 [6]	Small	Yes	91.75	93.69
CurricularFace, R50 [9]	Small	Yes	92.48	94.55
UniFace, $L_{\text{uce-mb-}\lambda}$, R50 + CASIA	Small	Yes	**92.75**	**95.17**
FaceNet [17]	Large	No	70.49	86.47
RegularFace [30]	Large	No	75.61	91.13
UniformFace [7]	Large	No	79.98	95.36
ArcFace, R100 [6]	Large	No	81.03	96.98
CurricularFace, R100 [9]	Large	No	81.26	97.26
CosFace [25]	Large	No	82.72	96.65
UniFace, $L_{\text{uce-mb-}\lambda}$, R100 + Glint360K	Large	No	**84.87**	**97.85**
SphereFace2 [27]	Large	Yes	89.84	91.94
CosFace, R100 [6, 25]	Large	Yes	97.91	97.91
MV-AM-Softmax [26]	Large	Yes	98.00	98.31
SphereFace-R v1 [12]	Large	Yes	98.03	98.30
SphereFace-R v2 [12]	Large	Yes	98.04	98.48
SphereFace [13]	Large	Yes	98.16	98.46
ArcFace, R100 [6]	Large	Yes	98.35	98.48
Circle Loss, R100 [20]	Large	Yes	98.50	98.73
CurricularFace, R100 [9]	Large	Yes	98.71	98.64
Partial FC, r=0.1, R100 [2]	Large	Yes	98.94	99.10
Partial FC, r=1.0, R100 [2]	Large	Yes	99.13	98.98
UniFace, $L_{\text{uce-mb-}\lambda}$, R100 + Glint360K	Large	Yes	**99.27**	**99.19**

Table 4. Comparisons on the MegaFace Challenge 1.

The proposed UCE loss matches well with the expectation of real face recognition applications and achieves clear improvement over the state-of-the-art methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 82261138629 and 62006156, Guangdong Basic and Applied Basic Research Foundation under Grants 2023A1515010688 and 2022A1515012125, and Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20220531101412030.

References

[1] 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.

[2] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.

[3] Mingqiang Chen, Lizhe Liu, Xiaohao Chen, and Siyu Zhu. Gb-cosface: Rethinking softmax-based face recognition from the perspective of open set classification. In *Proceedings of the Asian Conference on Computer Vision*, pages 670–686, 2022.

[4] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021.

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, 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.

[6] 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.

[7] 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 (CVPR)*, June 2019.

[8] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[9] 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.

[10] 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.

[11] Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. *Advances in neural information processing systems*, 31, 2018.

[12] Weiyang Liu, Yandong Wen, Bhiksha Raj, Rita Singh, and Adrian Weller. Sphereface revived: Unifying hyperspherical face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] Yandong Wen, Weiyang Liu, Adrian Weller, Bhiksha Raj, and Rita Singh. Sphereface2: Binary classification is all you need for deep face recognition. In *International Conference on Learning Representations*, 2022.

[28] 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.

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

[30] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. Regularface: Deep face recognition via exclusive regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] 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.