IEEE Conference on Computer Vision and Pattern Recognition 2019

RegularFace: Deep Face Recognition via Exclusive Regularization

Kai Zhao Jingyi Xu Ming-Ming Cheng *
TKLNDST, CS, Nankai University
kz@kaizhao.net cmm@nankai.edu.cn

Abstract

We consider the face recognition task where facial images of the same identity (person) is expected to be closer in the representation space, while different identities be far apart. Several recent studies encourage the intra-class compactness by developing loss functions that penalize the variance of representations of the same identity. In this paper, we propose the 'exclusive regularization' that focuses on the other aspect of discriminability – the inter-class separability, which is neglected in many recent approaches. The proposed method, named RegularFace, explicitly distances identities by penalizing the angle between an identity and its nearest neighbor, resulting in discriminative face representations. Our method has intuitive geometric interpretation and presents unique benefits that are absent in previous works. Quantitative comparisons against prior methods on several open benchmarks demonstrate the superiority of our method. In addition, our method is easy to implement and requires only a few lines of python code on modern deep learning frameworks.

1. Introduction

Face recognition is one of the most widely studied topics in computer vision, and recently significant improvement has been made with the convolutional neural networks (CNNs) becoming the workhorse. In general, there are two sub-tasks in face recognition: 1) face identification that attributes a given facial image to a known identity, and 2) face verification that determines whether a pair of facial images belongs to the same identity. There are also two testing protocols for face recognition: the open-set protocol and closed-set protocol. In the open-set circumstance, testing identities may not exist in the training set; while in close-set setting, training images and testing images are drawn from the same identities. The open-set face recognition is more challenging and closer to real-world applications because it's infeasible to collect all identity faces for training.

*M.M. Cheng is the corresponding author.

It is widely accepted that learning discriminative feature representation is the key to accurate open-set face recognition [3, 21, 28, 14]. The Intra-class compactness and inter-class separability are two important factors to feature discriminability: representations belonging to the same identity are expected to be closer in the representation space; while representations of different identities are expected to be scattered away. Many recent works insist on designing novel loss functions to improve the intra-class compactness of deep features. *Center loss* [28] improves the intra-class compactness by imposing extra loss term that penalizes the Euclidean distance between samples and their representation centers. Then in *SphereFace*, Liu *et al*. proposed the A-Softmax loss [14] that imposes an angular margin to concentrate the samples in a sphere manifold. Here 'softmax loss' represents softmax normalization followed by cross entropy loss. Similar to *SphereFace*, *CosFace* [27] and *ArcFace* [4] also impose angular margins to the decision boundaries of original softmax loss, leading to further performance improvement. These methods focus on the intra-class compactness by clamping representations of the same identity, either in the Euclidean space (*center loss*) or in the sphere space (*SphereFace*, *CosFace*, *ArcFace*).

In this paper we consider the other side of discriminability: inter-class separability. Apart from intra-class compactness that shortens the distance between representations of the same identity, the inter-class separability, on the other side, aims at distancing samples of different identity classes. Specifically, we impose a regularization term, named *exclusive regularization*, to parameters of the classification layer. The regularization term explicitly enlarges the angle between parametric vectors of different identity classes, leading to 'exclusive' classification vectors. Consequently, these regularized classification vectors will, in turn, scatter the samples of different identities in the representation space. Our contributions are concluded as follows:

- First, we propose to quantitatively evaluate the inter-class separability with angular distance between identity centers.
- Second, we present a novel *exclusive regularization* term which explicitly enlarges the angular distance be-

tween different identities. To the best of our knowledge, we are the first to enhance feature discrimination by promoting inter-class separability for face recognition.

- Third, our method is orthogonal with, and therefore, can be seamlessly plugged into, existing approaches to improving the performance with less effort.
- And last, we test the proposed method on LFW [3], YouTube face (YTF) [29] and MegaFace challege[10], achieving promising performance.

2. Related Work

There has been a significant improvement in face recognition due to the use of CNNs [5, 26, 20, 24, 15]. Based on the loss function in use, there are two major types of approaches: softmax-free methods and softmax loss based methods.

Softmax-free Methods. In softmax-free methods, face pairs are fed to the model to train feature embedding with pairwise annotations (*e.g*. whether a pair of facial images come from the same identity). Since the identity label is invisible during training, the model cannot utilize classification losses, *e.g*. the *softmax-loss*, as supervision. Chopra *et al*. proposed Siamese networks [3] with the contrastive loss to learn contrastive representations. In Siamese networks, two facial images are successively fed into two identical networks to obtain their respective embeddings, and the *contrastive loss* penalizes the distance between two embeddings when the input images are paired. Hu *et al*. [7] designed a discriminative deep metric with a margin between positive and negative face pairs. Florian *et al*. proposed the *Triplet loss*[5] which accepts three images as input at once, two of which are paired (the anchor and the positive) and the other is an outlier (the negative). The Triplet loss minimizes the embedding distance between paired images, and meanwhile, maximizes the distance between the negative sample and others. Note that both *contrastive loss* and *triplet loss* require a carefully designed pair selection and procedure [5, 17].

Softmax-based Methods. Softmax-based methods usually accept identity labels as supervision. Therefore, classification losses, typically the *softmax-loss* (or its invariants), can be used as supervision. [24] adds extra loss terms by adding fully connected layers and loss functions to each convolutional layer, consequently enhancing the supervision. Very recently, Wen *et al.* [28] proposed the *center-loss* that penalizes the Euclidean distance between embeddings and their corresponding centers. The model is then jointly supervised by *center-loss* and the softmax-loss, to emphasize the intra-class compactness in the embedding manifold. *SphereFace* [14] introduced another variant of softmax-loss, the *angular margin softmax loss* (A-softmax loss) that brings an angular margin to the decision boundary of original softmax-loss. Specifically, *SphereFace* uses a multiplier to impose multiplicative margin to the original decision boundaries. Another study *ArcFace* [4] used an additive angular margin, leading to further performance improvement. Similar idea is also presented in *CosFace* [27] which narrows the decision margin in the cosine manifold.

There are also many researchers trying to combine the philosophy of the aforementioned two kinds of methods. For example, [23] proposed to jointly supervise the deep model with identification signal (softmax loss) and verification signal (triplet loss).

3. Observation and Motivation

Our method is mainly inspired by the recent work of *center-loss* [28] and *SphereFace* [14]. We start by analysing [14, 28], and then illustrate how we are motivated to propose the *exclusive regularization*.

3.1. Softmax Loss and Variants

Center-Loss. It penalizes the Euclidean distance between feature embeddings and their corresponding centers, with the purpose of imposing intra-class compactness in the representation space. The center penalty is defined as

$$\mathcal{L}_{center} = \frac{1}{2}\sum_{i=1}^{N} \|x_i - c_{y_i}\|_2^2 , \tag{1}$$

where $x_i \in \mathbb{R}^K$ is the feature embedding of sample i, c_{y_i} is the embedding center of samples whose identity label is y_i ($y_i \in \{1, ..., C\}$). The embedding centers of each identity are iteratively updated during training.

Softmax Loss & Angular Softmax Loss. We start from the original softmax loss and then introduce its angular invariants. Given embedding vector x_i, the posterior of x_i belonging to identity c is:

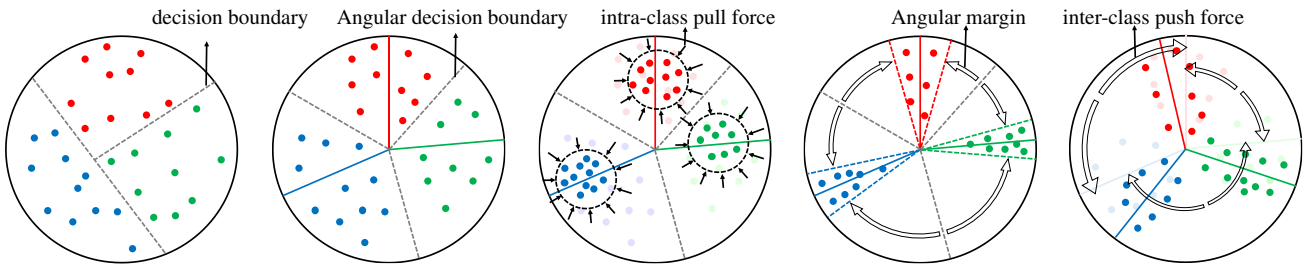
$$p_c(x_i) = \frac{e^{W_c^T x_i + b_c}}{\sum_{j=1}^{C} e^{W_j^T x_i + b_j}}, \tag{2}$$

where W is a $K \times C$ matrix that maps x to posterior probabilities, and b is the bias term. K and C are dimension of feature embeddings and number of identities, respectively. Obviously we have $\sum_{c=1}^{C} p_c = 1$. Given the identity label y_i, the softmax loss is

$$l(x_i, y_i) = -\sum_{c=1}^{C} 1(y_i = c) \cdot \log p_c(x_i). \tag{3}$$

$1(\cdot)$ is an indicator function that values to 1 when the condition is true, and otherwise values to 0. Then we zero the bias and normalize each column of W to deduce the angular softmax loss, where the posterior is given by:

$$p_c(x_i) = \frac{e^{\|x_i\| \cos(\phi_{i,c})}}{\sum_{j=1}^{C} e^{\|x_i\| \cos(\phi_{i,j})}}. \tag{4}$$



(a) Softmax loss (b) Angular softmax loss (c) Center loss [28] (d) SphereFace [14] (e) RegularFace

Figure 1. Illustration of face embeddings trained under various loss functions, points in color indicate different identities. (a) Softmax loss learns separable decision boundaries. (b) Angular softmax loss learns angularly separable decision boundaries. (c) *Center loss* [28] 'pulls' embeddings of the same identity towards their center, in order to obtain compact and discriminative representations. (d) *SphereFace* [14] (A-Softmax loss) proposes the 'angular margin' to clamp representations within a narrow angle. (e) Our proposed *RegularFace* introduces 'inter-class push force' that explicitly 'pushes' representations of different identities far way.

In Eq.(4), $\phi_{i,j}$ is the angle between feature embedding x_i and weight vector W_j. The decision boundaries of angular softmax-loss is shown in Fig. 1 (b). Obviously, minimizing the softmax-loss is equivalent to minimizing ϕ_{i,y_i}. Therefore, weight vector W_j can be regarded as the cluster center of all x_i with $y_i = j$.

A-Softmax Loss. *SphereFace* [14] introduces an angular margin to the decision margin of *angular softmax loss* in Eq.(4) so as to compress embeddings of the same identity class in the hypersphere space (Fig. 1 (d)). Novelly, the posterior p_c is defined as:

$$p_c(x_i) = \frac{e^{\|x_i\|\cos(m\cdot\phi_{i,y_i})}}{e^{\|x_i\|\cos(m\cdot\phi_{i,y_i})} + \sum_{j\neq y_i} e^{\|x_i\|\cos(\phi_{i,j})}}, \quad (5)$$

where $m \in \mathbb{Z}_+ = \{1, 2, ...\}$ is a factor used to control the margin. When $m = 1$, Eq.(5) reduces to Eq.(4). As depicted in Fig. 1 (d), A-Softmax loss can learn angularly compact intra-class representations in the sphere manifold.

3.2. Inter-class Separability

Inter-class separability and intra-class compactness are two key factors to discriminability. Many classic methods [5, 23] simultaneously take these two factors into consideration. However, current softmax-based face recognition approaches such as center loss [28] and SphereFace [14] care mainly about the intra-class compactness, either in Euclidean manifold (center loss) or in sphere manifold (SphereFace). The inter-class separability has not been paid for special attention by recent softmax-based face recognition methods.

Many recent works [14, 28, 15, 27] first perform experiments on a tiny dataset, say MNIST [13], to geometrically demonstrate the discriminability of learned representations. Usually, these demonstrative experiments restrict representations in a low dimension space (2D or 3D) to ease the visualization. In the case that there are relatively redundant clusters (identities) than representation dimensions, the clusters tend to stretch so as to decrease the classification error. In the demonstrative experiment of [28], the author trained a model on the MNIST dataset with 2D representations. The cluster centers of representations are nearly uniformly distributed and hold the maximal inter-cluster distances in a 2D plane (shown in Fig.3 of [28]).

Partially misguided by these demonstrative experiments, we may mistakenly assume that the cluster centers (W_j) are, at least to some extent, evenly distributed in the representation space so that generally the cluster centers present large inter-class separability. One notable fact is that we commonly have more redundant dimensions relative to the number of identities, to guarantee better performance. For example, recent works [27, 14, 28] usually use 512 dimensional representations and train the model with 10K identities. In this case, the cluster centers may not be so well distributed.

We quantitatively evaluate the inter-class separability based on the classification matrix $W \in \mathbb{R}^{K\times C}$ that maps the representations to identity confidences. W_i is the i-th column of W which represents the weight vector for the i-th identity class. We measure the inter-class separability of cluster centers by

$$\begin{aligned} Sep_i &= \max_{j\neq i} \cos(\varphi_{i,j}) \\ &= \max_{j\neq i} \frac{W_i \cdot W_j}{\|W_i\| \cdot \|W_j\|}, \end{aligned} \quad (6)$$

where $\varphi_{i,j}$ is the angle between W_i and W_j. Ideally the cluster centers are expected to be uniformly distributed and be as far away (small cos value) from others as possible. In other words $mean(Sep)$ and $std(Sep)$ are expected to be as small as possible. Quantitative comparisons of models trained with different loss functions are listed in Tab.1. All the models are trained on the CASIA-WebFace dataset with ResNet20 as the backbone architecture.

Methods	mean(Sep)	std(Sep)
Softmax Loss	0.286	0.0409
Center Loss[28]	0.170	0.134
SphereFace[14]	0.170	0.013
Random	0.16992	0.027

Table 1. Inter-class separability of different models. 'Random' means the model parameters are draw from a uniform distribution.

3.3. Motivation of Exclusive Regularization

The statistics in Tab. 1 reveals that the cluster centers in existing methods are not so well distributed. Therefore, we may potentially improve feature discrimination by enhancing the inter-class separability.

Inspired by the idea that 'cluster centers of different identities should stand far apart', we propose the 'exclusive regularization' to explicitly force the cluster centers W_j to move away from each other during training. Consequently, the regularization term will result in angularly separated weight vectors and separated representations as well.

As pointed out in Eq.(4), minimizing the angular softmax loss is equivalent to minimizing the angle between representations and corresponding cluster center W_j. Therefore, angularly separated weight vectors W_j will in turn pull representations of different identities apart from each other, making the representations more discriminative by enlarging their 'inter-class separability', as illustrated in Fig. 1 (e).

4. Exclusive Regularization

4.1. Formulation of Exclusive Regularization

Let $G_\theta(\cdot)$ be the all layers of the model except FC2, parameterized by θ. Matrix $W \in \mathbb{R}^{K\times C}$ denotes the parameter of FC2 (FC2 in Fig. 2) which maps representations to identity predictions. Given input image I_i, we obtain its feature representation x_i through

$$x_i = G_\theta(I_i). \tag{7}$$

According to Eq.(4), the angular softmax loss is:

$$\mathcal{L}_s(\theta, W) = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{\|x_i\|_2 \cos(\phi_{i,y_i})}}{\sum_j e^{\|x_i\|_2 \cos(\phi_{i,j})}}, \tag{8}$$

where y_i is the identity label, and ϕ_{i,y_i} is the angle between feature embedding x_i and classification vector W_{y_i}.

As aforementioned, the parameter W_j for identity class j can be regarded as the *cluster center* of all x_i with $y_i = j$. In the purpose of enlarging the angular distance between samples of different identities, we indirectly introduce the *exclusive regularization* that enlarges the angle between cluster center of identities. Following Eq.(6), the regularization term is defined as:

$$\mathcal{L}_r(W) = \frac{1}{C}\sum_i \max_{j\neq i} \frac{W_i \cdot W_j}{\|W_i\| \cdot \|W_j\|}. \tag{9}$$

We jointly supervise the model with angular softmax loss and *exclusive regularization*, the overall loss function is

$$\mathcal{L}(\theta, W) = \mathcal{L}_s(\theta, W) + \lambda \mathcal{L}_r(W), \tag{10}$$

where λ is a balance factor between the two terms. Note that the angular softmax loss $\mathcal{L}_s$ can be replaced by other advanced loss functions, *e.g. A-Softmax loss*[14] or *center-loss*[28], to perform joint optimization.

As discussed in Sec. 3.2, the angular softmax loss will pull sample representations x_i towards their cluster center W_{y_i}. Meanwhile, the *exclusive regularization* term will push different cluster centers W_j and W_i $(i \neq j)$ apart. Consequently, with the joint supervision of the **inter-class push force** and the **intra-class pull force**, the model will learn a discriminative representation that puts special emphasis on the inter-class separability.

4.2. Optimize with Projected Gradient Descent

Optimizing loss function in Eq.(10) can be formulated as:

$$(\theta^*, W^*) = \underset{(\theta, W)}{\operatorname{argmin}} \mathcal{L}(\theta, W). \tag{11}$$

For θ we update it with the standard stochastic gradient descent (SGD) method:

$$\theta^{t+1} = \theta^t - \alpha \frac{\partial \mathcal{L}_s(\theta^t, W)}{\partial \theta^t}, \tag{12}$$

in which α is the learning rate.

As described in Sec. 3.1, W is normalized in a hypersphere surface so as to derive the angle-based loss. The standard SGD iteration will drive W out of the hypersphere surface. Therefore, we use the projected gradient descent[1] to update W:

$$\begin{cases} \hat{W}^{(t+1)} = W^t - \alpha \frac{\partial \mathcal{L}}{\partial W^t} \\ W^{(t+1)} = Normalize(\hat{W}^{(t+1)}). \end{cases} \tag{13}$$

The second step of Eq.(13) is called the 'project step' that projects the updated parameters back to the nearest boundary of constraints. Since W is constrained to the sphere surface, we simply perform L2 normalization on columns of W.

4.3. The Gradient of Exclusive Regularization

Let W_j be the j-th column of W, and it is constrained by $|W_j|_2^2 = 1$. The gradient of $\mathcal{L}_r$ w.r.t W_j is:

$$\frac{\partial \mathcal{L}_r(W)}{\partial W_j} = W_{j'} + \sum_{W_i \in \mathbb{C}} W_i \tag{14}$$

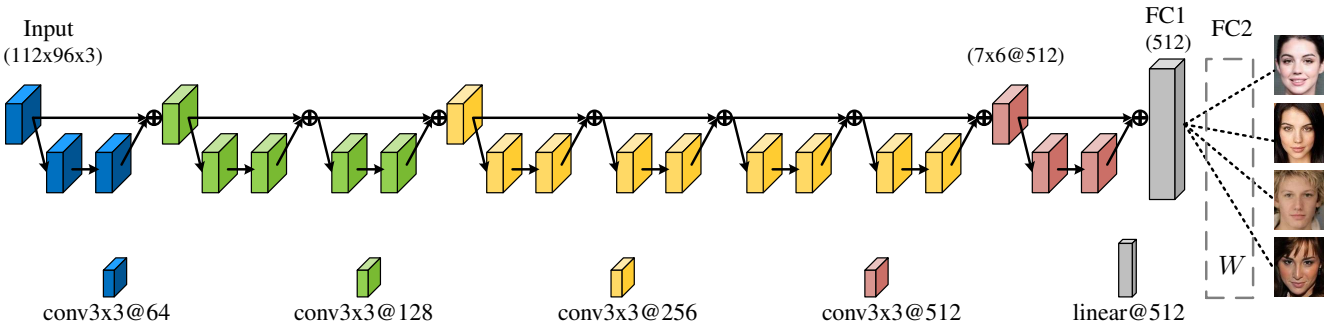


Figure 2. The ResNet20 architecture. 'conv3x3@$\mathcal{X}$' represents a 3×3 convolutional layer that outputs $\mathcal{X}$ feature maps, and $\bigoplus$ represents element-wise sum. W is a matrix that maps the facial representation to probabilities of input image belonging to identities.

where $W_{j'}$ is the nearest neighbor of W_j:

$$j' = \underset{i \in \{1,\dots,C\}, i \neq j}{\operatorname{argmax}} W_j \cdot W_i.$$

$\mathbb{C}$ is the collection of columns whose nearest neighbor is W_j:

$$\forall\, W_i \in \mathbb{C}, \underset{k \in \{1,\dots,C\}, k \neq i}{\operatorname{argmax}} W_i \cdot W_k = j.$$

5. Experiments

5.1. Implementation Details

Here we give several details that are critical to reproducing the method and the performance.

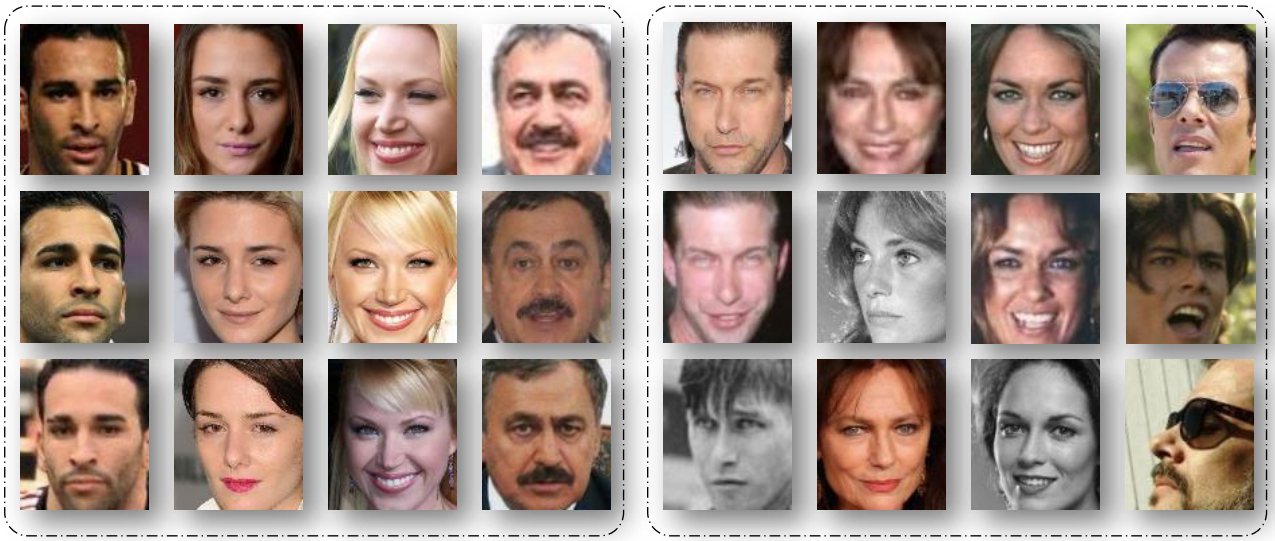
Network Settings. A grown body of research on CNN architecture designing [25, 6, 11] clearly reveals that deeper networks consistently present better performance. In *SphereFace* the author has tested 5 residual networks with different depth: 4, 10, 20, 36 and 64. The performance continuously increases when the networks get deeper. However, deeper networks require more GPU memory and are computationally costly, thus need more training time. For the compromise between performance and time-efficiency, we implement our proposed method based on the ResNet20 architecture, similar architecture is also used in[28, 4]. Our network, as shown in Fig. 2, accepts 112×96 RGB image as input. After 4 residual blocks which totally contain 20 convolutional layers, the shape of the output feature map is $7 \times 6 \times 512$. Then the first fully connected layer (FC1) maps the feature map to a 512D vector, which is used to calculate similarity scores in the testing phase. During training, another fully connected layer, FC2, is appended to the back of FC1 in order to perform classification.

Training data. We use the publicly available CASIA-WebFace[30] and VGGFace2 [2] datasets (after excluding the images of identities appearing in testing sets) to train our CNN models. CASIA-WebFace has 494,414 face images belonging to 10,575 different identities and the VGGFace2 has 3.1 million images belonging to 8,631 identities. Some example images of the two datasets can be found in Fig. 3. As shown in Fig. 3, there are low-resolution and profile facial images in the WebFace dataset.

Preprocessing. Face alignment is a common preprocessing operation for face recognition. Following these previous works [26, 22, 24, 28, 14], we perform face alignment that guarantees all the eyeballs stay at the same position in the image. The facial landmarks are detected with MTCNN [31] and images are cropped to 112x96 according to detected landmarks. If there are multiple faces detected, we keep the face that is nearest to the image center, and discard all other faces. If there is no face detected, we delete the image if it is a training sample and perform center crop if it is a testing sample.

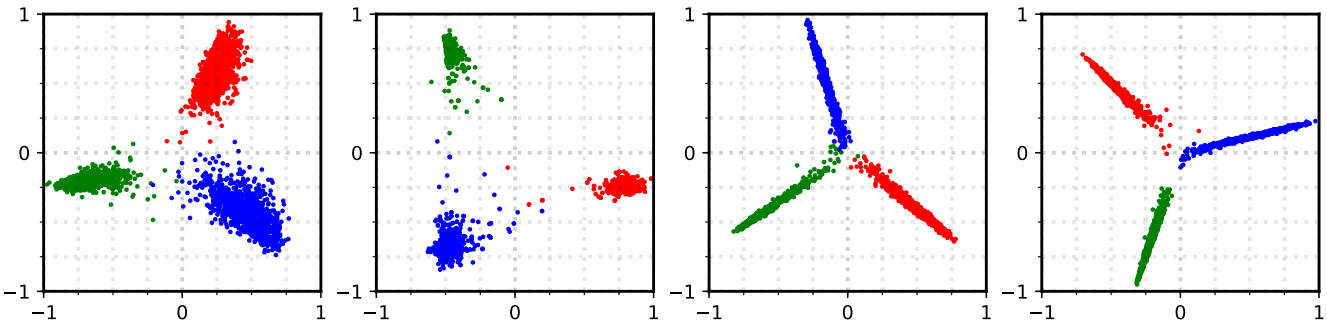
Evaluation Protocol. The performance of the proposed method as well as other competitors is tested on three commonly used face recognition benchmarks: LFW[9], YTF[29] and the MegaFace challenge[10]. We extract the 512 dimensional deep features from the output of FC2 (Fig. 2). For all experiments, the final representation of a testing face is obtained by combining its original face features and its horizontally flipped features. Consequently, we obtain a 1024 dimensional vector for each facial image. For LFW and YTF datasets, we calculate the cosine distance of the two features as the similarity score, and then perform standard 10-fold cross-validation. The testing set is equally divided into 10 folds. 9 of the 10 folds are used as validation set to tune the best threshold and the accuracy is tested on the other fold. For the MegaFace dataset, we use the official evaluation tools[10].

Different Loss Formulas. As aforementioned in Sec. 4.1, the softmax loss $\mathcal{L}_s$ in Eq.(10) can be replaced by other advanced loss functions, say, *center loss*[28] or *angular margin softmax loss*[14]. Therefore, it's reasonable to combine the proposed regularization term $\mathcal{L}_r$ with other other loss



VGGFace2 CASIA-WebFace

Figure 3. Example faces from VGGFace2[2] (left) and CASIA-WebFace[30] (right) datasets, where images of the same column belong to the same identity. Facial images in both datasets present different expressions, lightness, and ages. In general, the image quality of VGGFace2 dataset is better than that of WebFace, because there are more profile faces or low-resolution images in the WebFace dataset, as depicted in the figure.



(a) Softmax loss (b) Center Loss [28] (c) SphereFace [14] (d) RegularFace

Figure 4. Representations of 3 digits MNIST classification. (a) Softmax loss learns separable representations; (b) Center loss[28] enhances intra-class compactness in the Euclidean space; (c) Sphereface[14] clamps intra-class representations in the sphere space; (d) Guided by the proposed exclusive regularization, inter-class representations present a mutual exclusive parttern, evidenced by large inter-class angles.

functions so as to learn both inter-class separable and intra-class compact representations.

In the experiments, we compare three different loss function formulas: Softmax Loss + $\mathcal{L}_r$ (RegularFace + SM), Softmax Loss + Center Loss + $\mathcal{L}_r$ (RegularFace + [28]), Softmax Loss + angular margin + $\mathcal{L}_r$ (RegularFace + [14]). The quantitative results on LFW, YTF datasets (Tab. 2) and the MegaFace challenge (Tab. 3) reveal that the proposed *RegularFace* method can consistently improves the performance.

5.2. Demonstrative Experiment on MNIST

We conduct experiments on MNIST dataset to demonstrate the geometric character of the proposed method. As discussed in Sec. 3.2, in real face recognition systems [15, 14, 8], the representation dimension is redundant compared with number of identities. To make our case similar to face recognition, we train LeNet[12] on a subset of the MNIST [13] dataset containing only 3 digits, and constrain the representations to a 2D space for the ease of visualization. We train LeNet with three different loss funtions: *center loss*[28], *angular margin softmax loss*[14] and softmax with our proposed *exclusive regularization*. Then we visualize the representations of all testing samples in Fig. 4.

As shown in Fig. 4 (a), the representations come into separatable clusters when the model is trained by softmax loss, and there are clear margins between representations of different digit classes. By penalizing the Euclidean distance between representations and their class-specific centers, *center loss* forces representations of the same class to

get more close and compact (Fig. 4 (b)). Different from the *center loss* that compresses representations in the Euclidean manifold, *angular margin softmax loss* clamps representations of the same class in the hypersphere manifold by imposing an angular margin on the decision boundaries (Fig. 4 (c)).

Both center loss and angular margin softmax loss try to improve model discriminability by emphasizing the intra-class compactness. While on the other side, our proposed *exclusive regularization* method enhances the discriminability by enlarging the inter-class separability. As shown in Fig. 4 (d), when the model is trained with exclusive regularization, representations of different classes tend to be angularly far away from each other.

Method	Data	LFW	YTF
DeepFace [26] (3)	4M	97.35	91.4
FaceNet [22]	4M	**99.65**	**95.1**
DeepID2+ [24]	4M	98.70	-
DeepID2+ [24] (25)	4M	99.47	93.2
Center Loss [28]	0.7M	99.28	94.9
Softmax Loss (SM)	WebFace	97.88	90.1
Center Loss [28]		98.91	93.4
L-Softmax [15]		99.01	93.0
SphereFace [14]		99.26	94.1
RegularFace+SM		99.02	91.9
RegularFace+[28]		99.18	93.7
RegularFace+[14]		**99.33**	**94.4**
Softmax Loss (SM)	VGGFace2	98.55	93.4
Center Loss [28]		99.31	94.3
L-Softmax [15]		99.35	94.1
SphereFace [14]		99.50	95.9
RegularFace+SM		99.32	94.7
RegularFace+[28]		99.39	95.1
RegularFace+[14]		**99.61**	**96.7**

Table 2. Performance comparison on the LFW [9] dataset. ($\mathcal{X}$) means the method ensemble $\mathcal{X}$ models. **RegularFace+[X]** represents the joint supervision of *exclusive regularization* and the loss function proposed in relevant paper.

5.3. Experiments on LFW and YTF

LFW dataset[9] includes 13,233 face images from 5749 different identities. YTF dataset [29] includes 3,424 videos from 1,595 different individuals, with an average of 2.15 videos per person. The clip durations vary from 48 frames to 6,070 frames, with an average length of 181.3 frames. Both LFW and YTF contain faces with large variations in pose, expression and illumination. We train *Center loss*, *SphereFace* and our proposed *RegularFace* using the network architecture in Fig. 2. The original and left-right flipped faces are fed into the model and the respective outputs are concatenated as the final representation. The performance on LFW and YTF datasets is evaluated under the *unrestricted with labeled outside data* protocol [8].

As depicted in Tab. 2, with the help of *exclusive regularization*, our proposed *RegularFace* method can significantly improve the performance of original softmax loss. The **RegularFace+SM** combination outperforms the original softmax loss with a very clear margin. Furthermore, when combined with other intra-class compactness oriented methods, *e.g. center loss* [28] or *SphereFace* [14], the **RegularFace+** [28] and **RegularFace+** [14] methods achieve the state-of-the-artaccuracy. This is mainly because models under joint supervision are able to learn both intra-class compact and inter-class separable representations.

5.4. MegaFace Challenge1 on FaceScrub

The MegaFace challenge [10] is a relatively new dataset which aims at benchmarking the performance of face recognition approaches at million scale distractors. The MegaFace dataset is divided into two subsets: (1) the gallery set containing more than 1 million images from 690K identities, and (2) the probe set which is composed of two existing datasets: Facescrub [18] and FGNet [19]. There are two evaluation protocols in MegaFace according to the scale of training data (**Small** and **Large**). The training set is considered as small when it contains less than 0.5M images, otherwise it is considered as large.

We test our method under both small and large protocols on the facescrub probe set. Our method consistently improves the performance of the original softmax loss, *center loss* and *SphereFace*. The results are given in Tab. 3.

Method	Protocol	Rank1 Acc	Ver.
Softmax loss (SM)	Small	52.86	65.93
L-Softmax [15]		67.13	80.42
Center Loss [28]		65.23	76.52
SphereFace [14]		69.62	83.16
RegularFace+SM		65.91	78.21
RegularFace+[28]		68.37	81.25
RegularFace+[14]		**70.23**	**84.07**
Softmax loss(SM)	Large	61.72	70.52
Center Loss [28]		70.29	87.01
SphereFace [14]		74.82	89.01
RegularFace+SM		72.91	88.37
RegularFace+[28]		73.27	89.14
RegularFace+[14]		**75.61**	**91.13**

Table 3. Verification and identification performance (%) on MegaFace[10] challenge 1. **Rank-1 Acc** means the rank 1 identification accuracy and **Ver** means the verification TAR at 10^{-6} FAR. All the methods are based on the ResNet20 architecture for fair comparison. We train models on the WebFace[30] dataset for **Small** protocol, and VGGFace2[2] dataset for **Large** protocol.

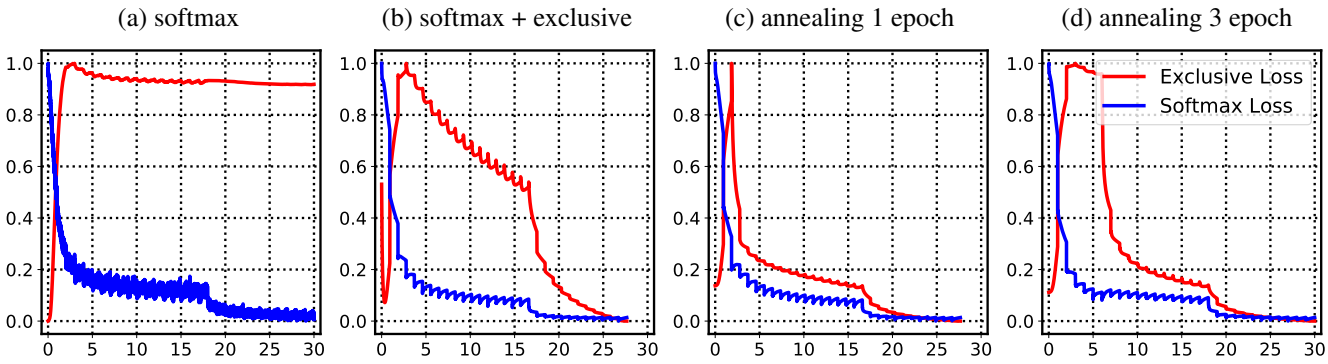


Figure 5. Exclusive loss under different annealing Strategies: (a): Softmax loss ($\lambda = 0$, without exclusive regularization). (b): Softmax loss + exclusive regularization, the exclusive term vibrates at the start of training. (c): Softmax + exclusive regularization with 1 epoch's annealing, the exclusive term becomes more stable. (d): Softmax + exclusive regularization with 3 epoch's annealing.

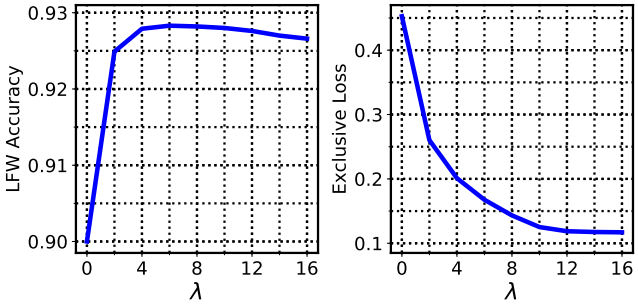


Figure 6. LFW[9] accuracy (left) and converged exclusive loss (right) under different λ.

5.5. Parameter Discussions

The only hyper-parameter our method brings is the weight factor λ in Eq.(10). Given larger λ, the *exclusive regularization* plays more important role in training. If λ is small, however, the softmax loss will dominate the learning procedure. We train the proposed method with different λ, and record the accuracy on LFW as well as the converged exclusive term.

When $\lambda = 0$, the loss described in Eq.(10) reduces to *angular softmax loss* (Fig. 1 (b) and Eq.(4)). As shown in Fig. 6 (a), with λ getting larger, the performance on LFW rapidly increases and achieves the peak performance at $\lambda = 6$. However, the converged exclusive loss $\mathcal{L}_r$ continuously decreases with the increasing of λ (Fig. 6 (a)). When $\lambda = 12$, $\mathcal{L}_r$ reaches 0.0610, nearly equals the optimal value in Tab. 1.

5.6. The Annealing Strategy

To further study the interaction between softmax loss and exclusive loss, we record the training losses during the training process. Results are shown in Fig. 5. In Fig. 5 (a) model is supervised only by the softmax loss ($\lambda = 0$), and in Fig. 5 (b) $\lambda = 1$.

Interestingly, we find that the softmax loss and exclusive loss may 'fight' with each other at the very beginning of training, as evidenced by the unstable vibration of exclusive loss shown in Fig. 5 (b). It quickly decreases to a local minimum at the very beginning, and then rebounds to a high level. After that, the exclusive loss stably goes down until convergence.

To stabilize the training procedure and make the losses decrease smoothly, we utilize the *annealing* strategy which is also referred in [16, 14] to balance two conflict loss terms at the beginning of training. The annealing strategy is a kind of 'warm up' that gradually fortifies the weight of exclusive regularization at starting epochs of training.

Suppose λ is the weight of exclusive regularization, t is time step (epoch), and N is the number of epochs for annealing. The effective exclusive weight λ^* warms up linearly:

$$\lambda^*(t) = \begin{cases} \frac{t}{N} \cdot \lambda, & t <= N \\ \lambda, & otherwise. \end{cases} \tag{15}$$

6. Conclusion

In this paper, we propose the 'exclusive regularization' that explicitly enlarges inter-class distance for discriminative face recognition. Different from existing methods that focus on the intra-class compactness, the proposed regularization term penalizes the angular distance between different cluster centers, leading to large inter-class margins. Comprehensive comparisons on several large open face benchmarks show that the proposed method can consistently improve the performance of the existing methods and outperforms state-of-the-arts, demonstrating the superiority of our algorithm.

Acknowledgements. This research was supported by NSFC (61572264), the national youth talent support program, Tianjin Natural Science Foundation (17JCJQJC43700, 18ZXZNGX00110) and the Fundamental Research Funds for the Central Universities (Nankai University, NO. 63191501).

References

[1] Stephen Boyd, Lin Xiao, and Almir Mutapcic. Subgradient methods. *Lecture notes of EE392o, Stanford University*, 2003.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74, 2018.

[3] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE conf Comput Vis Pattern Recog.*, volume 1, pages 539–546. IEEE, 2005.

[4] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018.

[5] Schroff Florian, Kalenichenko Dmitry, and Philbin James. Facenet: A unified embedding for face recognition and clustering. In *IEEE conf Comput Vis Pattern Recog.*, volume 1, 2015.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conf Comput Vis Pattern Recog.*, 2016.

[7] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *IEEE conf Comput Vis Pattern Recog.*, pages 1875–1882, 2014.

[8] Gary B. Huang and Erik Learned-Miller. Labeled faces in the wild: updates and new reporting procedures. Technical report, University of Massachusetts, Amherst, 2014.

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

[10] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE conf Comput Vis Pattern Recog.*, pages 4873–4882, 2016.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Adv Neural Inform Process Syst.*, pages 1097–1105. 2012.

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. 86(11):2278–2324, 1998.

[13] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *AT&T Labs*, 2, 2010.

[14] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE conf Comput Vis Pattern Recog.*, volume 1, 2017.

[15] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. pages 507–516, 2016.

[16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[17] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Adv Neural Inform Process Syst.*, pages 4826–4837, 2017.

[18] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *IEEE Conf Image Process (ICIP)*, pages 343–347. IEEE, 2014.

[19] Gabriel Panis, Andreas Lanitis, Nicholas Tsapatsoulis, and Timothy F Cootes. Overview of research on facial ageing using the fg-net ageing database. *Iet Biometrics*, 5(2):37–46, 2016.

[20] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, page 6, 2015.

[21] Swami Sankaranarayanan, Azadeh Alavi, Carlos Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. *IEEE Conference Biometrics Theory, Applications and Systems*, 2016.

[22] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE conf Comput Vis Pattern Recog.*, pages 815–823, 2015.

[23] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *IEEE conf Comput Vis Pattern Recog.*, volume 1, 2014.

[24] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *IEEE conf Comput Vis Pattern Recog.*, pages 2892–2900, 2015.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE conf Comput Vis Pattern Recog.*, pages 2818–2826, 2016.

[26] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE conf Comput Vis Pattern Recog.*, pages 1701–1708, 2014.

[27] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE conf Comput Vis Pattern Recog.*, 2018.

[28] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision.*, pages 499–515. Springer, 2016.

[29] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE conf Comput Vis Pattern Recog.*, pages 529–534. IEEE, 2011.

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

[31] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.