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

Mingqiang Chen[✉], Lizhe Liu, Xiaohao Chen, and Siyu Zhu

Alibaba, Hangzhou, China
{mimingqiang.cmq,lizhe.llz,xiaohao.cxh,siting.zsy}@alibaba-inc.com

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 5.30%, 0.70%, and 0.36% 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 [1–4]. 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 [5–16] and proxy-based methods [17–29]. The proxy-free methods directly compress the intra-class distance and expand the inter-class distance based on pair-wise learning [5–8] or triplet learning [9–13,15,16]. However, when dealing with a large amount of training data, the hard-mining operation which is crucial for

---

M. Chen and L. Liu—These authors contributed equally to this work.

$$S(W_1, X_1) - m > S(W_2, X_1)$$
$$S(W_2, X_2) - m > S(W_1, X_2)$$

(a) Softmax Training Objective

$$S(X_{a1}, X_{a2}) > T$$
$$S(X_{a1}, X_{b1}) < T$$
$$S(X_{a2}, X_{b1}) < T$$
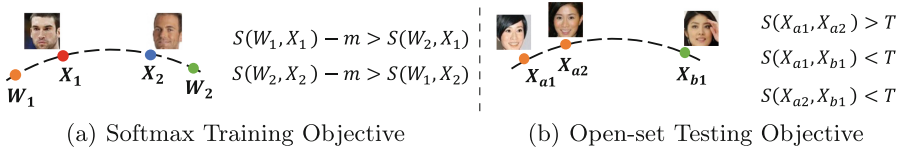
(b) Open-set Testing Objective

Fig. 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 **(a)**, X_1 and X_2 is the given training sample, m is the margin parameter. In **(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.

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.

Despite the great success of softmax-based face recognition, this strategy has its limitation from the perspective of the open set classification [30–33]. As is shown in Fig. 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 [1]. 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 Fig. 1(b). The inconsistency of the objective of training and testing limits the performance.

To reduce the impact of this inconsistency, 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 [18,19]. Typically, the similarity between two samples is represented by the cosine similarity of their corresponding feature vectors. Large-margin-based methods [18,20,21,23] are proposed to further compress the intra-class distance and expand the inter-class distance. Recently, the dynamic schemes for the scale parameter [34] and the margin parameter [26,35] have been studied and further improved the model performance.

From the perspective of training strategy, Lu et al. [36] proposed an optimal sampling strategy to address the inconsistency between the direction of gradient descent and optimizing the concerned evaluation metric. For face feature alignment, DAM [37] proposed a Discrepancy Alignment Metric, which introduces local inter-class differences for each face feature obtained from a pre-trained model, in the face verification stage. However, none of these methods consider introducing the global boundary in the testing process into the training objective.

In this paper, we propose a novel face recognition loss named global boundary CosFace (GB-CosFace), which resolve the above-mentioned inconsistencies well and can be easily applied for end-to-end training on face recognition task. In our GB-CosFace loss, 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. We further prove that CosFace [20,21] 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 testing objective of the open set classification while inheriting the advantages of the softmax-based loss. To the best of our knowledge, we are the first work which introduces a global boundary into the training objective for face recognition.
– 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 5.30%, 0.70%, and 0.36% 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

The training framework of the general softmax-based face recognition is shown in Fig. 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 extracting face features from 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 hyper-sphere. 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.
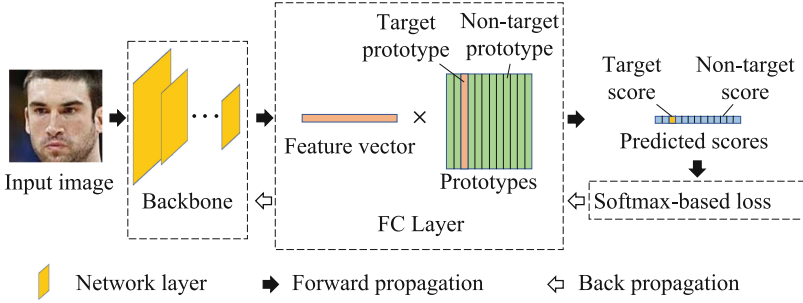
Fig. 2. The training framework of the general softmax-based face recognition.

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:

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

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".

We can reach several common softmax-based losses from Eq. 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 $\mathcal{O}_S$.

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

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 Eq. 2, we can visualize the decision boundaries of normalized softmax loss [19], CosFace [20,21], and ArcFace
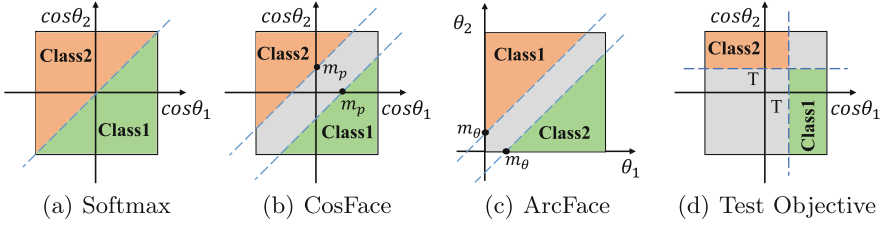
(a) Softmax   (b) CosFace   (c) ArcFace   (d) Test Objective

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

[23] under binary classification case, as is shown in Fig. 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 Fig. 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.

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

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 Eq. 2 as follows:

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

where we divide the training objective into the target score $\mathcal{O}_T$ and the non-target scores $\mathcal{O}_N$ respectively. p_y is the target score, where $p_y = cos\theta_y$. p_i is the non-target score, where $p_i = cos\theta_i$. 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 Eq. 2, we take the smooth form of $\mathcal{O}_T$ and $\mathcal{O}_N$ as the antetype of the proposed loss.

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

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 Sect. 2.3) of the softmax-based loss.

However, the proposed antetype introduces another problem: the setting of hyper-parameter p_v. First, the inappropriate setting of p_v may 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 Sect. 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 can be seen in Fig. 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. Therefore, an adaptive scheme for the global boundary is necessary.
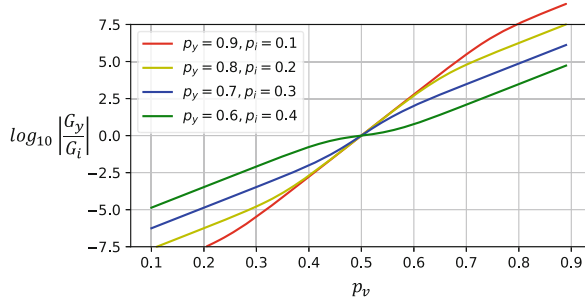
Fig. 4. The ratio of the target gradient to the non-target gradient varies with p_v under binary classification case using different p_y and p_i. Hyper-parameter s and m are set to 32 and 0.15 respectively. Note that the ordinate is the base 10 logarithm of the ratio.

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.

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 $\hat{p}_v$:

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

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.

Compatible with CosFace. In Eq. 8, if we take α as 0, the proposed loss will fully conform to Property 1 and Property 2 (stated in Sect. 2.3) of softmax-based loss. Through the following analysis, we can further find that Cosface [20,21] is a special case of the proposed loss when $\alpha = 0$.

The gradients based on CosFace is calculated as follows.

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

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 Eq. 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 $\hat{p}_v$ (Eq. 6), the difference of the proposed loss (Eq. 5) and CosFace only lies on the margin and the scale. The more detailed proof is included in the supplementary material.

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

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

where $p_n = \frac{1}{s} \log \sum_i e^{sp_i}$. The value of p_v is in accordance with Eq. 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 [23] and CosFace [20,21] are determined by the following Eq. 12 and Eq. 13 respectively.

$$|arccos(P \cdot P_1) - arccos(P \cdot P_2)| = m \qquad (12)$$

$$|P \cdot P_1 - P \cdot P_2| = m \qquad (13)$$

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.

(a) Boundary of Softmax-based Losses

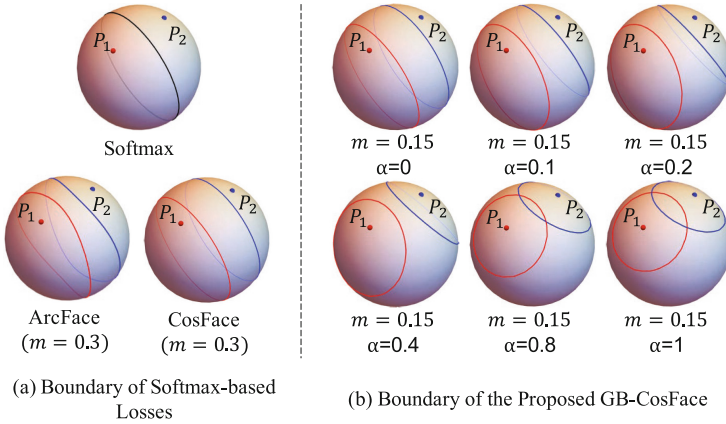(b) Boundary of the Proposed GB-CosFace

Fig. 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.(Color figure online)

For normalized softmax loss, the boundary is determined by Eq. 12 or Eq. 13 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 Fig. 5(a).

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

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

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

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

We show the boundaries of the proposed GB-CosFace loss in Fig. 5(b), where p_{vg} is fixed to 0.62 (a reasonable value according to the experiments in Sect. 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 Fig. 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 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 Fig. 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.

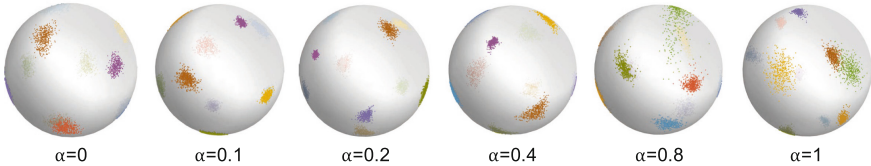α=0   α=0.1   α=0.2   α=0.4   α=0.8   α=1

Fig. 6. Visualization of the toy experiments on the proposed GB-CosFace. Different colors represent 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 Fig. 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 Sect. 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 Sect. 4.2.

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 MS1MV3 [38], a refined version of MS1M [39] as our training set for all the following experiments. This is a large-scale face recognition dataset containing 5.1M face images of 93K celebrities. We use several popular benchmarks as the validation set, including LFW [40], CFP-FP [41], CPLFW [42], AgeDB-30 [43], and CALFW [44]. And we use IJB-B [45] and IJB-C [46] as the testing sets.

Implementation Details. We use ResNet50 [47] and ResNet100 as the backbone for the following experiments. The BN-FC-BN structure is added after the last convolution layer to output 512-dimensional face feature vectors. For data pre-possessing, all face images are set to 112×112 and normalized by utilizing five facial points following recent papers [23,26]. Each RGB pixel is normalized to $[-1, 1]$. Random horizontal flip is the only data augmentation method employed in the training process. For optimization, we adopt the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and weight decay of 1e-4. 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.

4.1   Face Recognition

Analysis of Gradient Blance. Figure 7 shows the gradients and the global boundary p_v in the training process. Throughout the training process, the gradient of the target score G_T and the gradients sum of the non-target scores G_N maintain a same convergence trend, and the values of G_T and G_N are approximately equal after 50k iterations. The change trend of the global boundary parameter p_v during the training process is consistent with the gradients G_T and G_N, and eventually converges to 0.62. This result shows that our adaptive global boundary strategy can guarantee the stability of model training and keep G_T and G_N balanced throughout the training process, which is consistent with our discussion in Sect. 3.2.

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 fa ce verification.
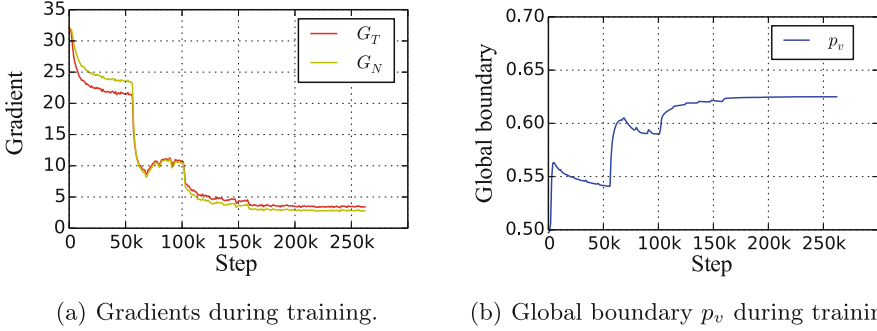
(a) Gradients during training.   (b) Global boundary p_v during training.

Fig. 7. G_T is the gradient of the target score, G_N is the gradients sum of the non-target scores, and p_v is the global boundary in Eq. 11.

The results are shown in Table 1. We achieve the best results on two of the five benchmarks. Even though both datasets are highly-saturated, our GB-CosFace surpasses the recent methods on CFP-FP and CPLFW, and achieves comparable results on other three datasets.

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.

The results are shown in Table 2. We achieve SOTA results on IJB-B and IJB-C. Compared to CosFace, our GB-CosFace improves 6.07%, 4.07% and 0.41% at TAR@FAR=1e-6, 1e-5, 1e-4 on IJB-B, and improves 5.30%, 0.70% and 0.36% at TAR@FAR=1e-6, 1e-5, 1e-4 on IJB-C.

Table 1. 1:1 verification accuracy is reported on the LFW, CFP-FP, AgeDB-30, CALFW, CPLFW datasets. Backbone network: ResNet100.

Method	Validation dataset				
	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
CosFace [21] (CVPR18)	99.81	98.12	98.11	95.76	92.28
ArcFace [23] (CVPR19)	**99.83**	98.27	98.28	95.45	92.08
Sub-center ArcFace [48] (ECCV20)	**99.83**	98.80	**98.45**	–	–
BroadFace [28] (ECCV20)	**99.83**	98.63	98.38	**96.20**	93.17
CurricularFace [25] (CVPR20)	99.80	98.37	98.32	**96.20**	93.13
URFace [49] (CVPR20)	99.78	98.64	–	–	–
CosFace+SCF [50] (CVPR21)	99.80	98.59	98.26	96.18	93.26
MagFace [26] (CVPR21)	**99.83**	98.46	98.17	96.15	92.87
GB-CosFace	99.80	**98.84**	98.31	96.15	**93.55**

Table 2. The face verification accuracy on IJB-B and IJB-C. We evaluated the TAR@FAR from 1e-4 to 1e-6. Backbone network: ResNet100.

Method	IJB-B(TAR)			IJB-C(TAR)		
	1e-6	1e-5	1e-4	1e-6	1e-5	1e-4
CosFace [21] (CVPR18)	36.49	88.11	94.80	85.91	94.10	96.37
ArcFace [23] (CVPR19)	38.28	89.33	94.25	89.06	93.94	96.03
Sub-center ArcFace [48] (ECCV20)	-	-	**95.25**	-	-	96.61
BroadFace [28] (ECCV20)	46.53	90.81	94.61	90.41	94.11	96.03
CurricularFace [25] (CVPR20)	-	-	94.80	-	-	96.10
GroupFace [51] (CVPR20)	**52.12**	91.24	94.93	89.28	94.53	96.26
CosFace+DAM [37] (ICCV21)	-	-	94.97	-	-	96.45
CosFace+SCF [50] (CVPR21)	-	91.02	94.95	-	94.78	96.22
MagFace [26] (CVPR21)	40.91	89.88	94.33	89.26	93.67	95.81
GB-CosFace	42.56	**92.18**	95.21	**91.21**	**94.80**	**96.73**

4.2   Ablation Study

To analyze the effect of the adaptive boundary strategy and the setting of hyper-parameter α, we train ResNet-50 networks on MS1MV3 with different settings and evaluated the TAR@FAR=1e-4 on IJB-C.

Hyperparameter Setting. Compared to CosFace, we introduce another hyper-parameter α in Eq. 8. Since the settings of the scale parameter s and the margin parameter m have been studied in detail in the previous works [20, 21, 23], we empirically set $s = 32$ and $m = 0.16$ (equivalent to $m = 0.32$ in CosFace), and focus on the setting of α. For more detailed theoretical analysis, please refer to Sect. 3.3.

Table 3. The results of the proposed GB-CosFace under different settings of α.

Settings	IJB-C(TAR)
FAR=1e-4, R50, adaptive p_v, $\alpha = 0$	96.10
FAR=1e-4, R50, adaptive p_v, $\alpha = 0.05$	96.15
FAR=1e-4, R50, adaptive p_v, $\alpha = 0.15$	96.24
FAR=1e-4, R50, adaptive p_v, $\alpha = 0.25$	**96.35**
FAR=1e-4, R50, adaptive p_v, $\alpha = 0.35$	96.33
FAR=1e-4, R50, adaptive p_v, $\alpha = 0.60$	96.08

We conduct the controlled experiment where the value of α is set from 0 to 0.6 and other parameters are fixed. The results are shown in Table 3. When the value of α gradually increases from 0 to 0.25, the performance of the model

gradually improves and the model performs best with $\alpha = 0.25$. After the value of α exceeds 0.35, the model performance obviously degenerates with the increase of α. Overall, the performance of the model can maintain relatively good results as the value of α is between 0.15 and 0.35. This result is consistent with the previous discussion and the toy experiments in Sect. 3.3.

Effect of the Adaptive Boundary Strategy. To evaluate the effectiveness of the adaptive boundary strategy, we compare the fixed boundary strategy and the proposed adaptive boundary strategy in Sect. 3.2. We fix the p_v in our GB-CosFace(Eq. 11) to different values and keep the other experimental settings the same as Sect. 4.1. Since p_v converges to 0.62 in the experiment in Sect. 4.1, we choose $p_v = 0.62$ and additionally choose values near 0.62.

Table 4. Comparison of the results of the proposed adaptive global boundary strategy and the fixed global boundary strategy.

Settings	IJB-C(TAR@FAR$=$1e-4)
FAR$=$1e-4, R50, $\alpha = 0.15$, $p_v = 0.50$	91.19
FAR$=$1e-4, R50, $\alpha = 0.15$, $p_v = 0.58$	**96.27**
FAR$=$1e-4, R50, $\alpha = 0.15$, $p_v = 0.62$	96.19
FAR$=$1e-4, R50, $\alpha = 0.15$, $p_v = 0.66$	96.17
FAR$=$1e-4, R50, $\alpha = 0.15$, $p_v = 0.74$	95.09
FAR$=$1e-4, R50, $\alpha = 0.15$, adaptive p_v	96.24

The results are shown in Table 4. For the fixed boundaries, the model performs best when $p_v = 0.58$ and gets worse rapidly when the p_v changes, eg. the TAR decreases to 91.19% when $p_v = 0.50$. What's more, if we reduce p_v to 0.42 or increase it to 0.82, the training will not converge. For the adaptive boundary strategy, the performance is very close to the best fixed boundary strategy result. This indicates that for the fixed boundary strategy, the model performance is sensitive to the value of p_v, a very careful setting of p_v is required to obtain good results. While the adaptive global boundary strategy does not require careful tuning of hyper-parameters to achieve a similar performance. This result is consistent with the previous discussion and the toy experiments in Sect. 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 from the perspective of open set classification, called the global boundary CosFace(GB-CosFace). Our GB-CosFace aligns the training objective with

the face recognition testing process while inheriting the good properties of the softmax-based loss. Comprehensive experiments indicate that our GB-CosFace has an obvious improvement over general softmax-based losses.

References

1. Wang, M., Deng, W.: Deep face recognition: a survey. Neurocomputing **429**, 215–244 (2021)
2. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
3. Sun, Y., Liang, D., Wang, X., Tang, X.: DeepID3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
4. Wen, Y., Zhang, K., Li, Z., Qiao, Yu.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
5. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2005)
6. Sun, Y.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems (NIPS) (2014)
7. Ustinova, E., Lempitsky, V.: Learning deep embeddings with histogram loss. In: Advances in Neural Information Processing Systems (NIPS) (2016)
8. Han, C., Shan, S., Kan, M., Wu, S., Chen, X.: Face recognition with contrastive convolution. In: European Conference on Computer Vision (ECCV) (2018)
9. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
10. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Association (BMVC) (2015)
11. Ge, W.: Deep metric learning with hierarchical triplet loss. In: European Conference on Computer Vision (ECCV) (2018)
12. Zhong, Y., Deng, W.: Adversarial learning with margin-based triplet embedding regularization. In: International Conference on Computer Vision (ICCV) (2019)
13. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
14. Rippel, O., Paluri, M., Dollar, P., Bourdev, L.: Metric learning with adaptive density discrimination. In: International Conference on Learning Representations (ICLR) (2015)
15. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Advances in Neural Information Processing Systems (NIPS) (2016)
16. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: International Conference on Computer Vision (ICCV) (2017)
17. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)

18. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: deep hypersphere embedding for face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
19. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM International Conference on Multimedia (ACM) (2017)
20. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**(7), 926–930 (2018)
21. Wang, H., et al.: Cosface: large margin cosine loss for deep face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
22. Sun, Y., et al.: Circle loss: a unified perspective of pair similarity optimization. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
24. Zheng, Y., Pal, D.K., Savvides, M.: Ring loss: convex feature normalization for face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
25. Huang, Y., et al.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
26. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: a universal representation for face recognition and quality assessment. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
27. Chen, B., Deng, W., Du, J.: Noisy softmax: improving the generalization ability of DCNN via postponing the early softmax saturation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
28. Kim, Y., Park, W., Shin, J.: BroadFace: looking at tens of thousands of people at once for face recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 536–552. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_31
29. Deng, J., Guo, J., Yang, J., Lattas, A., Zafeiriou, S.: Variational prototype learning for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
30. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1757–1772 (2012)
31. Geng, C., Huang, S.J., Chen, S.: Recent advances in open set recognition: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3614–3631 (2020)
32. Ge, Z., Demyanov, S., Chen, Z., Garnavi, R.: Generative openmax for multi-class open set classification. arXiv preprint arXiv:1707.07418 (2017)
33. Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., Naemura, T.: Classification-reconstruction learning for open-set recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
34. Zhang, X., Zhao, R., Qiao, Y., Wang, X., Li, H.: Adacos: adaptively scaling cosine logits for effectively learning deep face representations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
35. Liu, H., Zhu, X., Lei, Z., Li, S.Z.: Adaptiveface: adaptive margin and sampling for face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

36. Lu, J., Xu, C., Zhang, W., Duan, L.Y., Mei, T.: Sampling wisely: deep image embedding by top-k precision optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)

37. Liu, J., et al.: DAM: discrepancy alignment metric for face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

38. Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S.: Lightweight face recognition challenge. In: International Conference on Computer Vision Workshops (ICCVW) (2019)

39. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-celeb-1M: a dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 87–102. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_6

40. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition (2008)

41. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2016)

42. Zheng, T., Deng, W.: Cross-pose LFW: a database for studying cross-pose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Technical report (2018)

43. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: AgeDB: the first manually collected, in-the-wild age database. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)

44. Zheng, T., Deng, W., Hu, J.: Cross-age LFW: a database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)

45. Whitelam, C., et al.: IARPA janus benchmark-b face dataset. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)

46. Maze, B., et al.: IARPA janus benchmark-C: face dataset and protocol. In: International Conference on Biometrics (ICB) (2018)

47. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

48. Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center ArcFace: boosting face recognition by large-scale noisy web faces. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 741–757. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_43

49. Shi, Y., Yu, X., Sohn, K., Chandraker, M., Jain, A.K.: Towards universal representation learning for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6817–6826 (2020)

50. Li, S., Xu, J., Xu, X., Shen, P., Li, S., Hooi, B.: Spherical confidence learning for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)

51. Kim, Y., Park, W., Roh, M.C., Shin, J.: Groupface: learning latent groups and constructing group-based representations for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)