An Equalized Margin Loss for Face Recognition

Jingna Sun, Wenming Yang, Jing-Hao Xue, and Qingmin Liao

***Abstract*—In this paper, we propose a new loss function, termed the equalized margin (EqM) loss, which is designed to make both intra-class scopes and inter-class margins similar over all classes, such that all the classes can be evenly distributed on the hypersphere of the feature space. The EqM loss controls both the lower limit of intra-class similarity by exploiting hard-sample mining and the upper limit of inter-class similarity by assuring equalized margins. Therefore, using the EqM loss, we can not only obtain more discriminative features, but also overcome the negative impacts from the data imbalance on the inter-class margins. We also observe that the EqM loss is stable with the variation of the scale in normalized Softmax. Furthermore, by conducting extensive experiments on LFW, YTF, CFP, MegaFace and IJB-B, we are able to verify the effectiveness and superiority of the EqM loss, compared with other state-of-the-art loss functions for face recognition.**

***Index Terms*—Face recognition; equalized margin (EqM) loss; intra-class scope; inter-class margin; deep learning**

I. Introduction

CONVOLUTIONAL neural networks (CNNs) have achieved great success in many fields [1]–[5], of which the most significant one is arguably face recognition [6]–[13], [13]–[20]. Face recognition contains two tasks: face identification and face verification. Face identification is to identify which person a face image belongs to, while face verification focuses on whether two face images belong to a same person. For face identification or face verification, there are two scenarios of evaluation [21], [22]. One scenario is called closed-set face recognition, where the identity of a test face has been predefined in the training set so that we can obtain the recognition accuracy by predicting the label of the test image and then comparing the predicted label to the true label. Closed-set face recognition has a defect that the training set usually cannot cover all possible face labels. In the other scenario called open-set face recognition, the identity of a test face has not necessarily been predefined in the training set, and we achieve face recognition through measuring the similarity of two faces between their features obtained by a face-recognition network. Due to the limit of the closed-set evaluation, recent studies often adopt the open-set face recognition. In this paper, we also apply the open-set face recognition to verify the effectiveness of different loss functions.

Therefore, to improve the accuracy of face recognition, we want to find a feature space in which the features from images of the same person (i.e. the intra-class features) are as close to each other as possible, while the features for different persons are as far away from each other as possible. This leads to two objectives that we aim to achieve while developing a face recognition network: enhancing intra-class similarity and reducing inter-class similarity. Many studies [3], [14], [17]–[19], [23]–[25] have achieved promising face recognition performances by pursuing the above two objectives. Here we review their achievement in face recognition from four methodological aspects, metric learning, hard-sample mining, margin enlarging, and data-imbalance mitigating, as follows.

Metric learning methods [12], [14], [23] mainly focus on reducing intra-class distance and enlarging inter-class distance through optimizing the distance metric of features. There are two metrics often used, the Euclidean distance and the cosine distance. When the facial features are normalized, the Euclidean distance can be regarded as the cosine distance. Metric learning can promote face recognition accuracy, however it was highly dependent on how to choose suitable face image pairs or triplets to constrain the learning. Additionally, its efficiency was very low compared with other methods, especially when the amount of training data is enormous.

To some extent, the identification of hard samples can be a reasonable measure for the capacity of a face recognition system. [26], [27] paid more attention to the hard samples by giving them heavier punishments. Hard-sample mining methods usually contain two steps: first, define what hard samples are; and second, give more severe penalties to the hard samples. In this way, the face recognition network can learn more discriminative features.

Recently, margin-based methods have become the mainstream to improve the performance of face recognition. Many methods [17]–[19], [24], [25] had achieved start-of-the-art face recognition performance. These methods optimized the angle between features and added a constant margin between classes to obtain discriminative features. [28] found that the l_2-norm of features can correspond to the quality of face images, and thus to eliminate the influence of different quality images, they proposed to constrain the l_2-norm of features to a constant and achieved great improvement in recognition. SphereFace [21] was proposed to normalize the wights of the classifier and added a margin between classes. NormFace [29] normalized both the weights and the features. [17]–[19] also normalized the wights and the features and put forward a constant margin additionally to make sure a reasonable inter-class distance.

There are also some studies [30]–[32] focusing on mitigating the harms brought by the data imbalance. These studies paid more attention to the people with few images (i.e. the minority class) to avoid the undesirable bias of the face recognition network toward the people with many images (i.e. the majority class).

Jingna Sun, Wenming Yang, Qingmin Liao are with Shenzhen Key Lab. of Info. Sci&Tech/Shenzhen Engineering Lab. of IS&DCP, Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China. e-mail: (sunjn17@mails.tsinghua.edu.cn, yangelwm@163.com, liaoqm@sz.tsinghua.edu.cn).

Jing-Hao Xue is with the Department of Statistical Science, University College London, UK. e-mail: (jinghao.xue@ucl.ac.uk).

The four approaches discussed above improve the face recognition accuracy from different aspects. In this paper, we aim to exploit their advantages while avoiding their defects. We observe that margin-based methods only constrain a margin between classes while no considering intra-class discrepancy. This can be undesirable when data are imbalanced, because the majority classes can occupy much more space in the hypersphere as their size and diversity are much larger than those of the minority classes, while margin-based methods adopt a constant margin regardless of specific classes. We provide a simple illustration in Fig. 1(a), where the red dots represent the samples from the majority class while the green dots for the minority class. It is clear that there are two main problems which will affect the performance of face recognition. Firstly, the network will bias toward the majority class. Secondly, it is hard for margin-based methods to get a compact majority class.

Therefore, in our EqM loss, while absorbing the strength of margin-based methods to constrain the inter-class distance, we also exploit the superiority of hard-sample mining to obtain a small intra-class scope. Moreover, we constrain the intra-class scope to overcome the adverse effect of imbalanced data. As shown in Fig. 1(b), we control the majority class and the minority class with the same intra-class range. With constraints to make all classes alike on both the intra-class scope and the inter-class margin, we aim to achieve a fair distribution of features for all classes, and thus to relieve the negative effect of data imbalance. In our experiments, not only is observed the superior recognition accuracy from applying our proposed EqM loss, but also we find that the EqM loss is stable when the dataset is class-imbalanced and the s in Eq.(1) changes.

In short, main contributions of this paper are threefold.

1) We propose an equalized margin (EqM) loss to improve the performance of face recognition. Using the EqM loss, we can appropriately reduce the intra-class distance and expand the inter-class margin, through setting balanced intra-class and inter-class scopes over all classes. The EqM loss exploits the advantages of hard-sample mining and margin controlling, and addresses the data imbalance issue, in face recognition.

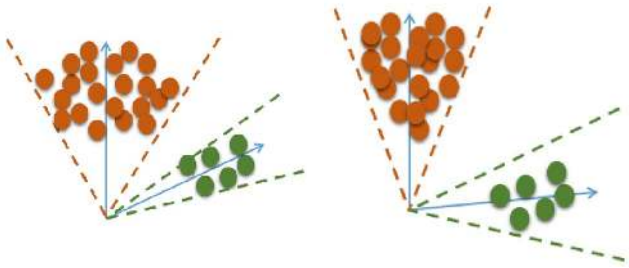
2) We find that our EqM loss is stable with the change of s in Eq.(1), a parameter adopted by the margin-based methods to make the network easier to converge. We also verify that the two hyper-parameters of the EqM loss can adequately control the intra-class scope and the inter-class scope. That is, the EqM loss can be more flexible and stable compared with other state-of-the-art loss functions for face recognition.

3) Through extensive experiments on LFW, YTF, CFP, MegaFace and IJB-B to evaluate the performance of different loss functions for face recognition, we observe that our EqM loss performs better than or is comparable to other state-of-the-art loss functions.

II. Related Work

Margin-Based Methods. There are many studies [17]–[19], [21] focus on enlarging the margin between different classes. The form of these loss functions can be summarized as

$$L_{margin} = -\log(p), \tag{1}$$



(a) (b)

Fig. 1. (a) The distributions of a majority class and a minority class in margin-based methods. (b) The target distributions of the majority class and the minority class by the limit of our EqM loss. The red dots represent the samples from the majority class and the green dots are the samples from the minority class.The blue arrows are the corresponding class weights in the classifier. The class scopes are marked by the dotted lines.

in which

$$p = \frac{1}{1 + \sum_{j \neq y_i}^{C} e^{s\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}},$$

where s is the parameter in the scale layer in [29] and C represent the the number of classes, respectively; $\widetilde{x}_i$ indicates the ith normalized sample feature whose label is y_i; $\widetilde{w}$ is the normalized weights of the network, with $\widetilde{w}_j$ the jth column of the weights $\widetilde{w}$ for the jth class.

The difference among the loss functions in studies [17]–[19], [21], [33] is mainly with the form of $\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})$:

$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = \cos\theta_{i,j} - \cos(m\theta_{i,y_i}), \tag{2}$$
$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = g(\theta_{i,j}) - g(m\theta_{i,y_i}), \tag{3}$$
$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = \cos\theta_{i,j} - \cos(\theta_{i,y_i} + m), \tag{4}$$
$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = \cos\theta_{i,j} - \cos\theta_{i,y_i} + m, \tag{5}$$

in which $\theta_{i,j}$ and θ_{i,y_i} represent the angles between $\widetilde{x}_i$ and $\widetilde{w}_j$, and $\widetilde{x}_i$ and $\widetilde{w}_{y_i}$, respectively. Eq.(2) is for the angular softmax (A-Softmax) loss [21], which adds a hyper-parameter m ($m \geq 1$) to produce different decision boundaries for different classes and thus enlarge the inter-class margin. Eq.(3) is called GA-Softmax, where $g(\cdot)$ can be a function in the form of linear, cosine and sigmoid [33]. Eq.(4) shows the idea of ArcFace in [19], which adds a fixed angular margin m between classes to attain a better convergence than Eq.(2). The motivation of the additive margin (AM) loss in [17], [18] can be written as Eq.(5), in which the margin m represents the fixed cosine margin between classes. The difference of the scale s in Eq.(1) between the studies [17]–[19], [21] is that the s of the A-Softmax loss is the l_2-norm of x_i, while the s of the other loss functions is a constant.

In short, these studies all introduce a margin to expand the inter-class margin, although from different perspectives. However, these margin-based methods focus on the large inter-class distance while ignoring the difference in the class characteristics.

Hard-Sample Mining Methods. As the name suggests, hard-sample mining methods pay more attention to hard samples, by punishing hard samples distinctively from easy samples through rescaling their corresponding weights. It

can help the network extract more discriminative features by increasing the importance of hard samples. For hard-sample mining methods, a classical approach is the focal loss [26]:

$$L_{focal} = (1-p)^{\gamma}\log(p), \tag{6}$$

with p as in Eq.(1) and

$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = \cos\theta_{i,j} - \cos\theta_{i,y_i}.$$

The focal loss determines the hard samples according to the value of p and designs a hyper-parameter γ to modulate the relative importance of easy samples and hard samples. The smaller the p, the heavier the punishment. A recent work [27] also punishes hard samples more, in which the hard samples are defined as the samples which satisfy the following condition: $\cos\theta_{i,y_i} < \cos\theta_{i,j}$. Hard-sample mining methods are conducive to promote the capability of the network. In [26], [27], the hard samples are defined by considering both the intra-class distance and inter-class variance; in our work, we define the hard samples by only considering the intra-class distance for compact individual classes.

Imbalanced Data Problem. It is well known that imbalanced data can hamper the network to learn more discriminative features. Many studies [30]–[32] apply various measures to mitigate the imbalanced data problem. [30] alleviates the influence of the long tail of the data by reducing intra-personal variance and enlarging inter-personal differences within a mini-batch. [31] optimizes the network by metric learning of clusters, which can draw balanced class boundaries. [32] pays more attention to the minority classes in a mini-batch from several aspects. All the methods above-mentioned relieved the harms caused by imbalanced data and obtained improved recognition accuracy. In this paper, we also consider the influence of imbalanced data, in particular the impact of the majority class on the value of m in the margin-based methods.

III. Proposed Method

A. The EqM Loss

In this paper, we propose a new loss function named the equalized margin (EqM) loss for face recognition, which exploits the advantages of both hard-sample mining methods and margin-based methods, and relieves the harms caused by imbalanced data. Mathematically, the novelty of the EqM Loss is with a new definition of function $\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})$:

$$\begin{aligned}\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = \cos\theta_{i,j} - \cos\theta_{i,y_i} \\ + |\cos\theta_{i,y_i} - t_1| + |\cos\theta_{i,j} - t_2| + t_1 - t_2,\end{aligned} \tag{7}$$

where t_1 and t_2 are two hyper-parameters. As $\cos\theta_{i,y_i}$ can be regarded as the measure of intra-class similarity (the larger the better) and $\cos\theta_{i,j}$ as inter-class similarity (the smaller the better), t_1 represents the lower limit that we expect the intra-class similarity not to be lower (i.e. we expect the intra-class compactness to be sufficiently large), and t_2 expresses the upper limit that we expect the inter-class similarity not to be higher (i.e. we expect the inter-class margin to be sufficiently large). This intuition also leads to different signs of t_1 and t_2 in the offset term $t_1 - t_2$. Although we can treat t_1 and t_2 as lower/upper limits, using the absolute value in $|\cos\theta_{i,y_i} - t_1|$ and in $|\cos\theta_{i,j} - t_2|$ actually allows for some relaxation.

To obtain a deep understanding of the EqM loss, we can analyze Eq.(7) in four scenarios by considering the different intervals of $\cos\theta_{i,j}$ and $\cos\theta_{i,y_i}$. Recall that we expect $\cos\theta_{i,y_i}$ to be as large as possible in order to obtain compact intra-class features, and we expect $\cos\theta_{i,j}$ to be as small as possible to attain a large inter-class margin.

1) $\cos\theta_{i,y_i} \geq t_1$, $\cos\theta_{i,j} \geq t_2$. In this case, Eq.(7) is transformed as

$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = 2(\cos\theta_{i,j} - t_2). \tag{8}$$

That is, the EqM loss is now irrelevant to $\cos\theta_{i,y_i}$ and minimising the loss is focused on reducing the value of $\cos\theta_{i,j}$ to the expected limit t_2; in other words, it is to increase the inter-class margin such that ϕ goes down to zero.

2) $\cos\theta_{i,y_i} \geq t_1$, $\cos\theta_{i,j} \leq t_2$. In this situation, we find that

$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = 0. \tag{9}$$

This is the desirable case that the intra-class similarity is sufficiently large ($\cos\theta_{i,y_i} > t_1$) and the inter-class similarity is sufficiently small ($\cos\theta_{i,j} < t_2$), hence Eq.(9) makes sense.

3) $\cos\theta_{i,y_i} \leq t_1$, $\cos\theta_{i,j} \geq t_2$. In this case, we have

$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = 2(\cos\theta_{i,j} - \cos\theta_{i,y_i} + t_1 - t_2). \tag{10}$$

In contrast to the case 2 above, this case is undesired, as both $\cos\theta_{i,y_i}$ and $\cos\theta_{i,j}$ do not satisfy the limits that we expect. Hence in this case, the EqM loss suggests for optimization of both the intra-class distance and inter-class margin. For example, an ideal result from the training is to make $\cos\theta_{i,y_i}$ and $\cos\theta_{i,j}$ reach t_1 and t_2, respectively, such that ϕ goes down to zero.

4) $\cos\theta_{i,y_i} \leq t_1$, $\cos\theta_{i,j} \leq t_2$. This is the opposite case of case 1, hence we have

$$\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = 2(-\cos\theta_{i,y_i} + t_1), \tag{11}$$

and minimizing the EqM loss is focused on increasing the value of $\cos\theta_{i,y_i}$ to its expected limit t_1, i.e. on increasing the intra-class similarity, such that ϕ goes down to zero.

Based on the analysis above, we can make two remarks. Firstly, the aim of minimizing Eq.(7) is to ensure the intra-class similarity to be close to or above the limit t_1, as well as to ensure the inter-class margin to be close to or above the limit t_2. These two hyper-parameters of limits provide an equal control on the intra-class scopes and the inter-class scopes, unbiased for all classes. That is, the use of t_1 and t_2 demonstrates that we treat every class fairly to circumvent the negative influence of data imbalance. Because of this, we call the proposed loss function the equalized margin (EqM) loss. A schematic diagram of the EqM loss is shown in Fig. 2.

Secondly, the EqM loss pays attention to different points in different scenarios. In case 1 and case 4, the EqM loss only focuses on either enlarging the inter-class margin or increasing the intra-class similarity. In case 3, the EqM loss optimizes the inter-class margin and the intra-class compactness simultaneously. This adaptive scheme to different situations of $\cos\theta_{i,j}$ and $\cos\theta_{i,y_i}$ is different from those of [17]–[19], [21], and is expected to be more flexible and powerful for face recognition.

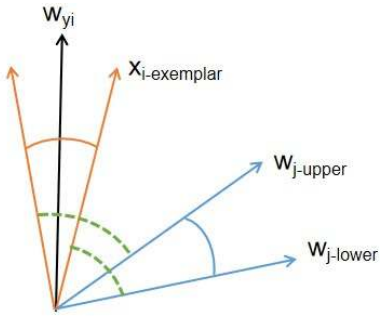


Fig. 2. A schematic diagram of the EqM loss. The orange solid arrows represent the samples x_i that satisfy the intra-class limit t_1; and the blue solid arrows indicate the jth class weights w_j that satisfy the inter-class limit t_2 for x_i. The green dotted arcs illustrate the inter-class margin between x_i and w_j, controlled by the inter-class limit t_2. The orange arc illustrates the intra-class scope of the samples and the blue arc presents the corresponding range of the weights, controlled by the intra-class limit t_1.

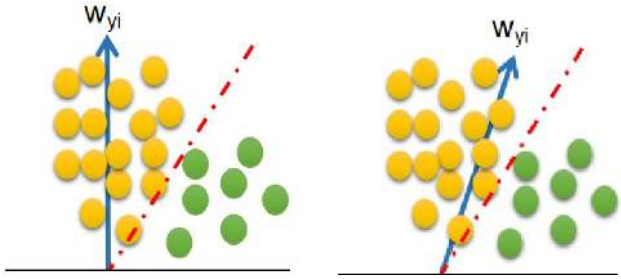


Fig. 3. The influence of hard-sample mining. The yellow solid dots and the green solid dots represent the easy samples and the hard samples, respectively. The blue solid arrow indicates the weights of class y_i. (a) The origin distribution of class y_i. (b) The distribution of class y_i after penalizing hard-samples more heavily.The dotted red line is the boundary between easy samples and hard samples.

B. The superiority of the EqM loss

Easy Sample vs. Hard Sample. In this paper, hard samples are defined as the samples satisfying $\cos\theta_{i,y_i} < t_1$. As usual, we punish the hard samples more heavily to compensate the quantitative advantage of the easy samples. The degree of punishment should be modest because excessive punishment will make the weights overly bias toward hard samples, which may reduce the recognition accuracy of easy samples. We illustrate this in Fig. 3. This phenomenon can be intuitively explained from the perspective of back-propagation. When the Softmax loss is employed to train a face recognition network, the updating of w_{y_i} can be expressed as

$$\frac{\partial L}{\partial \widetilde{w}_{y_i}} = -\frac{1}{N}\sum_{i=1}^{N} s\widetilde{x}_i(1-p), \tag{12}$$

where N is the number of samples in a mini-batch. From Eq.(12), we can find that the updating of w_{y_i} is based on a weighted sum of samples from class y_i. When the punishment on hard samples is too severe, w_{y_i} will be biased toward the hard samples heavily and the easy samples will be negligible. Hence we should be able to adjust the penalty for hard samples

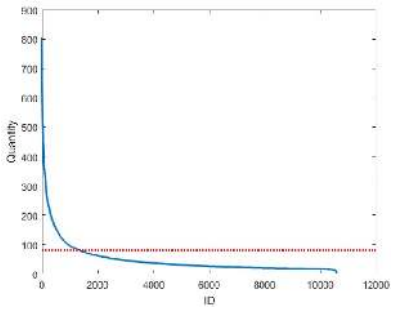


Fig. 4. The quantity distribution (in blue) of each class in CASIA-WebFace. The x-axis indexes the class and the y-axis is the quantity of each class. The red dotted line indicates that "quantity = 80".

to reach an appropriate balance point between easy samples and hard samples.

In the EqM loss, we implement t_1 to adjust the relation between easy samples and hard samples. When the intra-class similarity $\cos\theta_{i,y_i}$ for a sample $\widetilde{x}_i$ is lower than t_1, we treat it as a hard sample and punish the samples more heavily. In contrast, for an easy sample with intra-class similarity $\cos\theta_{i,y_i}$ higher than t_1, from Eq.(8) and Eq.(9), we can see the easy sample is not involved in the optimization. In Section IV-C, we will also find that the recognition accuracy is not always better with a larger value of t_1, which demonstrates that a balance point is needed between easy samples and hard samples.

Imbalanced Data. In Fig. 4, we show the quantity of each class in CASIA-WebFace [34]. [30] focused on the long tail of the data and demonstrated that the long tail affects the performance of the recognition network. In the paper, we pay more attention to the "heavy head" of Fig. 4 and discuss the influence of the "heavy head" on the inter-class margin. In the experiments, the "heavy head" refers to the majority classes, which will limit the increase of m in the margin-based methods. Take the AM loss in Eq.(5) as an example and apply the experimental settings for CASIA-WebFace in Section IV-B. To illustrate the influence of the "heavy head", we randomly select 80 face images from each class who has more than 80 image to form a balanced training dataset, as indicated by the red dotted line in Fig. 4. In the experiments, we train the network using the original CASIA-WebFace dataset and the balanced CASIA-WebFace dataset (i.e. the new dataset with each class having less than or equal to 80 images), respectively, and obtain various test accuracy with an increasing m.

The test accuracy on LFW [35] with different m by using AM loss in Eq.(5) is shown in Fig. 5. From the figure, we can observe that the turning point of the orange curve (using the network trained with the original CASIA-WebFace dataset) is smaller than the blue line (for the balanced CASIA-WebFace dataset), which demonstrates that the "heavy head" in the original dataset hinders the increase of m, although in general we prefer a larger m (i.e. a larger inter-class margin). In Table I, we also present the test accuracies on four test datasets (details in Sec IV-A). From the table, we can observe that the network trained with the balanced dataset obtains better test

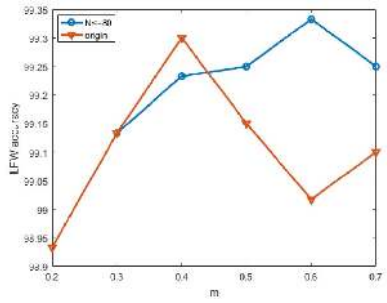
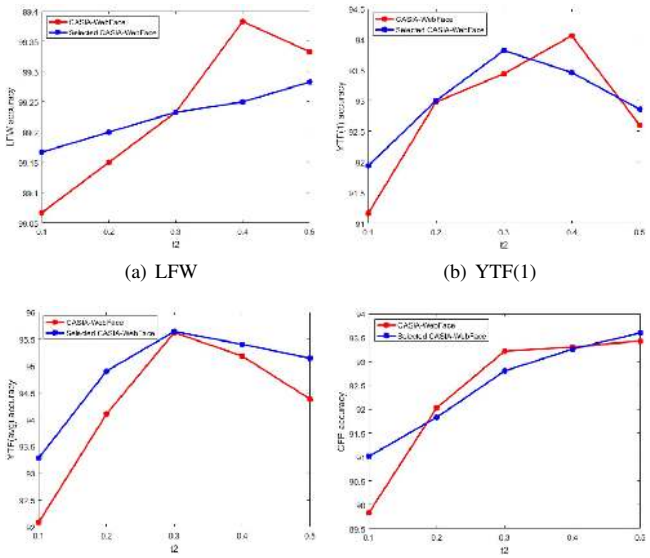


Fig. 5. The test accuracy on LFW with an increasing m. The orange curve is for the accuracy obtained from a network trained with the original CASIA-WebFace dataset, and the blue curve is for the balanced CASIA-WebFace dataset with each class has less than or equal to 80 images.



(a) LFW (b) YTF(1)

(c) YTF(avg) (d) CFP

Fig. 6. Four test accuracies (%) obtained from the networks trained with two different training datasets. The red curve is for the original CASIA-WebFace, and the blue curve is for the balanced CASIA-WebFace.

results on all the four test sets than the network trained with the original imbalanced dataset.

TABLE I
THE TEST RESULTS WITH DIFFERENT TRAINING DATASETS. YTF(1) IS THE TEST RESULTS FROM THE FIRST FRAME OF THE VIDEO. YTF(AVG) IS THE TEST RESULT BY USING THE AVERAGE FEATURES OF ALL THE VIDEO FRAMES.

Training dataset	m	LFW	YTF(1)	YTF(avg)	CFP
original CASIA-WebFace	0.4	99.30	92.42	94.06	92.49
balanced CASIA-WebFace	0.6	99.33	93.08	94.24	92.49

We illustrate the situation in Fig. 1(a) and intuitively speculate that this is because the AM loss in Eq.(5) only constrains a stable inter-class margin while not limiting the intra-class scopes for the majority classes. In Fig. 1(a), the majority class occupies more space than the minority class, and the large scope of the majority class hinders the expansion of inter-class margin m and leads to an imbalanced spatial distribution.

In contrast, our EqM loss constrains both the intra-class scope and the inter-class scope, which will reach a more uniform distribution for the majority and minority classes, as illustrated in Fig. 1(b). We also apply the EqM loss under the same experimental settings as for Table I, and show the results in Fig. 6. From Fig. 6, we can find that the trend of accuracy versus t_2 is almost the same for the imbalanced and balanced training datasets, which demonstrates that the EqM loss is not much affected by the negative influence of the "heavy head" on the inter-class margin. Moreover, from our experiments, we find that, when $t_1 = 0.9$ and $t_2 = 0.3$, the best results are obtained by using the balanced CASIA-WebFace dataset for training; when $t_1 = 0.8$ and $t_2 = 0.3$, the best results are attained by using the original CASIA-WebFace for training; and these best results are comparable, unlike the results of the AM loss in Table. I where using the balanced training data performs consistently better. From the above discussion, we can state that the EqM loss mitigates the harms on the inter-class margin caused by the "heavy head" of imbalanced data.

C. Optimization

In this subsection, we verify that the EqM loss is easy to optimize with the stochastic gradient descent algorithm. The EqM loss is different from the Softmax loss in the formulation of $\phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})$, so we rewrite their difference in the optimization process as follows:

$$\frac{\partial L_{margin}}{\partial w} = \frac{\partial L_{margin}}{\partial p}\frac{\partial p}{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}\frac{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}{\partial w},$$
$$\frac{\partial L_{margin}}{\partial x_i} = \frac{\partial L_{margin}}{\partial p}\frac{\partial p}{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}\frac{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}{\partial x_i};$$

for the EqM loss, we have

$$\frac{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}{\partial w_{y_i}} = \begin{cases} 0 & \cos\theta_{i,y_i} > t_1, \cos\theta_{i,j} > t_2 \\ 0 & \cos\theta_{i,y_i} > t_1, \cos\theta_{i,j} < t_2 \\ -2x_i & \cos\theta_{i,y_i} < t_1, \cos\theta_{i,j} > t_2 \\ -2x_i & \cos\theta_{i,y_i} < t_1, \cos\theta_{i,j} < t_2 \end{cases} \tag{13}$$

$$\frac{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}{\partial w_j} = \begin{cases} 2x_i & \cos\theta_{i,y_i} > t_1, \cos\theta_{i,j} > t_2 \\ 0 & \cos\theta_{i,y_i} > t_1, \cos\theta_{i,j} < t_2 \\ 2x_i & \cos\theta_{i,y_i} < t_1, \cos\theta_{i,j} > t_2 \\ 0 & \cos\theta_{i,y_i} < t_1, \cos\theta_{i,j} < t_2 \end{cases} \tag{14}$$

$$\frac{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}{\partial x_i} = \begin{cases} 2w_j & \cos\theta_{i,y_i} > t_1, \cos\theta_{i,j} > t_2 \\ 0 & \cos\theta_{i,y_i} > t_1, \cos\theta_{i,j} < t_2 \\ 2w_j - 2w_{y_i} & \cos\theta_{i,y_i} < t_1, \cos\theta_{i,j} > t_2 \\ -2w_{y_i} & \cos\theta_{i,y_i} < t_1, \cos\theta_{i,j} < t_2; \end{cases} \tag{15}$$

and for the Softmax loss, we have

$$\frac{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}{\partial w_{y_i}} = -x_i, \tag{16}$$

$$\frac{\partial \phi(\widetilde{w}_{y_i}, \widetilde{x}_i, \widetilde{w}_j)}{\partial w_j} = x_i, \tag{17}$$

$$\frac{\partial \phi(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}{\partial x_i} = w_j - w_{y_i}. \tag{18}$$

From these equations, we can observe that the EqM loss only needs to optimize the parameters in certain intervals where the gradient is nonzero, while the Softmax loss has to always update the parameters over the whole ranges of their potential values. The update scheme for the EqM loss is reasonable and more precisely-targeting. In Eq.(13) and Eq.(14), the weights w_{y_i} are updated when the intra-class compactness is lower than the threshold t_1 and the updating of w_j is only required when the inter-class similarity is higher than the threshold t_2. Additionally, as Eq.(15) shows, the updating of x_i is different in the four intervals. These formulae also indicate the effect of t_1 and t_2: t_1 decides the lower limit of intra-class similarity and t_2 represents the upper limit of inter-class similarity. That is, we decompose the objective of "reducing intra-class distance and enlarging inter-class variance" into two tasks and thus can control them in a more explicit and flexible way than the AM loss.

IV. Experiments

A. Datasets

Training Datasets. To avoid the bias due to a particular training dataset, in the experiments, we use two different training datasets. One is the CASIA-WebFace dataset [34], which contains about 0.49 million face images from 10,575 subjects. The other is the MS-Celeb-1M dataset [36], which has about 100k identities with 10 million face images. The MS-Celeb-1M dataset contains a large number of noisy face images; we clean the dataset and retain about 4.5 million face images. In terms of the amount of data, CASIA-WebFace is a small dataset and MS-Celeb-1M is a large dataset.

Test Datasets. To reach a relatively fair evaluation, in the experiments, we test the loss functions presented in this paper on three different types of datasets: LFW [35], CFP [37] and YTF [38]. LFW [35] contains about 13,000 images and has a list of 6000 pairs to verify. The face images in LFW [35] have a higher resolution and many studies [17]–[19], [21] have attained nearly perfect recognition accuracy. CFP [37] have two protocols: one is to test frontal images vs. frontal images, and the other is to test frontal images vs. profile images. In the experiments, to distinguish from LFW, we use the frontal vs. profile protocol, which contains 7,000 pairs with 3,500 same pairs and 3,500 non-same pairs for 500 different subjects. YTF [38] is applied to test video face images, and it has 5,000 video pairs. To obtain the recognition accuracy for YTF, We adopt two ways: testing the features after averaging all frames and testing the feature of the first face image of the corresponding video. The two ways measure the recognition performance from different aspects. We expect that it is generally hard to correctly identify a person by using only one video image. These three datasets focus on different aspects of face recognition. In Fig. 7, we show some examples of face images from the three datasets. It is clear that using the three datasets is able to evaluate the performance of different loss functions more comprehensively and fairly.

To further verify the effectiveness of our EqM loss, we additionally list the results on two large and challenging datasets: MegaFace [39] and IJB-B [40]. MegaFace [39] contains 1M distractors. In this paper, the probe set is Facescrub [41] and we show the 'Rank-1' identification accuracies with training datasets CASIA-WebFace [34] and MS-Celeb-1M [36]. The IARPA Janus Benchmark-B face challenge (IJB-B) [40] contains 67000 face images, 7000 face videos, and 10000 non-face images, and there is no overlap between IJB-B and other popular face datasets.

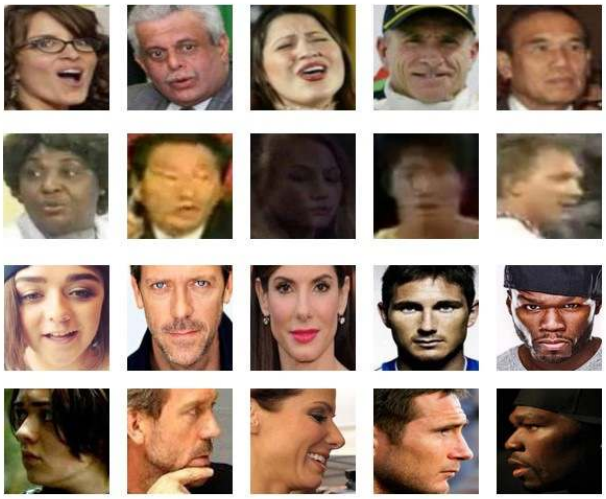


Fig. 7. Sample images from the three test datasets (LFW, YTF and CFP). The first row (LFW): high-resolution faces. The second row (YTF): blurring and low-resolution faces. The third and fourth rows (CFP): frontal vs. profile faces.

B. Experimental Settings

Data Enhancements. We detect and align faces using MTCNN [42]. All face images are pre-whitened and resized to 160 x 160. The face images are randomly flipped left and right and rotated to obtain a more stable recognition network.

Train. There are two different settings for the training on CASIA-WebFace [34] and MS-Celeb-1M [36], respectively. For the small dataset CASIA-WebFace [34], the batch size is 90, and the weight decay is 0.0005. The learning rate is initialized to 0.1 and divided by ten at 60k, 120k iterations, and the training process stops after 140k iterations. For the large dataset MS-Celeb-1M [36], the learning rate is initialized to 0.05 and divided by ten after 80k and 140k iterations, and we stop the training process after 180k iterations. It is worth noting that we use the network Inception-ResNet-V1 [14], [43], [44] as the base face recognition network.

Test. In the test stage, we compute the Euclidean distance between normalized features of 512 dimensions. All the test accuracy is obtained after ten-fold cross-validation. We extensively compare the performance of the loss functions by using four types of evaluations: for high-resolution faces in LFW [35], for the first images of faces with low resolution in YTF [38], for the average set-based face features in YTF [38], and for the faces with large pose variation in CFP [37].

We also list the Rank-1 identification accuracies in MegaFace [39], and in IJB-B [40], we draw the ROC curves

and use the AUC values to further discuss the effectiveness of EqM loss.

C. The Influence of t_1 and t_2

In this subsection, we discuss the influence of parameters t_1 and t_2 on the recognition accuracy. We use the case of $s = 64$ as an example. The recognition network is trained using CASIA-WebFace [34] and the values of t_1 and t_2 are changed in steps of 0.1. We plot the changes in accuracy with different values of t_1 and t_2 in Fig. 8, from which we can observe the following two patterns.

1) The influence of inter-class similarity upper limit t_2. From Fig. 8, we can observe that, when $t_2 = 0.1$, the accuracies of the four test results (LFW, YTF(1), YTF(avg) and CFP) are all among the lowest, regardless of the value of t_1. This demonstrates that an excessively large inter-class margin can be harmful to obtain discriminative features. We also observe that, when $t_2 > 0.3$, with the increase of t_2, the accuracies for different test datasets change in slightly different patterns, but nonetheless the accuracies in general decrease when t_2 is sufficiently large. All of these suggest that a appropriate inter-class margin can be attained by using a modest value of t_2.

2) The influence of intra-class similarity lower limit t_1. In Fig. 8, when we fix t_2, the largest t_1 ($t_1 = 1.0$) often does not result in the best test accuracy, and a larger t_1 often does not necessarily bring better performance. This matches the analysis of easy samples vs. hard samples in Section III-B. On the other hand, when t_1 sets a relaxed demand for the intra-class compactness (e.g. $t_1 = 0.7$), the accuracies are often among the lowest. All of these elaborate that an adequate intra-class compactness is preferred when we train the face recognition network, and this can be implemented by using a suitable value of t_1.

In summary, we can adjust the values, and thus influences, of t_1 and t_2 so as to achieve desirable test performance. Actually from the experiments, the range in which the two parameters are to adjust is reasonably small (often $t_1 > 0.5$ and $t_2 \in [0.1, 0.5]$). This indicates the flexibility and feasibility for the proposed EqM loss in practice.

D. Results on LFW, YTF, CFP

Now we compare the EqM loss with state-of-the-art loss functions [3], [17]–[19], [21], [24], [28], [45], [46] used in the face recognition community.

1) Trained with CASIA-WebFace: We show the test results in Table II by using the network trained with CASIA-WebFace [34]. Here, the normalized weights and features are applied in the Softmax loss. From Table II, we can observe that the EqM loss attains the highest accuracy on all the four test protocols. This verifies the effectiveness and general applicability of the EqM loss.

2) Trained with MS-Celeb-1M: There are many differences between the CASIA-WebFace dataset [34] and the MS-Celeb-1M dataset [36] used as the training sets in the experiments. The first difference is with the quantity of classes: CASIA-WebFace [34] contains 10,575 subjects while MS-Celeb-1M [36] has about 79k classes. The second difference is with the degree of class imbalance. According to our statistics, the maximum and minimum numbers of images in each class are 806 and 1 in CASIA-WebFace [34], while in MS-Celeb-1M [36] the numbers are 133 and 1. This demonstrates that the imbalanced data problem in CASIA-WebFace [34] is severer than in MS-Celeb-1M [36]. In the following, we will observe the different performances between the uses of these two datasets for training.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS TRAINED WITH CASIA-WEBFACE

	LFW	YTF(1)	YTF(avg)	CFP
Softmax loss	98.67	91.28	93.78	90.47
Center loss [3]	99.08	92.62	95.36	92.77
L_2-constrained loss [28]	98.82	92.36	95.44	92.60
Ring loss [45]	98.98	92.18	94.72	91.61
ArcFace [19]	99.20	93.30	95.38	93.37
AM loss [17], [18]	99.28	93.28	95.40	93.27
SphereFace [21]	99.17	92.58	95.46	92.57
SphereFace+ [46]	99.20	92.58	95.30	92.43
L-Softmax [24]	99.02	92.42	95.46	92.41
EqM loss ($t_1 = 0.8$, $t_2 = 0.3$)	**99.33**	**93.82**	**95.70**	**93.73**

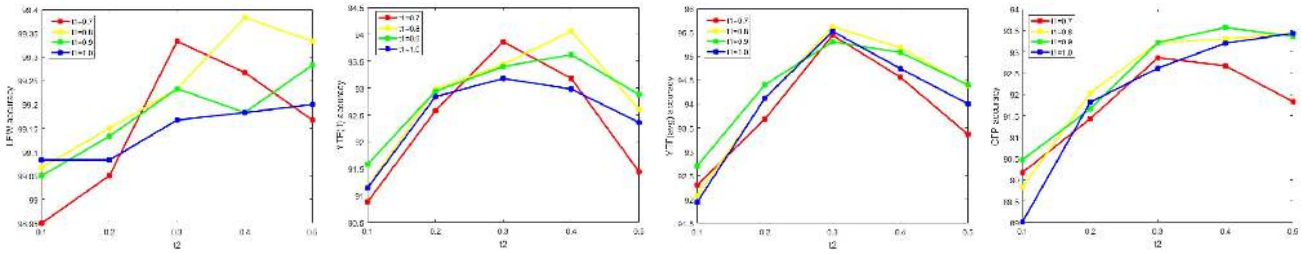
Comparing Table II and Table III, we can find that in Table III, L_2-constrained loss [28], Ring loss [45] and ArcFace [19] obtain equivalent or better performances compared with the AM loss [17], [18], while in Table II, their performances are much worse than the AM loss [17], [18]. We speculate that this is because the data of CASIA-WebFace [34] is more unbalanced, which is harmful to obtain a better performance. In Table III, the EqM loss achieves the best performances on YTF(1) and obtains the second best performance on LFW and CFP, which is comparable with the state-of-the-art loss functions. This indicates the consistency of the EqM loss in the cases of training with data of various degrees of imbalance.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS TRAINED ON MS-CELEB-1M

	LFW	YTF(1)	YTF(avg)	CFP
Softmax loss	99.28	94.18	95.66	91.33
Center loss [3]	99.32	94.00	96.04	92.36
L_2-constrained loss [28]	99.47	94.68	**96.74**	92.61
Ring loss [45]	99.50	94.58	96.52	92.00
ArcFace [19]	**99.58**	94.64	96.60	92.20
AM loss [17], [18]	99.50	**94.76**	96.52	92.01
SphereFace [21]	99.52	94.70	96.66	**92.44**
SphereFace+ [46]	99.52	94.32	**96.74**	92.33
L-Softmax [24]	99.35	93.48	95.96	92.33
EqM loss ($t_1 = 1.0$, $t_2 = 0.3$)	99.55	**94.76**	96.38	92.33

E. Results on MegaFace and IJB-B

In Table IV, we show the results on MegaFace [39] obtained by the state-of-the-art methods. With the training dataset CASIA-WebFace [34], our method achieves the best result, which indicates the superiority of EqM loss. With the training dataset MS-Celeb-1M [36], our method obtains comparable result with the best method.



(a) LFW (b) YTF(1) (c) YTF(avg) (d) CFP

Fig. 8. The influence of t_1 and t_2 on the four types of test results.

TABLE IV
RESULTS ON THE MEGAFACE WITH TRAINING DATASET CASIA-WEBFACE

Methods	CASIA-WebFace	MS-Celeb-1M
Softmax loss	56.4859	76.2512
Center loss [3]	64.4771	62.3003
L_2-constrained loss [28]	64.1142	75.0621
L-Softmax [24]	64.6936	57.5016
Ring loss [45]	61.3447	77.7578
ArcFace [19]	73.8765	83.7656
AM loss [17], [18]	77.0413	**86.2122**
SphereFace [21]	64.1899	79.5068
SphereFace+ [46]	64.9571	79.8825
EqM loss	**79.9838**	85.9151

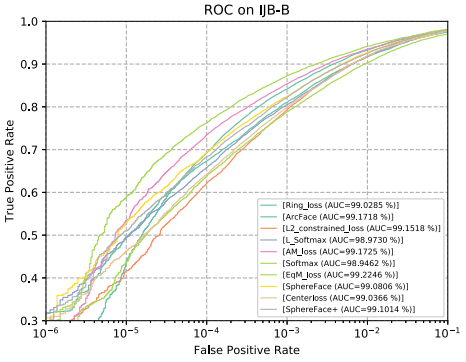


Fig. 9. Results on IJB-B with training dataset CASIA-WebFace.

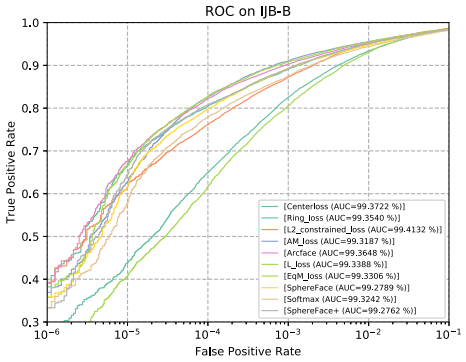


Fig. 10. Results on IJB-B with training dataset MS-Celeb-1M.

Fig. 9 and Fig. 10 show the ROC curves on IJB-B [40] with training datasets CASIA-WebFace [34] and MS-Celeb-1M [36]. To better show the values in the figures, we list the results in Table V and Table VI. In Fig. 9, the EqM loss gets the best AUC value and almost the best ROC curve. With the training dataset MS-Celeb-1M [36], the EqM loss gets a comparable ROC curve.

V. DISCUSSION

In this section, we discuss the influence of scale s in Eq.(1) on the performances of the EqM loss and the AM loss, and show that the EqM loss is stable with the change of s. Many previous studies, such as [17]–[19], [29], usually set the value of s according to the experimental results. However, the range

TABLE V
ACCURACIES ON IJB-B WITH DIFFERENT FPR VALUES BY TRAINING DATASET CASIA-WEBFACE.

Methods	1e-06	1e-05	0.0001	0.001	0.01	0.1
Softmax loss	23.01	43.59	63.80	78.79	90.25	97.01
Center loss [3]	28.76	46.35	64.12	79.91	91.93	97.72
L_2-constrained loss [28]	24.49	41.70	62.12	79.31	91.79	97.96
Ring loss [45]	31.85	49.66	67.36	81.05	91.71	97.49
ArcFace [19]	11.11	43.83	69.35	84.15	93.30	98.00
AM loss [17], [18]	24.92	53.47	73.44	85.41	93.50	97.98
SphereFace [21]	30.63	52.84	69.35	82.46	92.71	97.90
SphereFace+ [46]	29.77	51.02	68.33	82.21	92.54	97.99
L-Softmax [24]	31.74	49.46	65.86	80.47	91.86	97.62
EqM loss	29.39	59.04	76.25	87.22	94.15	98.15

TABLE VI
ACCURACIES ON IJB-B WITH DIFFERENT FPR VALUES BY TRAINING DATASET MS-CELEB-1M.

Methods	1e-06	1e-05	0.0001	0.001	0.01	0.1
Softmax loss	33.30	57.84	77.91	87.55	94.40	98.21
Center loss [3]	22.22	43.85	64.89	82.53	93.41	98.67
L_2-constrained loss [28]	39.04	61.70	76.11	87.28	95.14	98.57
Ring loss [45]	38.21	67.10	80.45	89.08	95.11	98.42
ArcFace [19]	39.29	67.73	82.21	90.30	95.24	98.47
AM loss [17], [18]	36.94	62.79	82.66	91.04	95.54	98.34
SphereFace [21]	35.85	62.21	79.81	89.29	94.76	98.30
SphereFace+ [46]	33.32	64.48	80.78	89.26	95.09	98.36
L-Softmax [24]	19.45	40.91	61.66	80.56	93.09	98.50
EqM loss	35.77	66.57	82.73	90.83	95.57	98.36

of s is too large to be exhaustively searched. Although [18] proposed a lower limit on the value of s, [18] did not provide a guide to users about how to choose a suitable s.

Firstly, we investigate the influence of s on the AM loss in Eq.(5) together. For a fixed value of s (30, 64 or 100), we change the value of parameter m in steps of 0.1 to obtain the best result considering the four test accuracies (LFW, YTF(1), YTF(avg) and CFP). All the experiments are implemented with the training dataset CASIA-WebFace [34]. The recognition results are shown in Table VII, from which we can observe that all the four accuracies are in a decreasing pattern (although it is slight oscillating for LFW). That is, a lower s is preferred. However, it is also known that the network with the AM loss cannot converge if the value of s is too small.

Secondly, we explore the influence of s on the proposed EqM loss in Eq.(7). The results are shown in Table VIII, where the accuracies are obtained with varying the values of t_1 and t_2 in steps of 0.1. Considering the factor '2' in our EqM loss (see Eqs.(8), (10), (11)), in Table VIII we also list the results for $s = 15$, as it is equivalent to $s = 30$ in Table VII, to be fair with both tables. We can clearly observe that, for each of the four test protocols, the results from using different values of s are very close to each other. That is, the proposed EqM loss is more stable than the AM loss, and can obtain comparable accuracies in a large range of s. As it is time-consuming to find a suitable s through experiments, the insensitivity of the EqM loss to s can help to largely reduce the time complexity of network training in practice. Additionally, the best accuracies in Table VIII are higher than the best accuracies in Table VII. This also verifies that the EqM loss is more effective than the AM loss.

TABLE VII
THE INFLUENCE OF s ON THE AM LOSS FUNCTION IN EQ.(5)

s	m	LFW	YTF(1)	YTF(avg)	CFP
30	0.3	99.28	93.28	95.40	93.27
64	0.4	99.30	92.42	94.06	92.49
100	0.4	99.22	91.28	93.44	90.49

TABLE VIII
THE INFLUENCE OF s ON THE EQM LOSS IN EQ.(7)

s	t_1	t_2	LFW	YTF(1)	YTF(avg)	CFP
15	0.8	0.3	99.33	93.46	95.58	93.26
30	0.8	0.3	99.33	93.82	95.70	93.73
64	0.8	0.3	99.23	93.44	95.62	93.21
100	0.8	0.4	99.28	93.38	95.42	93.69

Now we start to explain the above patterns by examining Eq.(1). To make the explanation more intuitive, we only consider the average $\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})$ of the C classes, and rewrite Eq.(1) as

$$L_{margin} = -\log \frac{1}{1 + (C-1)e^{s\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i})}}. \quad (19)$$

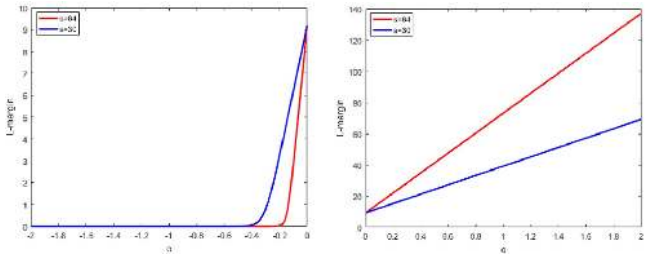
Moreover, for simple illustration, we set $C = 10001$ and draw the function of Eq.(19) into two panels in Fig. 11: Fig. 11(a) is for $\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) \leq 0$ and Fig. 11(b) is for $\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) > 0$; in each panel, there are two curves: one for $s = 30$ and the other for $s = 64$. In Fig. 11(a), each curve can be further divided into two parts: one part is the smooth area whose slope is almost 0, and the other part is the steep area whose slope is large and almost constant. We use P to denote the demarcation point of $\bar{\phi}$ between the two areas. A smaller s produces a smaller P, which is more conducive to extract discriminative features. Fig. 11(b) shows constant slopes with steep areas only.

The AM loss in Eq.(5) shifts the original curve of Eq.(19) to the left by adding a positive margin m, as shown in Fig. 12 for the two optimal pairs $(s = 30, m = 0.3)$ and $(s = 64, m = 0.4)$ as listed in Table VII. From Fig. 12, we can observe the followings. Firstly, using the AM loss enlarges the scope of the steep area to obtain more discriminative features, as verified by the results in Table VII. Secondly, The demarcation points P of the two curves become almost same (to be precise, we can adjust m to make the same P), but the test accuracies for $s = 64$ in Table VII are much lower than $s = 30$. In Fig 12, the only difference between the two curves is that the red curve (for $s = 64$) has a larger slope in the steep area than the blue curve (for $s = 30$). If we treat the samples in the steep area as hard samples (due to their large losses), then Fig 12 implies that the larger $s = 64$ punishes the hard samples more severely than the smaller $s = 30$. On the one hand, if Eq.(5) penalizes much more the hard samples when the value of s is larger, the face recognition network will bias toward optimizing the hard samples, over-enlarge the space of each class and thus fail to satisfy the inter-class margin m required by the AM loss; this will leads to a drop in the recognition accuracy. On the other hand, if we adjust the inter-class margin m to relieve the problem above-mentioned, it may also result in the decline of the recognition accuracy. This dilemma can only be addressed if we can secure a suitable value of s.

For the EqM loss in Eq.(7), as we discussed in Section III-A about the four cases, the minimization of the EqM loss will lead to $\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) = 0$. That is, the EqM loss makes $\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) \geq 0$. In other words, the optimization of the EqM loss has moved the whole curve to the steep area and eliminated the smooth area of Eq. 1. Specifically, t_1 adjusts the intra-class compactness and make a suitable boundary between hard samples and easy samples; t_2 controls the inter-class margin to avoid the overlarge margin; t_1 and t_2 jointly to adjust the point P, making the EqM loss more flexible than the AM loss when the value of s changes. Additionally, from Eq.(11), we can see that the inter-class distance is irrelevant to the hard samples, which ensures a proper inter-class distance.

VI. CONCLUSIONS

In this paper, we propose a new loss function called the EqM loss, in order to ensure an appropriate balanced point between reducing the intra-class distance and enlarging the inter-class margin. The two hyper-parameters of the EqM loss enable a flexible control to balance the penalization of hard samples and the extraction of discriminative features. In the experiments, we also find that the EqM loss can ease the harms caused by imbalanced data, is more stable with the change of



(a) $\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) \leq 0$ (b) $\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) > 0$.

Fig. 11. The curve of Eq.(19). The red curve is for $s = 64$ and the blue curve is for $s = 30$

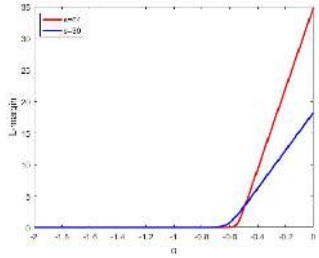


Fig. 12. The curve of Eq.(5) when $\bar{\phi}(\widetilde{w}_j, \widetilde{x}_i, \widetilde{w}_{y_i}) \leq 0$

s in Eq.(1), and performs better than other state-of-the-art loss functions for face recognition.

ACKNOWLEDGMENT

This work was partly supported by the National Natural Science Foundation of China (No.61471216 and No.61771276), the National Key Research and Development Program of China (No.2016YFB0101001), and the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (No.JCYJ20170307153940960 and No.JCYJ20170817161845824)

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[2] ——, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[3] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," *arXiv preprint arXiv:1412.6856*, 2014.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[6] C.-K. Hsieh, S.-H. Lai, and Y.-C. Chen, "Expression-invariant face recognition with constrained optical flow warping," *IEEE Transactions on Multimedia*, vol. 11, no. 4, pp. 600–610, 2009.

[7] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[9] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1222–1228, 2004.

[10] Z. Li, W. Liu, D. Lin, and X. Tang, "Nonparametric subspace analysis for face recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 961–966, 2005.

[11] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 755–761, 2009.

[12] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[13] R. Gao, F. Yang, W. Yang, and Q. Liao, "Margin loss: Making faces more separable," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 308–312, 2018.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[15] B. Leng, K. Yu, and Q. Jingyan, "Data augmentation for unbalanced face recognition training sets," *Neurocomputing*, vol. 235, pp. 10–14, 2017.

[16] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, 2017.

[17] F. Wang, W. Liu, H. Liu, and J. Cheng, "Additive margin Softmax for face verification," *arXiv preprint arXiv:1801.05599*, 2018.

[18] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," *arXiv preprint arXiv:1801.09414*, 2018.

[19] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *arXiv preprint arXiv:1801.07698*, 2018.

[20] F. Yang, W. Yang, R. Gao, and Q. Liao, "Discriminative multidimensional scaling for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 388–392, 2018.

[21] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 212–220.

[22] M. Gunther, S. Cruz, E. M. Rudd, and T. E. Boult, "Toward open-set face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 71–80.

[23] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Faces in-the-wild Workshop/Challenge*, vol. 4, no. 6, 2017.

[24] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *ICML*, 2016, pp. 507–516.

[25] J. Sun, W. Yang, R. Gao, J.-H. Xue, and Q. Liao, "Inter-class angular margin loss for face recognition," *Signal Processing: Image Communication*, vol. 80, p. 115636, 2020.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[27] X. Wang, S. Wang, S. Zhang, T. Fu, H. Shi, and T. Mei, "Support vector guided softmax loss for face recognition," *arXiv preprint arXiv:1812.11317*, 2018.

[28] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[29] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L_2 hypersphere embedding for face verification," *arXiv preprint arXiv:1704.06369*, 2017.

[30] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *IEEE International Conference on Computer Vision*, 2017, pp. 5409–5418.

[31] C. Huang, Y. Li, C. Change Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5375–5384.

[32] Q. Dong, S. Gong, and X. Zhu, "Class rectification hard mining for imbalanced deep learning," in *IEEE International Conference on Computer Vision*, 2017, pp. 1851–1860.

[33] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao, and L. Song, "Deep hyperspherical learning," in *Advances in neural information processing systems*, 2017, pp. 3950–3960.

[34] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[35] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.
[36] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 87–102.
[37] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.
[38] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 529–534.
[39] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *CVPR*, 2016, pp. 4873–4882.
[40] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "Iarpa janus benchmark-b face dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 90–98.
[41] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 343–347.
[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 1–9.
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
[45] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5089–5097.
[46] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song, "Learning towards minimum hyperspherical energy," in *Advances in Neural Information Processing Systems*, 2018, pp. 6222–6233.