GB-CosFace: Rethinking Softmax-based Face Recognition from the Perspective of Open Set Classification

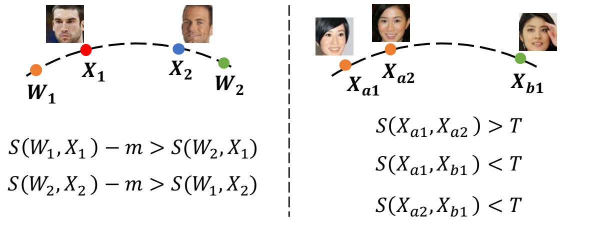
Lizhe Liu[1†] Mingqiang Chen[1†] Xiaohao Chen[1] Siyu Zhu[1] Ping Tan[12]
[1]Alibaba Group [2]Simon Fraser University [†]Equal contribution

Abstract

State-of-the-art face recognition methods typically take the multi-classification pipeline and adopt the softmax-based loss for optimization. Although these methods have achieved great success, the softmax-based loss has its limitation from the perspective of open set classification: the multi-classification objective in the training phase does not strictly match the objective of open set classification testing. In this paper, we derive a new loss named global boundary CosFace (GB-CosFace). Our GB-CosFace introduces an adaptive global boundary to determine whether two face samples belong to the same identity so that the optimization objective is aligned with the testing process from the perspective of open set classification. Meanwhile, since the loss formulation is derived from the softmax-based loss, our GB-CosFace retains the excellent properties of the softmax-based loss, and CosFace is proved to be a special case of the proposed loss. We analyze and explain the proposed GB-CosFace geometrically. Comprehensive experiments on multiple face recognition benchmarks indicate that the proposed GB-CosFace outperforms current state-of-the-art face recognition losses in mainstream face recognition tasks. Compared to CosFace, our GB-CosFace improves 1.58%, 0.57%, and 0.28% at TAR@FAR=1e-6, 1e-5, 1e-4 on IJB-C benchmark.

1. Introduction

Research on the training objectives of face recognition (FR) has effectively improved the performance of deep-learning-based face recognition [32, 34, 39, 40]. According to whether a proxy is used to represent a person's identity or a set of training samples, face recognition methods can be divided into proxy-free methods [4, 8, 12, 22–24, 27, 29, 30, 35, 42, 48] and proxy-based methods [3, 5, 15, 17, 20, 31, 33, 36–38, 47]. The proxy-free methods directly compress the intra-class distance and expand the inter-class distance based on pair-wise learning [4, 12, 30, 35] or triplet learning [8, 22, 23, 27, 29, 42, 48]. However, when dealing with a large amount of training data, the hard-mining operation which is crucial for proxy-free methods becomes extremely difficult. Recently, proxy-based method have achieved great success and shown advantages in big data training. Most of them take a softmax-based multi-classification pipeline and use cross-entropy loss as the optimization objective. In these methods, each identity in the training set is represented by a prototype, which is the weight vector of the final fully connected layer. We refer to this type of method as the softmax-based face recognition method in this paper.



(a) Softmax Training Objective **(b) Open-set Testing Objective**

Figure 1. The difference of the objective between softmax-based training and the open set classification testing, where $S(\cdot)$ is the function to measure the distance between two samples, W_1 and W_2 are the prototypes of two identities respectively. In **Figure (a)**, X_1 and X_2 is the given training sample, m is the margin parameter. In **Figure (b)**, X_{a1} and X_{a2} are two testing samples of ID "a", and X_{b1} is a testing sample of ID "b". ID "a" and "b" are not included in the training data.

The traditional softmax-cross-entropy loss designed for the close set multi-classification problem is not suitable for an open set classification problem such as face recognition. Current softmax-based face recognition methods have made various improvements to the training objective. One of the most vital improvements is to normalize the face features to the hypersphere for unified comparison [17, 37]. Typically, the similarity between two samples is represented by the cosine similarity of their corresponding feature vectors. Given the feature vector of a training sample and the prototypes of the training identities, the training objective is to compress the distance between the feature vector and the corresponding prototype and to push away the other prototypes from the feature vector. To further compress the intra-class distance and expand the inter-class distance, large-margin-based methods [5,17,36,38] are proposed. Recently, the dynamic schemes for the scale parameter [44] and the margin parameter [16, 20] have been studied and further improved the model performance.

Despite the great success of softmax-based face recognition, this strategy has its limitation from the perspective of the open set classification [9, 10, 26, 43]. As is shown in **Figure** 1(a), the training objective of softmax-based multi-classification is to make the predicted probability of the target category larger than other categories. However, face recognition is an open set classification problem where the test category generally does not exist in the training category [39]. A typical requirement for a face recognition model is to determine whether two samples belong to the same identity by comparing the similarity between them with a global threshold T, as is shown in **Figure** 1(b). The inconsistency of the objective of training and testing limits the performance. Based on this consideration, we believe that the consistency of the objective during training and testing needs to be considered in the loss design.

In this paper, we propose a novel face recognition loss named global boundary CosFace (GB-CosFace). In our GB-CosFace framework, the training objective is aligned with the testing process by introducing a global boundary determined by the proposed adaptive boundary strategy. First, we compare the objective difference between the softmax-based loss and the face recognition testing process. Then, we abstract the reasonable training objective from the perspective of open set classification and derive a antetype of the proposed loss. Furthermore, we combine the excellent properties of softmax-based losses with the proposed antetype loss and derive the final GB-CosFace formulation, and prove that CosFace [36, 38] is a special case of the proposed GB-CosFace. Finally, we analyze and explain the proposed GB-CosFace geometrically. The contributions of this paper are summarized as follows.

- We propose GB-CosFace loss for face recognition, which matches the objective of the testing process of the open set classification problem while inheriting the advantages of the softmax-based loss.

- We analyze the difference and connection between GB-CosFace and general softmax-based losses, and give a reasonable geometric explanation.

- Our GB-CosFace obviously improve the performance of softmax-based face recognition (e.g., improves 1.58%, 0.57%, and 0.28% at TAR@FAR=1e-6, 1e-5, 1e-4 on IJB-C benchmark compared to CosFace).

2. Softmax-based Face Recognition

To better understand the proposed GB-CosFace, this section review the general softmax-based face recognition.

2.1. Framework

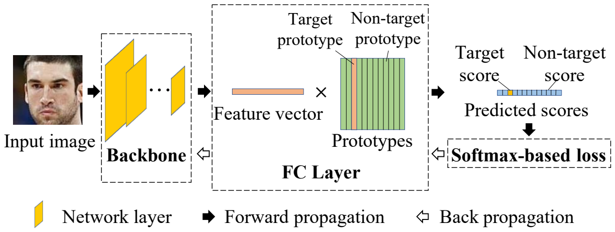


Figure 2. The training framework of the general softmax-based face recognition.

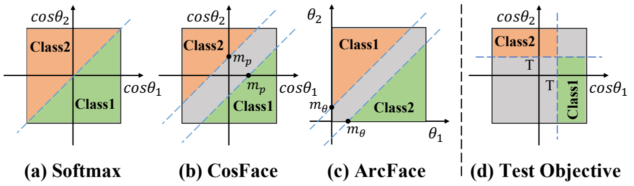
The training framework of the general softmax-based face recognition is shown in **Figure** 2. In this framework, each identity in the training set has its corresponding prototype. The prototypes are represented by the weight vectors of the final fully connected layer. Given a training sample, we call the prototype representing the identity of this sample "target prototype", and call other prototypes "non-target prototypes". After the facial feature extraction of the backbone, the predicted scores which represent the similarity between the feature vector and each prototype are calculated through the final fully connected layer (FC layer). The similarity between the feature vector and the target prototype is called "target score", and the other predicted scores are called "non-target scores". Generally, the output feature vector and the prototypes are normalized to the unit hypersphere. Therefore, the predicted scores are usually represented by the cosine of the feature vector and the prototype. In training, the softmax-based loss is adopted to optimize the backbone and the final FC layer through backpropagation.

2.2. Objective

For each iteration in n-class face recognition training, given a training sample and its label y, the general softmax-based loss is as follows:

$$\mathcal{L}_S = -log\frac{e^{s(cos(\theta_y+m_\theta)-m_p)}}{e^{s(cos(\theta_y+m_\theta)-m_p)}+\sum_i e^{scos_{\theta i}}} \qquad (1)$$

where θ_y is the arc between the predicted feature vector and the target prototype, θ_i is the arc between the predicted feature vector and the non-target prototype, y is the index of the target identity, i is the index of the non-target identities, $i \in [1, n]$ and $i \neq y$. There are three hyper-parameters: the scale parameter "s", and the two margin parameters "m_θ" and "m_p".



(a) Softmax (b) CosFace (c) ArcFace (d) Test Objective

Figure 3. Decision boundaries of different loss functions under binary classification case. Figure (d) shows the expected decision boundary in the testing phase.

We can reach several common softmax-based losses from **Equ.** 1. E.g., normalized softmax loss will be reached if both m_θ and m_p are set as zero. ArcFace and CosFace will be reached if we respectively set m_p and m_θ as 0.

Softmax-based losses can be regarded as the smooth form of the following optimization objective $\mathcal{O}_S$.

$$\begin{aligned}\mathcal{O}_S &= ReLU(max(cos\theta_i) - (cos(\theta_y + m_\theta) - m_p)) \\ &= \lim_{s\to+\infty} -\frac{1}{s} log \frac{e^{s(cos(\theta_y+m_\theta)-m_p)}}{e^{s(cos(\theta_y+m_\theta)-m_p)} + \sum_{i=1,i\neq y}^{n} e^{scos\theta_i}} \\ &= \lim_{s\to+\infty} \frac{1}{s}\mathcal{L}_S\end{aligned} \tag{2}$$

Where the SoftPlus function is used as a smooth form of ReLU operator and $log\sum exp(\cdot)$ is used as a smooth form of $max(\cdot)$ operator. More detailed derivation is included in the supplementary material.

From this perspective, we can find that the training objective $\mathcal{O}_S$ constrains the target score to be larger than the maximum non-target score. The margin is introduced for a stricter constraints. However, this constraint is not completely consistent with the objective of the testing process. Based on **Equ.** 2, we can visualize the decision boundaries of normalized softmax loss [37], CosFace [36,38], and ArcFace [5] under binary classification case, as is shown in **Figure** 3 (a)-(c). In the testing phase, a global threshold T of the cosine similarity needs to be fixed to determine whether two samples belong to the same person, as is shown in **Figure** 3(d). We can see that, even if a margin is added, the decision boundaries of softmax-based losses do not completely match the expected boundary for testing.

2.3. Properties

Current face recognition models do not directly apply $\mathcal{O}_S$ as the training objective. On the one hand, $max(\cdot)$ operator only focuses on the maximum value and the gradients will only be backpropagated to the target score and maximum non-target score. On the other hand, if the argument of the RELU function is less than 0, no gradient will be backpropagated. As a smooth form of $\mathcal{O}_S$, the softmax-based loss can avoid the above problems. The success of softmax-based loss is due to its excellent properties.

Property 1. The gradients of the non-target scores are proportional to their softmax value.

For softmax-based loss, the backpropagated gradients will be assigned to all non-target scores according to their softmax value. This property ensures that each non-target prototype can play a role in training, and hard non-target prototypes get more attention.

Property 2. The gradient of the target score and the sum of the gradients of all non-target scores have the same absolute value and opposite signs.

$$\frac{\partial \mathcal{L}_S}{\partial cos(\theta_y + m_\theta)} = -\sum\nolimits_i \frac{\partial \mathcal{L}_S}{\partial cos(\theta_i)} \tag{3}$$

Softmax-based loss has balanced gradients for the target score and the non-target scores. This property can maintain the stability of training and prevent the training process from falling into a local minimum.

Considering the key role that these two properties play in face recognition training, we expect to inherit them in the loss design. In this paper, we add the consistency of training and testing to the loss design by introducing an adaptive global boundary. From the expected training objective, we derive our GB-CosFace framework and prove compatibility with CosFace. This compatibility allows the proposed loss to inherit the excellent properties of the general softmax-based loss while solving the inconsistency between the training and testing objective.

3. GB-CosFace Framework

3.1. Antetype Formulation

Based on the face recognition testing process which is shown in **Fig** 3(d), we propose to introduce a global threshold p_v as the boundary between target score and non-target scores. The target score is required to be larger than p_v while the maximum of the non-target scores is required to be less than p_v. Following this idea, we improve **Equ.** 2 as follows:

$$\begin{cases} \mathcal{O}_T = ReLU(p_v - (p_y - m)) \\ \mathcal{O}_N = ReLU(max(p_i) - (p_v - m)) \end{cases} \tag{4}$$

where we divide the training objective into $\mathcal{O}_T$ and $\mathcal{O}_N$ for the target score and the non-target scores respectively, p_y is the target score, $p_y = cos\theta_y$, p_i is the non-target score, $p_i = cos\theta_i$, and m is the margin parameter introduced for stricter constraints. The training objective is to minimize $\mathcal{O}_T$ and $\mathcal{O}_N$.

Inspired by the success of the softmax-based loss, similar to **Equ.** 2, we take the smooth form of $\mathcal{O}_T$ and $\mathcal{O}_N$ as the antetype of the proposed loss.

$$\begin{cases} \mathcal{L}_{T1} = -log\frac{e^{s(p_y-m)}}{e^{s(p_y-m)}+e^{sp_v}} \\ \mathcal{L}_{N1} = -log\frac{e^{s(p_v-m)}}{e^{s(p_v-m)}+\sum_i e^{sp_i}} \end{cases} \tag{5}$$

Where the loss for target score and non-target scores are represented as $\mathcal{L}_{T1}$ and $\mathcal{L}_{N1}$ respectively, p_v is the global boundary hyper-parameter, which also means "virtual score". For $\mathcal{L}_{T1}$, p_v is a virtual non-target score. For $\mathcal{L}_{N1}$, p_v is a virtual target score. Since we take $log\sum exp(\cdot)$ as the smooth form of $max(\cdot)$, the distribution of the gradients of non-target scores inherits **Property 1.** (stated in **Section** 2.3) of the softmax-based loss.

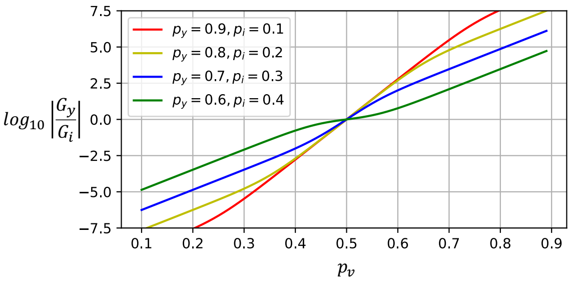


Figure 4. The ratio of the target gradient to the non-target gradient varies with p_v under binary classification case using different p_y and p_i. Hyper-parameter s and m are set to 32 and 0.15 respectively. Note that the ordinate is the base 10 logarithm of the ratio.

However, the proposed antetype introduces another problem: the setting of hyper-parameter p_v. First, the inappropriate setting of p_v can cause a serious gradient imbalance problem. Since we separate the constraints on the target score and the non-target scores, the gradient balance for target score and non-target scores is broken and the antetype loss no longer retains **Property 2.** (stated in **Section** 2.3). Second, considering the rapid rise of the exponential function and the amplification effect of the hyper-parameter "s", the model is extremely sensitive to the choice of the hyper-parameter p_v. As is shown in **Figure** 4, a slight change in p_v can cause an order of magnitude difference between the gradients for target score p_y and non-target scores p_i. Third, considering that the predicted scores of each sample are dynamic during the training process, the ideal threshold p_v should also be a dynamic parameter to adapt to different training stages. Therefore, an adaptive scheme for the global boundary is necessary.

3.2. Adaptive Global Boundary

To control the gradient balance and adapt the global boundary to different training stages, we propose an adaptive global boundary method. We believe that an ideal global boundary should meet the following conditions: **a)** Under this boundary setting, the gradients of the target score and the non-target scores should be roughly balanced from a global perspective; **b)** The global boundary should change slowly during the training process to keep the training stable while adapting to different training stages. Based on these two conditions, we make the following design.

3.2.1 Gradient Balance Control

We define $\hat{p}_v$ as the balanced threshold of the target score and the non-target scores which satisfies $\frac{\partial \mathcal{L}_{T1}}{\partial p_y} = -\sum_i \frac{\partial \mathcal{L}_{N1}}{\partial p_i}$. Based on this condition, we reach the following form of $\hat{p}_v$:

$$\hat{p}_v = (p_y + \frac{1}{s}log\textstyle\sum_i e^{sp_i})/2 \tag{6}$$

Ideally, for each iteration, to satisfy the above condition **a)**, we expect to calculate $\hat{p}_v$ for each sample in the data set and get the mean value as the threshold p_v. Considering the efficiency, we calculate the mean of $\hat{p}_v$ for each batch and update it by the momentum update strategy.

$$p_{vg} = (1-\gamma)p_{vg} + \gamma p_{vb} \tag{7}$$

Where $\gamma \in [0, 1]$ is the update rate, p_{vb} is the mean of p_v in a batch. A small γ can keep the stability of p_v. We empirically set γ to 0.01.

This dynamic threshold strategy makes the gradient balanced globally. However, for each sample, the problem of gradient imbalance can be very serious. Therefore, we modify the value of p_v to be the weighted sum of p_{vg} and $\hat{p}_v$ as follows.

$$p_v = \alpha p_{vg} + (1-\alpha)\hat{p}_v \tag{8}$$

Where α is a hyper-parameter and $\alpha \in [0, 1]$. When $\alpha = 0$, the gradients for the target score and the non-target scores are completely balanced. We can control the degree of the gradient imbalance by adjusting α.

3.2.2 Compatible with CosFace

In **Equ.** 8, if we take α as 0, the proposed loss will fully conform to **Property 1** and **Property 2** (stated in **Section** 2.3) of softmax-based loss. Through the following analysis, we can further find that Cosface [36, 38] is a special case of the proposed loss when $\alpha = 0$.

The gradients based on CosFace is calculated as follows.

$$\begin{aligned} \mathcal{G}_{T-CosFace} &= -\mathcal{G}_{N-CosFace} \\ &= -\frac{s \cdot \sum_i e^{sp_i}}{e^{s(p_y-m)} + \sum_i e^{sp_i}} \\ &= -\frac{s \cdot e^{sp_n}}{e^{s(p_y-m)} + e^{sp_n}} \end{aligned} \tag{9}$$

Where the gradient for the target score is represented as $\mathcal{G}_{T-CosFace}$, the sum of the gradients of the non-target scores is represented as $\mathcal{G}_{N-CosFace}$, and $p_n = \frac{1}{s} log \sum_i e^{sp_i}$.

For the proposed loss, based on **Equ.** 5, we can get the gradient for target score p_y ($\mathcal{G}_{T1}$) and the sum of the gradients for non-target scores p_i ($\mathcal{G}_{N1}$) when α is set to 0.

$$\mathcal{G}_{T1} = -\mathcal{G}_{N1} = -\frac{s \cdot e^{\frac{1}{2}sp_n}}{e^{\frac{1}{2}s(p_y-2m)} + e^{\frac{1}{2}sp_n}} \tag{10}$$

As the above equation shows, if we take p_v as $\hat{p}_v$ (**Equ.** 6), the difference of the proposed loss (**Equ.** 5) and CosFace only lies on the margin and the scale. The more detailed proof is included in the supplementary material.

3.2.3 Final Loss

For formal unity with CosFace, we rewrite the proposed loss into the following form.

$$\begin{aligned}\mathcal{L}_{GB-CosFace} = &-\frac{1}{2} log \frac{e^{2s(p_y-m)}}{e^{2s(p_y-m)} + e^{2sp_v}} \\ &- \frac{1}{2} log \frac{e^{2s(p_v-m)}}{e^{2s(p_v-m)} + e^{2sp_n}}\end{aligned} \tag{11}$$

Where $p_n = \frac{1}{s} log \sum_i e^{sp_i}$. The value of p_v is in accordance with **Equ.** 8. In training, p_v is a detached parameter which does not require gradients.

This is the final form of the proposed GB-CosFace. Under this formulation, the hyper-parameter α controls the degree of gradient imbalance. If we set α as 0, the gradients for the target score and the non-target scores are balanced, and the proposed GB-CosFace is equivalent to CosFace which has the margin of $2m$ and the scale of s.

3.3. Geometric Analysis

To analyze the properties of the proposed loss and compare it with other softmax-based losses, we analyze the loss boundaries in the binary classification case. The boundaries of ArcFace [5] and CosFace [36, 38] are determined by the following **Equ.** 12 and **Equ.** 13 respectively.

$$|arccos(P \cdot P_1) - arccos(P \cdot P_2)| = m \tag{12}$$

$$|P \cdot P_1 - P \cdot P_2| = m \tag{13}$$

Where P is the predicted normalized n-dimensional feature vector and n is the face feature dimension, P_1 and P_2 are the feature vectors of ID1 and ID2 respectively.

For normalized softmax loss, the boundary is determined by **Equ.** 13 or **Equ.** 12 with a zero margin. We set the angle between vector P_1 and P_2 as 60° and show the boundaries of normalized softmax, ArcFace, and CosFace in the 3D spherical feature space in **Figure** 5(a).

The boundaries of the proposed GB-CosFace can be determined according to **Equ.** 14.

$$\begin{cases} |P \cdot P_1| = p_v + m \\ |P \cdot P_2| = p_v + m \end{cases} \tag{14}$$

According to **Equ.** 6 and **Equ.** 8, in the binary classification case, p_v can be represented as follows.

$$p_v = \alpha p_{vg} + (1-\alpha)(P \cdot P_1 + P \cdot P_2)/2 \tag{15}$$

We show the boundaries of the proposed GB-CosFace loss in **Figure** 5(b), where p_{vg} is fixed to 0.625 (a reasonable value according to the experiments in **Section** 4) and m is fixed to 0.15.

In the face recognition problem, feature vectors of the same identity are expected to cluster together. However, by observing **Figure** 5(a), we can find that the boundaries in the case of binary classification do not meet this expectation. Only the positions near the line from point P_1 to point P_2 on the sphere can be effectively constrained. Fortunately, the training set has far more than two identities. Ideally, the prototypes of different identities will be evenly distributed on the sphere. The feature vectors of the same

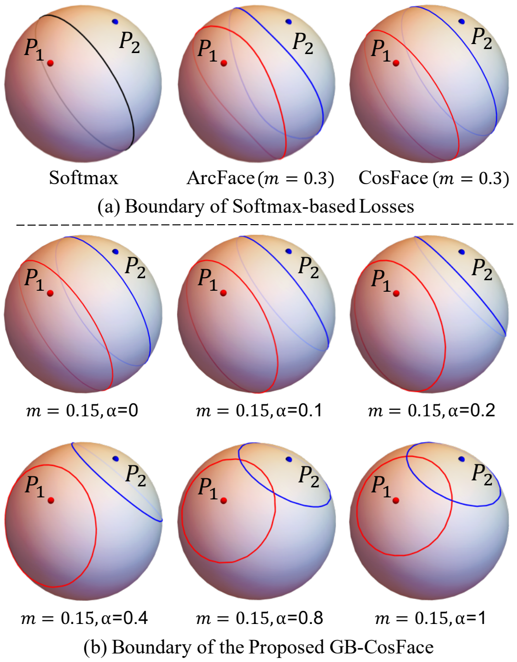


Figure 5. Boundaries of the softmax-based losses and the proposed GB-CosFace loss. P_1 and P_2 are two points at a distance of 60°. The red line and blue line are the target boundaries for P_1 and P_2 respectively. For the normalized softmax loss, the boundaries for P_1 and P_2 are coincident and represented in black color.

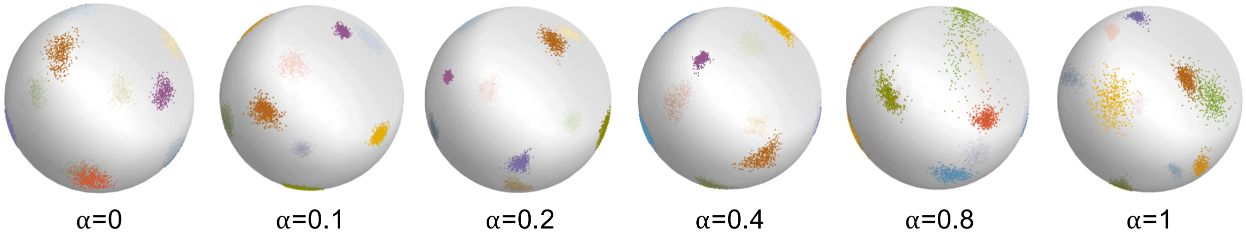


Figure 6. Visualization of the toy experiments on the proposed GB-CosFace. Different colors represent different identities.

identity will be constrained in all directions. But actually, it cannot be guaranteed that in the sparse high-dimensional spherical feature space, there are enough non-target prototypes evenly distributed around each training sample.

The proposed loss $\mathcal{L}_{GB}$ alleviates this problem by introducing a global boundary. As is shown in **Figure** 5(b), when $\alpha = 0$, the boundary is the same as CosFace. When $\alpha = 1$, the boundary is a circle on the sphere centered on P_1 or P_2 with a fixed radius completely determined by p_{vg} and the margin m. With the increase of α, the boundary is closer to the ideal open set classification objective. However, an excessively large α will cause blurring or even crossing of the boundaries between different identities.

To study the appropriate range of α, we conduct a toy experiment based on a seven-layer convolutional neural network on a small face recognition dataset containing ten identities. We set the feature dimension as three, and visualize the distribution of the feature vectors on the unit sphere under different α settings, as is shown in **Figure** 6. The margin is fixed to 0.15 and α is adjusted from 0 to 1. When $\alpha = 0$, our GB-CosFace is exactly the same as CosFace with the margin of 0.3, as indicated in **Section** 3.2. As α increases, e.g., $\alpha = 0.2$, the feature vectors of the same identity are more concentrated as expected. The model performance will deteriorate if α is further increased, e.g., $\alpha = 0.8$ or $\alpha = 1$. The setting of α will be studied in detail in the **Section** 4.3.

4. Experiments

In this section, we verify our GB-CosFace on two important face tasks: face recognition and face clustering. Furthermore, we conduct ablation experiments to verify the proposed strategies and the settings of the hyper-parameters.

Dataset. We employ MS1MV2 [6], a refined version of MS1M [11] as our training set for all the following experiments. This is a large-scale face recognition dataset containing 5.8M face images of 85K celebrities. We use several popular benchmarks as the validation set, including LFW [14], CFP-FP [28], CPLFW [45], AgeDB-30 [21], and CALFW [46]. And we use IJB-B [41] and IJB-C [19] as the testing sets. The details of the used datasets are shown in **Table** 1.

Dataset	#Identity	#Image	Split
MS1MV2 [6, 11]	85K	5.8M	train
LFW [14]	5749	13233	val
CFP-FP [28]	500	7000	val
AgeDB-30 [21]	568	16488	val
CALFW [46]	5749	12174	val
CPLFW [45]	5749	11652	val
IJB-B [41]	1845	76.8K	test
IJB-C [19]	3531	148.8K	test

Table 1. Details of used datasets.

Implementation Details. We use ResNet100 [13] as the backbone for all the following experiments. The BN-FC-BN structure is added after the last convolution layer to output 512-dimensional face feature vectors. For data prepossessing, all face images are set to 112 × 112 and normalized by utilizing five facial points following recent papers [5, 20]. Each RGB pixel is normalized to $[-1, 1]$. Random horizontal flip is the only data augmentation method employed in the training process. For optimization, the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and weight decay of $1e-4$ is adopted. We adopt the step learning rate decay strategy with an initial learning rate of 0.1. We train 24 epochs and divide the learning rate by 10 at 5, 10, 15, and 20 epochs. The training batch size is fixed to 512. Eight NVIDIA GPUS are employed for training. We fix the hyper-parameters s, m, α and γ as 32, 0.16, 0.15 and 0.01 respectively if not specified. For other methods, we follow the hyper-parameter settings in the original papers for fair comparison.

4.1. Face Recognition

We train the face recognition model on MS1MV2 [6] with the proposed GB-CosFace. **Figure** 7 shows the gradients and the global boundary p_v in the training process. It can be indicated that throughout the training process, the

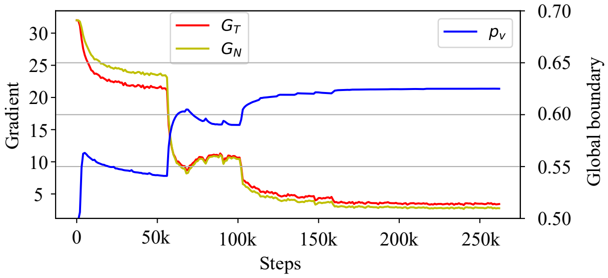


Figure 7. The value of the gradients and the global boundary p_v during the training process. In this figure, G_T is the gradient for the target score, G_N is the sum of gradients for the non-target scores, and p_v is the global boundary in **Equ.** 11.

gradients for the target score and the non-target scores are balanced, and the global boundary parameter p_v eventually converges to 0.625. Next, we test this model on multiple benchmarks as follows.

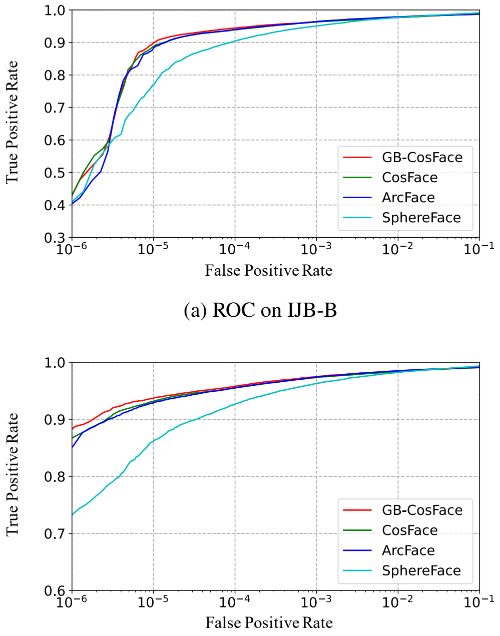
Results on Validation Datasets To compare with recent state-of-the-art competitors, we compare the results on several popular face recognition benchmarks, including LFW, CFP-FP, AgeDB-30, CALFW, and CPLFW. LFW focuses on unconstrained face verification. AgeDB-30 and CALFW are dedicated to large age variants face verification. CFP-FP and CPLFW aim at face verification with large pose changes.

Method	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
AM-Softmax [36]	99.50	95.10	95.68	94.38	89.48
CosFace [38]	99.78	98.26	98.17	**96.18**	92.18
SphereFace [17]	99.67	96.84	97.05	95.58	91.27
ArcFace [5]	99.81	**98.40**	98.05	95.96	92.72
GB-CosFace	**99.83**	**98.40**	**98.27**	96.02	**92.98**

Table 2. The face verification accuracy on the validation benchmarks.

The results are shown in **Tabel.** 2. We achieve the best results on four of the five benchmarks. The results on LFW, AgeDB-30, and CPLFW are improved by 0.02%, 0.1%, and 0.26% respectively compared to the current SOTA results.

Results on IJB-B and IJB-C. IJB is one of the largest and most challenging benchmarks to evaluate unconstrained face recognition. IJB-B contains 1845 identities with 55025 frames and 7011 videos. IJB-C is an extension of IJB-B which contains about 3.5K identities from 138K face images and 11K face videos. Based on these two benchmarks, we conduct the 1:1 face verification experiment using the official evaluation protocol. For each face image, we extract 512-dimensional features without flip operation.



(a) ROC on IJB-B

(b) ROC on IJB-C

Figure 8. ROC curves of face verification protocol on the IJB-B and IJB-C dataset.

Method	IJB-B(TAR)			IJB-C(TAR)		
	1e-6	1e-5	1e-4	1e-6	1e-5	1e-4
VGGFace2 [2]	-	67.10	80.00	-	74.70	84.00
CenterFace [40]	-	-	-	-	78.10	85.30
CircleLoss [31]	-	-	-	-	89.60	93.95
Softmax	46.73	75.17	90.06	64.07	83.68	92.40
SphereFace [17]	40.27	77.03	90.49	73.61	86.26	92.69
CosFace [38]	**48.94**	88.88	93.91	87.16	93.15	95.53
ArcFace [5]	38.72	88.39	93.99	86.54	92.90	95.52
GB-CosFace	44.97	**89.99**	**94.45**	**88.74**	**93.72**	**95.81**

Table 3. The face verification accuracy on IJB-B and IJB-C benchmarks.

The results are shown in **Table** 3. We achieve SOTA results on IJB-B and IJB-C. Compared to CosFace, our GB-CosFace improves 1.11% and 0.46% at TAR@FAR=1e-5, 1e-4 on IJB-B, and improves 1.58%, 0.57% and 0.28% at TAR@FAR=1e-6, 1e-5, 1e-4 on IJB-C. The ROC curves on IJB-B and IJB-C datasets are shown in **Figure** 8.

4.2. Face Clustering

The face clustering task aims to test the ability of the face recognition algorithm to recognize multiple instances of the same identity from various face images without a prior label. In this section, we compare the performances of CosFace and the proposed GB-CosFace through face clustering experiments to further investigate the feature representations learned by GB-CosFace. We compare the performance by applying various clustering methods to the feature representations of GB-CosFace and CosFace. For fair comparison, we re-implement the CosFace with the recommended hyper-parameters ($s = 30, m = 0.35$).

Method	Loss	IJB-B-512		IJB-B-1024		IJB-B-1845	
		F	NMI	F	NMI	F	NMI
K-means [18]	CosFace	68.54	89.39	68.33	89.96	**68.49**	90.35
	Ours	**68.80**	**89.55**	**68.56**	**90.19**	68.47	**90.57**
AHC [25]	CosFace	70.54	89.77	71.13	90.60	71.87	91.13
	Ours	**70.94**	**90.23**	**72.04**	**91.07**	**71.91**	**91.38**
DBSCAN [7]	CosFace	74.59	90.85	74.24	91.50	75.59	92.30
	Ours	**75.88**	**93.32**	**75.77**	**92.02**	**77.05**	**92.80**

Table 4. Comparison of clustering results based on different clustering methods.

We employ three sub-protocols (a total of seven) in the IJB-B dataset to test the ability of the algorithm to cluster on different scales. The identity numbers of the three sub-protocols are 512, 1024, and 1845 respectively, and the sample numbers are 18,251, 36,575, and 68,195 respectively. To illustrate the superiority of our method in clustering tasks, the typical clustering methods used in the experiment include K-means [18], AHC [25] and DBSCAN [7]. Normalized mutual information (NMI) and Bcubed F-measure [1] are the evaluation metrics selected in the clustering task. The experimental results are shown in **Table** 4. It can be indicated that our method surpasses CosFace on all the three sub-protocols and the facial feature representation obtained based on our GB-CosFace is more conducive to the clustering task.

4.3. Ablation Study

4.3.1 Effective of the Adaptive Boundary Strategy

p_v	IJB-B(TAR)			IJB-C(TAR)		
	1e-6	1e-5	1e-4	1e-6	1e-5	1e-4
0.585	42.77	87.48	94.02	85.43	92.69	95.52
0.625	40.92	87.53	94.24	84.46	92.99	95.79
0.665	44.56	87.98	93.94	85.71	92.73	95.38
Adaptive	**44.97**	**89.99**	**94.45**	**88.74**	**93.72**	**95.81**

Table 5. Comparison of the results of the proposed adaptive global boundary strategy and the fixed global boundary strategy.

To evaluate the effectiveness of the adaptive boundary strategy, we compare the fixed boundary strategy and the proposed adaptive boundary strategy in **Section** 3.2. We fix the p_v in our GB-CosFace(**Equ.** 11) to different values and keep the other experimental settings the same as **Section** 4.1. Since p_v converges to 0.625 in the above experiments in **Section** 4.1, we choose $p_v = 0.625$ for the experiment, and additionally choose two values near this value. The results are shown in **Table** 5. The proposed adaptive boundary strategy exceeds the performance of the fixed boundary strategy.

4.3.2 Hyperparameter Setting

Compared to CosFace, we introduce another hyperparameter α in **Equ.** 8. Since the settings of the scale parameter s and the margin parameter m have been studied in detail in the previous works [5, 36, 38], we empirically set $s = 32$ and $m = 0.16$ (equivalent to $m = 0.32$ in CosFace), and focus on the setting of α.

α	IJB-B(TAR)			IJB-C(TAR)		
	1e-6	1e-5	1e-4	1e-6	1e-5	1e-4
0.05	45.00	88.69	94.22	89.45	93.61	95.69
0.15	44.97	**89.99**	**94.45**	88.74	**93.72**	**95.81**
0.25	45.61	89.17	94.26	**89.53**	93.51	95.77
0.35	**50.54**	88.59	94.06	87.06	92.98	95.79

Table 6. The results of the proposed GB-CosFace under different settings of α.

As discussed in **Section** 3.3, α is an important parameter to control the gradient balance and the optimization objective. A too-large α will cause the boundaries of different identities to cross on the high-dimensional spherical feature space, resulting in unstable training. We conduct the controlled experiment where different values of α are set and other parameters are controlled unchanged. The results are shown in **Table** 6. Overall, the model performs best with $\alpha = 0.15$. Continue to increase α, the accuracy decreases. This result is consistent with the previous discussion and the toy experiments in **Section** 3.3.

5. Conclusion

In this work, we discuss the inconsistency between the training objective of the softmax-based loss and the testing process of face recognition and derive a new loss called global boundary CosFace(GB-CosFace) from the perspective of open set classification. Our GB-CosFace aligns the training objective with the face recognition testing process while inheriting the good properties of the softmax-based loss, based on the proposed adaptive global boundary strategy. Under this framework, CosFace is proved to be a special case of our GB-CosFace. Comprehensive experiments indicate that our GB-CosFace has an obvious improvement over general softmax-based losses.

References

[1] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 2009. 8

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 7

[3] Binghui Chen, Weihong Deng, and Junping Du. Noisy softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[4] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 1

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 5, 6, 7, 8

[6] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *International Conference on Computer Vision Workshops (ICCVW)*, 2019. 6

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 8

[8] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *European conference on computer vision (ECCV)*, 2018. 1

[9] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017. 2

[10] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision (ECCV)*, 2016. 6

[12] Chunrui Han, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Face recognition with contrastive convolution. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[14] 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. 6

[15] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[16] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 7

[18] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 1982. 8

[19] 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 *International Conference on Biometrics (ICB)*, 2018. 6

[20] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 6

[21] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 6

[22] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[23] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Association (BMVC)*, 2015. 1

[24] Oren Rippel, Manohar Paluri, Piotr Dollar, and Lubomir Bourdev. Metric learning with adaptive density discrimination. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[25] Warren S Sarle. Algorithms for clustering data. *Prentice Hall Advanced Reference Series : Computer Science*, 1990. 8

[26] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2012. 2

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[28] 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 *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016. 6

[29] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems (NIPS)*, 2016. 1

[30] Yi Sun. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems (NIPS)*, 2014. 1

[31] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 7

[32] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 1

[33] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[34] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[35] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In *Advances in neural information processing systems (NIPS)*, 2016. 1

[36] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 2018. 1, 2, 3, 4, 5, 7, 8

[37] 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 (ACM)*, 2017. 1, 3

[38] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 4, 5, 7, 8

[39] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 2021. 1, 2

[40] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision (ECCV)*, 2016. 1, 7

[41] 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 *Conference on Computer Vision and Pattern Recognition workshops (CVPRW)*, 2017. 6

[42] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[43] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[44] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[45] 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*, 2018. 6

[46] 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. 6

[47] Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[48] Yaoyao Zhong and Weihong Deng. Adversarial learning with margin-based triplet embedding regularization. In *International Conference on Computer Vision (ICCV)*, 2019. 1