UniformFace: Learning Deep Equidistributed Representation for Face Recognition

Yueqi Duan[1,2,3], Jiwen Lu[1,2,3,*], Jie Zhou[1,2,3]
[1]Department of Automation, Tsinghua University, China
[2]State Key Lab of Intelligent Technologies and Systems, China
[3]Beijing National Research Center for Information Science and Technology, China
duanyq14@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn

Abstract

In this paper, we propose a new supervision objective named uniform loss to learn deep equidistributed representations for face recognition. Most existing methods aim to learn discriminative face features, encouraging large inter-class distances and small intra-class variations. However, they ignore the distribution of faces in the holistic feature space, which may lead to severe locality and unbalance. With the prior that faces lie on a hypersphere manifold, we impose an equidistributed constraint by uniformly spreading the class centers on the manifold, so that the minimum distance between class centers can be maximized through complete exploitation of the feature space. To this end, we consider the class centers as like charges on the surface of hypersphere with inter-class repulsion, and minimize the total electric potential energy as the uniform loss. Extensive experimental results on the MegaFace Challenge I, IARPA Janus Benchmark A (IJB-A), Youtube Faces (YTF) and Labeled Faces in the Wild (LFW) datasets show the effectiveness of the proposed uniform loss.

1. Introduction

Face recognition has attracted much attention over the past three decades, and a variety of face recognition methods have been proposed in the literature [2, 1, 20, 25, 10, 9]. In general, there are four main procedures in a practical face recognition system: face detection, face alignment, face representation and face matching. As faces in the wild condition usually suffer from large variations which reduce inter-class separability and intra-class compactness, face representation plays a key role by extracting discriminative features to separate faces from different persons [25].

With the fast development of deep learning, recent years have witnessed significant improvement of convolutional neural networks (CNNs) based face representation [37, 34, 33, 35, 32, 30, 42, 24, 45, 23, 46, 41]. There are three key attributes that determine the discriminative power of the learned CNN features: training data, network architecture and loss function. The amount of employed data largely affects the training procedure of CNNs, where large-scale face datasets have been presented in recent years such as VGGFace [30], VGGFace2 [3], MS-Celeb-1M [11], IJB-A [18] and MegaFace [17, 28]. Moreover, data-augmentation methods have been developed to improve the performance and avoid overfitting [27]. Deeply learned features also benefit from the development of network architectures, where representative CNN models include AlexNet [19], VGG [30] and GoogLeNet [36]. Facing the tremendous increase in the amount and complexity of training data, deeper structures such as ResNet [13] and DenseNet [14] have been designed to strengthen the learning ability. The last attribute is to design highly efficient loss function, which provides effective gradients for learning discriminative CNN features [12, 34, 33, 32, 42, 24, 8, 45, 23, 7]. In this paper, we mainly focus on the third aspect of how to design a more effective loss function.

Softmax loss is widely used in training CNN features [37, 34], which is defined as a combination of the last fully connected layer, a softmax function and a cross-entropy loss [24]. However, we only learn separable features through softmax loss with limited discriminative power. To address the limitation, various supervision objectives have been proposed to enhance the discriminativeness of the learned features, such as contrastive loss [33], triplet loss [32], center loss [42], large-margin softmax (L-Softmax) loss [24] and range loss [45]. While most existing loss functions impose constraints of Euclidean margin, SphereFace [23] shows the effectiveness of angular margin by mapping faces on a hypersphere manifold with angular softmax (A-Softmax) loss. However, all these methods aim to enhance the discriminative power of the learned features,

* Corresponding author

(a) SphereFace-2D (b) UniformFace-2D (c) SphereFace-3D (d) UniformFace-3D

Figure 1. Comparison of SphereFace and UniformFace on the LFW dataset, where points in different colors represent the learned deep features from varying classes, and the numbers on the figure are located at the class centers. We apply LeNet-5 by modifying the dimention of the last hidden layer into 2 and 3, respectively, so that we can directly visualize the learned deep features on the 2D and 3D hypersphere manifold. In the toy examples, we only employ three identities from the LFW dataset for clear illustration. While both SphereFace and UniformFace achieve high accuracy on 2D and 3D situations, we observe that the class centers of UniformFace are more equidistributed than SphereFace for complete exploitation of the holistic feature space. (Best viewed in color.)

which fail to consider the distribution of faces in the holistic feature space and may suffer from high locality and unbalance.

In this paper, we argue that the distribution of the features should be considered as an essential property. On one hand, the learned features may be located very locally on the manifold, which fail to fully exploit the feature space. On the other hand, the minimum average inter-class distance for each class, i.e., the distance between one class and its nearest neighbouring class, may have large variance due to the unbalanced distribution, where some classes especially face the risk of being misclassified. Fig. 1 (a) and (c) show the visualization results of SphereFace on the LFW dataset [15], where the class centers are nonuniformly distributed. Even though faces could be multi-modal, highly non-uniform distribution (e.g. varying identities from the same modality are gathered together) still leads to less discriminativeness as the large gap between different modalities is a "waste" of feature space. To this end, we propose a new objective function named uniform loss to learn equidistributed representations for face recognition, which is an ideal goal to fully exploit the feature space. Motivated by the fact that the electric potential energy of like charges on the surface of a sphere is minimized when they are uniformly distributed, we consider the class centers as like charges with repulsion and formulate the objective of potential energy minimization as the uniform loss. Through the joint supervision of A-Softmax loss and uniform loss, the class centers of the learned features are uniformly spread on the hypersphere manifold, so that the minimum average inter-class distance is maximized with uniform distribution, and we term the learned features as *UniformFace*. We observe that the classes are more equidistributed on the hypersphere manifold in Fig. 1 (b) and (d) with the additional supervision signal of uniform loss. Experimental results on the MegaFace Challenge I [17], IJB-A [18], YTF [43] and LFW [15] datasets validate that the proposed uniform loss effectively boosts the performance of face recognition.

2. Related Work

Face recognition is a long-standing computer vision problem, where the methods can be mainly divided into two categories: hand-crafted representation and learning-based representation. Hand-crafted methods require strong prior knowledge for the researchers to engineer the feature extractors by hand. For example, Gabor wavelets [21] and LBP [1] firstly computed the textural information or gradient in local regions, and then generated holistic features for face representation. While hand-crafted methods are heuristic and data-independent, learning based methods learn face representations in a data-driven manner. For example, Cao *et al.* [4] presented a learning-based descriptor (LE) by learning an encoder in an unsupervised manner. Lei *et al.* [20] learned a LBP-like feature named discriminant face descriptor (DFD) with the LDA criterion. Duan *et al.* [10] proposed a context-aware local binary feature learning (CA-LBFL) method to obtain bitwise interacted binary codes for face recognition.

In recent years, deep face representation learning methods have achieved a series of breakthrough [37, 34, 33, 35, 32, 30, 42, 24, 45, 23, 46, 41]. Pioneering works include DeepFace [37] and DeepID [34], which employed softmax loss for training deeply learned features. Parkhi *et al.* [30] proposed a "very deep" VGG network and created a reasonably large face dataset. While softmax loss only guarantees the separability of the features, several new supervision objectives have been proposed to enhance the discriminative power [33, 32, 42, 24, 8, 23, 7]. For example, Sun *et al.* [33] proposed a joint identification-verification signal. Schroff *et al.* [32] demonstrated the effectiveness of triplet loss. Wen *et al.* [42] presented a center loss to improve the intra-class compactness. Zhang *et al.* [45] proposed a range loss to address the long-tailed distribution of train-

ing data. Liu *et al.* [24] enlarged the angular separability of features through a large-margin softmax (L-Softmax) loss. They also proposed an angular softmax (A-Softmax) loss by constraining the learned features on a hypersphere manifold [23]. However, these loss functions aim to enhance the discriminative power of the learned features, ignoring the distribution of features on the holistic feature space, which may lead to high locality and unbalance in feature distribution.

3. Proposed Approach

In this section, we first revisit the A-Softmax loss [23] which maps faces on a hypersphere manifold. Then, we detail the proposed uniform loss and introduce the deep equidistributed representation UniformFace. Lastly, we highlight the differences between the uniform loss and two relevant objectives, and demonstrate the necessity of simultaneous supervision.

3.1. Revisiting the A-Softmax Loss

The softmax loss has been widely applied in various visual recognition tasks, and its formulation is represented as follows:

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathbf{W}_{y_i}^T\mathbf{x}_i+b_{y_i}}}{\sum_{j=1}^{M}e^{\mathbf{W}_j^T\mathbf{x}_i+b_j}}. \tag{1}$$

In (1), $\mathbf{W}_j \in \mathbb{R}^d$ is the weights in the last fully connected layer of the jth class and d is the feature dimension. $b_j \in \mathbb{R}$ is the bias term which is omitted for the sake of concision below. $\mathbf{x}_i \in \mathbb{R}^d$ is the learned deep feature of sample i, and y_i is the ground truth class label. N and M are the number of samples and classes, respectively. The features learned by softmax loss have an intrinsic angular distribution, which suggests the cosine distance as the metric rather than the Euclidean distance [23]. To this end, a modified softmax loss is formulated as follows by constraining $\|\mathbf{W}_i\|$ to 1:

$$L_m = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|\mathbf{x}_i\|\cos(\theta_{y_i,i})}}{\sum_{j=1}^{M}e^{\|\mathbf{x}_i\|\cos(\theta_{j,i})}}, \tag{2}$$

where the decision boundaries depend on angles. Given a query point, we compare the angles with the weights of each class and choose the minimum one as the result, so that the features are evaluated on a hypersphere manifold. SphereFace [23] manipulates decision boundaries to produce angular margin through an A-Softmax loss, where the angle between the sample point and the target class is multiplied by the margin parameter m:

$$L_{a-s} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|\mathbf{x}_i\|\psi(\theta_{y_i,i})}}{e^{\|\mathbf{x}_i\|\psi(\theta_{y_i,i})}+\sum_{j\neq y_i}e^{\|\mathbf{x}_i\|\cos(\theta_{j,i})}}. \tag{3}$$

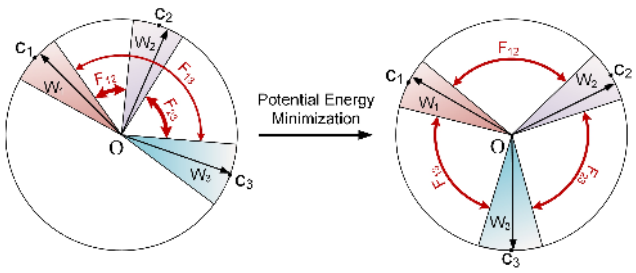


Figure 2. An illustration of learning uniformly distributed representations on the hypersphere manifold through potential energy minimization, where different colors represent varying classes. We define the repulsion F between classes which is inversely proportional to the square of the distance. In the figure, the linewidth represents the magnitude of the repulsion. Through the minimization of potential energy, the distribution of classes becomes uniform, and the maximum repulsion for each class obtained from the nearest neighbouring class is equal and minimized. The situation of 2D hypersphere manifold is shown for easy illustration.

In (3), $\psi(\theta_{y_i,i})$ is defined as $(-1)^k\cos(m\theta_{y_i,i})-2k$, $\theta_{y_i,i}\in[\frac{k\pi}{m},\frac{(k+1)\pi}{m}]$, $k\in[0,m-1]$ instead of $\cos(m\theta_{y_i,i})$ to address the limitation that $\theta_{y_i,i}$ is restricted in $[0,\frac{\pi}{m}]$. While the A-Softmax loss aims to enlarge the angular distance between classes and constrain the features from the same class to a smaller hypersphere area, it fails to consider the distribution of features in the holistic hypersphere manifold. As shown in Fig. 1, A-Softmax fails to completely exploit the holistic feature space, which leads to unbalanced distribution.

3.2. UniformFace

As the faces from different classes should be separated, we consider the class centers as like charges with equal quantity, where each center repels the others. In order to learn uniformly distributed class centers on the hypersphere manifold, we define the uniform loss as the potential energy of all the centers, and the class centers will be equidistributed through potential energy minimization. Fig. 2 shows an illustration of the proposed uniform loss. Motivated by Coulomb's law, we set the repulsion between two class centers $\mathbf{c}_{j_1}$ and $\mathbf{c}_{j_2}$ inversely proportional to the square of the distance:

$$F = \lambda\frac{1}{d(\mathbf{c}_{j_1},\mathbf{c}_{j_2})^2}, \tag{4}$$

where $d(\mathbf{c}_{j_1},\mathbf{c}_{j_2})$ is the distance between the centers $\mathbf{c}_{j_1}$ and $\mathbf{c}_{j_2}$. In this paper, we obey the conventional Coulomb's law by using the Euclidean distance rather than the angular distance. Moreover, we add one for each distance to prevent from too large repulsion, i.e., $d(\mathbf{c}_{j_1},\mathbf{c}_{j_2}) = ||\mathbf{c}_{j_1}-\mathbf{c}_{j_2}||_2+1$.

With the definition of (4), we obtain the potential energy

Algorithm 1: UniformFace

Input: Training set $\{\mathbf{x}_i\}$, training labels $\{y_i\}$, number of classes M, Parameters Θ of CNN, hyperparameter λ, and iteration numbers T.

Output: The parameters Θ.

```
Initialize Θ and the class centers c_j.
for iter = 1, 2, ··· , T do
    Sample a mini-batch from the training set.
    for j = 1, 2, ··· , M do
        Update the class centers c_j with (7).
    end for
    Update the parameters Θ with (8).
end for
return Θ.
```

of the center $\mathbf{c}_{j_1}$ affected by $\mathbf{c}_{j_2}$:

$$E = \int_{d(\mathbf{c}_{j_1},\mathbf{c}_{j_2})}^{\infty} \lambda \frac{1}{x^2} dx = \lambda \frac{1}{d(\mathbf{c}_{j_1},\mathbf{c}_{j_2})}, \tag{5}$$

where the potential energy of $\mathbf{c}_{j_2}$ is the same as $\mathbf{c}_{j_1}$. In order to learn equidistributed representations, we minimize the total potential energy of all the class centers as our uniform loss. As potential energy is scalar quantity, we formulate the uniform loss with the average of all the pairwise energies, which is represented as follows:

$$L_u = \frac{\lambda}{M(M-1)} \sum_{j_1=1}^{M} \sum_{j_2 \neq j_1} \frac{1}{d(\mathbf{c}_{j_1},\mathbf{c}_{j_2})}. \tag{6}$$

As the class centers $\mathbf{c}_j$ are continuously changing during the training procedure, we require to utilize the entire training set to update $\mathbf{c}_j$ in each iteration, which is not applicable in practice. Therefore, we employ a modified method by updating the centers on each mini-batch [42]:

$$\Delta \mathbf{c}_j = \frac{\sum_{i=1}^{n} \delta(y_i = j) \cdot (\mathbf{c}_j - \mathbf{x}_i)}{1 + \sum_{i=1}^{n} \delta(y_i = j)}, \tag{7}$$

where n is the number of samples in a mini-batch, $\delta(\cdot) = 1$ if the condition is true and $\delta(\cdot) = 0$ otherwise.

We employ the simultaneous supervision of A-Softmax loss and uniform loss to learn discriminative and equidistributed features as follows:

$$L = L_{a-s} + L_u, \tag{8}$$

where the parameter λ in L_u balance the weights of different terms, and SphereFace can be seen as a special case for $\lambda = 0$. We optimize the CNN by standard SGD. Algorithm 1 details the proposed UniformFace.

3.3. Discussion

In this subsection, we first compare the proposed uniform loss with two relevant supervision objectives: A-Softmax loss and center loss, and then discuss the necessity of simultaneous supervision.

Comparison with A-Softmax Loss and Center Loss: In recent years, several supervision signals have been proposed to learn more discriminative deep face representation, where the most relevant objectives are A-Softmax loss [23] and center loss [42]. A-Softmax aims to learn discriminative features on the hypersphere manifold. However, it fails to explicitly constrain the distribution on the holistic feature space, where the faces may be located locally and unbalanced. Center loss simply minimizes the distances between the intra-class faces and the corresponding class center, ignoring the inter-class relationships of class centers. The proposed uniform loss considers the inter-class repulsion and encourages equidistributed class centers on the hypersphere manifold, so that the feature space is completely exploited and the minimum distance between class centers can be maximized.

Necessity of Simultaneous Supervision: In UniformFace, we simultaneously employ the A-Softmax loss and uniform loss as the training objective. On one hand, if we only supervise CNN by A-Softmax loss, the faces will suffer from nonuniform distribution on the hypersphere manifold. On the other hand, if we simply utilize the uniform loss, the intra-class variations will be unconstrained only to guarantee the uniform distribution of the class centers. Therefore, it is of significant necessity to employ simultaneous supervision for discriminative and equidistributed deep representations.

4. Experiments

In this section, we conducted extensive experiments on four widely-used face recognition datasets to demonstrate the effectiveness of the proposed UniformFace, which included the MegaFace Challenge I [17], IJB-A [18], YTF [43] and LFW [15] datasets.

4.1. Implementation Details

Detailed Setup of CNN: We utilized MXNet package [6] through the experiments and employed ResNet [13] as the CNN architecture for all the datasets. Fig. 3 detailed the employed architecture of the CNN. Throughout the experiments, we set m to 4 for L_{a-s} as suggested in [23]. We fixed the parameter λ as 1 through cross-validation on the YTF and LFW datasets. The model was trained under the batchsize of 128 on four GTX 1080Ti GPUs for acceleration. We initialized the learning rate as 0.1, which was divided by 10 at the 16K, 24K iterations.

Preprocessing: We performed standard preprocessing

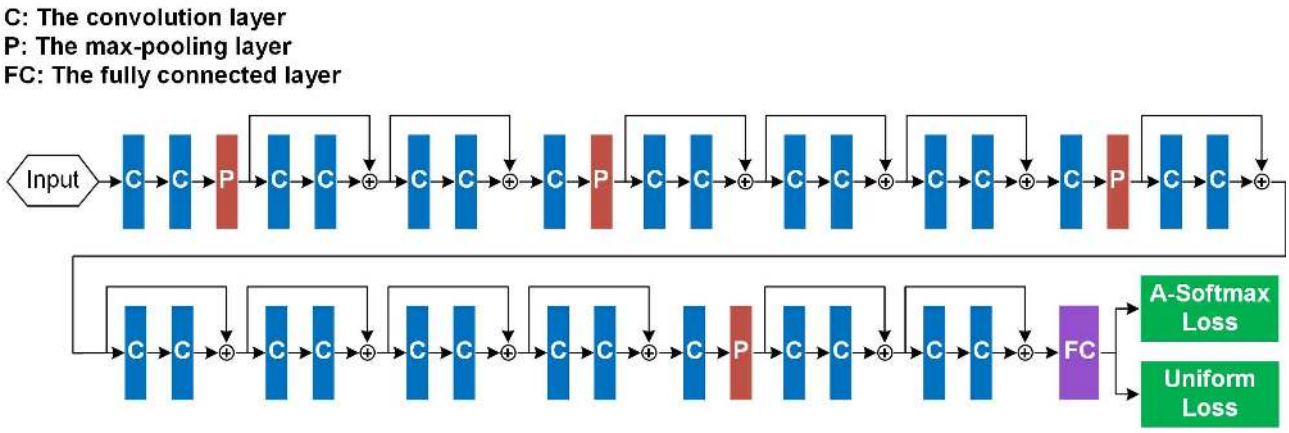


Figure 3. The CNN architecture adopted in UniformFace. The parameters of CNN are supervised by a joint signal of A-Softmax loss and uniform loss. The dimension of the fully connected layer is 512. (Best viewed in color.)

on faces. We detected and aligned each face from training sets and test sets with MTCNN [44] through five landmarks (two eyes, two mouth corners and nose), and cropped the image into 112×112. We also normalized each pixel in RGB images by subtracting 127.5 and then dividing by 128.

Training: We trained our model on the refined MS-Celeb-1M [11] and VGGFace2 [3] datasets. MS-Celeb-1M originally contained about 10M images from 100K identities. We removed the images which were far away from the class centers to improve the quality of the training data and cleared the identities with less than 3 images to relieve the long-tail distribution [8, 7]. The refined MS-Celeb-1M dataset contained 85K identities with 3.84M images. VGGFace2 consisted of 9,131 subjects with 3.31 million images. We employed the training split to optimize our UniformFace, containing 8,631 classes with 2.21M faces.

Testing: We extracted UniformFace from the output of the fully connected layer, and we concatenated the features of original faces and horizontally flipped faces as the final representation. Therefore, the dimension of the final representation is 1,024 for each face. We employ the nearest neighbour classifier with cosine distance for face identification and verification.

4.2. Datasets

We conducted experiments on four widely-used face recognition benchmark datasets, where we followed the standard evaluation protocols to evaluate the effectiveness of UniformFace.

The MegaFace Dataset: MegaFace [17] is a challenging benchmark dataset, which is designed to evaluate the performance of face recognition algorithms at the million scale. The MegaFace dataset consists of a gallery set and a probe set. The gallery set is a subset of Flickr photos from Yahoo, which consists of more than 1 million pictures from 690K individuals. The probe dataset contains two existing datasets: FaceScrub [29] and FGNet. FaceScrub is a publicly available dataset with 100K photos from 530 unique individuals, where 55,742 images are male and 52,076 images are female. FGNet is a face aging dataset containing 1,002 images of 82 identities. Each identity has multiple facial images at a wide range of ages (from 1 to 69).

The IJB-A Dataset: IJB-A [18] is an increasingly concerned public dataset which offers challenges to the field of face detection and face recognition by unconstrained image settings. The IJB-A dataset contains 5,397 images and 20,412 video frames split from 2,042 videos of 500 individuals with extreme pose, illumination and expression conditions. We employ 10 folders with different random collection of 333 subjects for training and 167 for testing. Face verification (1:1) and identification (1:N) are both evaluation protocols for IJB-A challenge. The verification protocol consists of about 1,756 positive pairs and 9,992 negative pairs in each folder, and the identification protocol contains 112 gallery templates and around 1,763 probe templates, where 55 randomly selected subjects are removed from the gallery for difficulty. Face verification tests true accepted rates (TAR) under varying false accepted rates (FAR). Face identification performance is measured by a Cumulative Match Characteristics (CMC) curve, which infers the identification rate within the top-K retrieval candidates.

The YTF Dataset: YTF [43] contains 3,425 videos of 1,595 different persons downloaded from YouTube, with varying variations of pose, illumination and expression, which is a popular dataset for unconstrained face recognition. In YTF, there are about 2.15 videos available for each person and a video clip has 181.3 frames on average.

The LFW Dataset: LFW [15] is a famous web-collected image dataset for face recognition, which contains 13,233 images from 5,749 different identities. The images are captured from the web in wild conditions, varying in pose, illumination, expression, age and background, lead-

Table 1. Rank-1 identification accuracy (%) with 1M distractors and verification TAR at 10^{-6} FAR (%) on the MegaFace dataset.

Method	Protocol	@Rank-1	@FAR=10^{-6}
YouTu Lab	Large	83.29	91.34
NTechLAB-facenx	Large	73.30	85.08
Vocord-DeepVo3	Large	**91.76**	94.96
DeepSense V2	Large	81.30	95.99
Shanghai Tech	Large	74.05	86.37
Google-FaceNet	Large	70.50	86.47
Beijing FaceAll-N	Large	64.80	67.12
Beijing FaceAll	Large	63.98	63.96
CosFace [41]	Large	82.72	**96.65**
GRCCV	Small	77.68	74.89
DeepSense	Small	70.98	82.85
SIAT - MMLAB	Small	65.23	76.72
Center Loss [42]	Small	65.23	76.52
L-Softmax [24]	Small	67.13	80.42
SphereFace [23]	Small	72.73	85.56
CosFace [41]	Small	77.11	89.88
SphereFace* [23]	Large	76.65	92.32
UniformFace	Large	**79.98**	**95.36**

ing to large intra-class variations.

4.3. Experiments on MegaFace

We evaluated the proposed UniformFace on FaceScrub of MegaFace Challenge 1, including both face identification and face verification tasks. We followed the protocol of *large* training set as the training dataset contains more than 0.5M images, where the identities appearing in FaceScrub were removed from the training set. We employed the original test set of MegaFace for fair comparisons.

Comparison with the State-of-the-Arts: Table 1 shows the experimental results on the MegaFace dataset compared with the existing deep learning based methods. In the face identification task, the similarity between the probe face and each gallery face is computed, where 1M distractors exist in the gallery set to make the task more challenging. We report the Rank-1 identification accuracy in the table to meet the practical demand. In the face verification task, we need to decide whether a pair of faces are of the same identity. The TAR is reported with 10^{-6} FAR.

We observe that the proposed UniformFace achieves comparable results with the state-of-the-art deep learning based methods. In Table 1, SphereFace* is to train the network only with A-Softmax loss, fixing the same network structure and training data for fair comparisons. We can see that UniformFace outperforms SphereFace* as the class centers are more equidistributed. Under the supervision of uniform loss, the hypersphere manifold is completely exploited, and the minimum distance between class centers can be maximized. The comparison shows the effectiveness of the proposed uniform loss through the final recognition rates on the MegaFace dataset. Fig. 4 shows the CMC and ROC curves of different methods on the MegaFace dataset.

Table 2. The comparison of the minimum average inter-class distances with or without the uniform loss.

Method	Mean	Variance	Least 1,000
SphereFace*	1.13	0.10	0.45
UniformFace	**1.45**	**0.06**	**0.55**

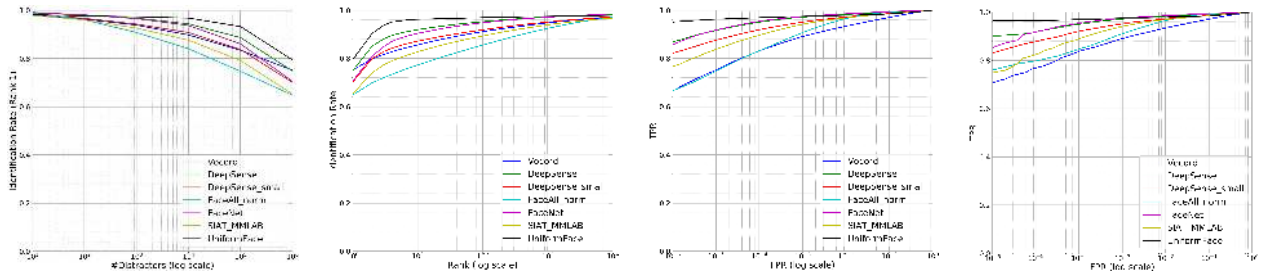
Table 4. Comparison of Rank-1 accuracy (%) with more baselines including SphereFace (SF), ArcFace (AF) and CosFace (CF).

SF	Ours (SF)	AF	Ours (AF)	CF	Ours (CF)
76.65	**79.98**	79.14	**81.46**	81.59	**83.53**

Evaluation of Uniformity: One of the most essential property for UniformFace is the equidistributed class centers. In the previous experiments, we show that the utilization of uniform loss L_u successfully boosts the face recognition rate. However, a more direct evaluation is required to show the improvement in feature distribution. In order to better evaluate the uniformity of the learned representation, we conducted an experiment for comparing the distribution of class centers supervised with or without the uniform loss.

For each class center, we computed its nearest distance to other class centers, which can be considered as the minimum average inter-class distance for the selected center. The minimum average inter-class distance represented the similarity of the selected class with the most dangerous class. For the M minimum distances from all the M classes, we compared their means, variances and means of the least 1,000 inter-class distances between SphereFace* and UniformFace. Table 2 illustrates that uniform loss leads to large and uniform minimum average inter-class distances as the class centers are more equidistributed, where the mean value increases by 28% (from 1.13 to 1.45) and the variance decreases by 40% (from 0.10 to 0.06). Moreover, while nonuniform distribution suffers from locality where some classes gather in local spaces, our UniformFace relieves such locality with large least 1,000 inter-class distances.

Adaptation to More Baselines: Recently, more and more methods have been presented in the angular space and achieved outstanding performance, such as SphereFace [23], ArcFace [7] and CosFace [41]. Technically, the proposed uniform loss can be generally applied to these methods as it is designed based on angular space instead of the specific SphereFace. Table 4 shows that the proposed uniform loss successfully boosts the performance of all the baselines, which presents its good generalization ability.



(a) CMC-distractors (b) CMC-rank (c) ROC-1M (d) ROC-10K

Figure 4. Comparison of (a) CMC curves with Rank-1 accuracy under varying numbers of distractors, (b) CMC curves with 1M distractors under varying rank-K, (c) ROC curves with 1M distractors, and (d) ROC curves with 10K distractors.

Table 3. Verification TAR at 10^{-2} and 10^{-3} FAR (%) and Rank-1 and Rank-5 identification accuracy (%) on the IJB-A dataset.

Method	@FAR=10^{-2}	@FAR=10^{-3}	@Rank-1	@Rank-5
DCNN [5]	78.7 ± 4.3	-	85.2 ± 1.8	93.7 ± 1.0
DCNN (fusion) [5]	83.8 ± 4.2	-	90.3 ± 1.2	96.5 ± 0.8
Triplet Similarity [31]	79.0 ± 3.0	59.0 ± 5.0	88.0 ± 1.5	95.0 ± 0.7
PAM [26]	73.3 ± 1.8	55.2 ± 3.2	77.1 ± 1.6	88.7 ± 0.9
3DMM [38]	60.0 ± 5.6	-	76.2 ± 1.8	89.7 ± 1.0
LSFS [40]	72.9 ± 3.5	51.0 ± 6.1	82.2 ± 2.3	93.1 ± 1.4
DR-GAN [39]	77.4 ± 2.7	53.9 ± 4.3	85.5 ± 1.5	94.7 ± 1.1
PRN [16]	96.5 ± 0.4	91.9 ± 1.3	$\mathbf{98.2 \pm 0.4}$	$\mathbf{99.2 \pm 0.2}$
SphereFace* [23]	92.3 ± 1.6	88.4 ± 4.2	93.2 ± 1.3	96.5 ± 1.1
UniformFace	$\mathbf{96.9 \pm 0.8}$	$\mathbf{92.3 \pm 1.7}$	$\mathbf{97.9 \pm 0.5}$	$\mathbf{98.8 \pm 0.2}$

4.4. Experiments on IJB-A

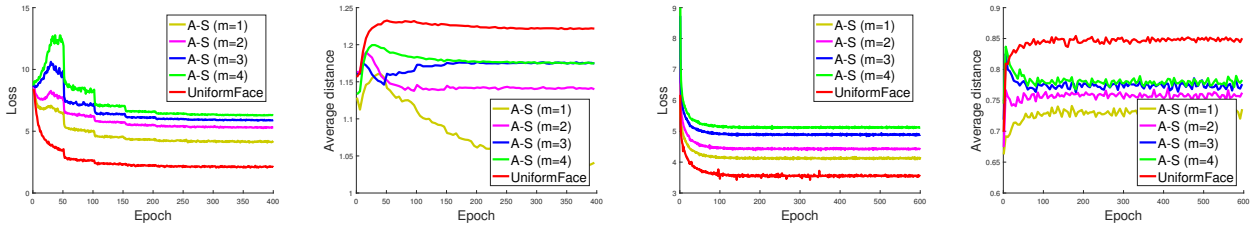
We evaluated our UniformFace on both verification and identification tasks, where we reported the TAR at 10^{-2} and 10^{-3} FAR for the verification task, and Rank-1 and Rank-5 accuracy for the identification task. Table 3 shows the experimental results of UniformFace and existing methods on the IJB-A dataset. In the compared methods, PAM [26], 3DMM [38] DR-GAN [39] and PRN [16] are recent pose-aware face recognition methods, which effectively address the extreme pose variance of the faces. However, UniformFace achieves very competitive results compared with these methods as a general face recognition method. While PAM, 3DMM, DR-GAN and PRN exploit strong prior information of poses, UniformFace enhances the robustness by encouraging equidistributed representation. As aforementioned, the minimum average inter-class distances are maximized with uniform distribution, which leads to stronger robustness. Moreover, the uniform loss successfully boosts the performance on the IJB-A dataset, which demonstrates its effectiveness on faces with large pose variations.

Table 5. Verification rate (%) of UniformFace compared with the state-of-the-art methods on the YTF and LFW datasets.

Method	Data	Model	YTF	LFW
DeepFace [37]	4M	3	91.4	97.4
FaceNet [32]	200M	1	95.1	99.7
VGG [30]	2.6M	1	97.3	99.0
DeepID2+ [35]	300K	1	-	98.7
DeepID2+ [35]	300K	25	93.2	99.5
Center Loss [42]	0.7M	1	94.9	99.3
Range Loss [45]	1.5M	1	93.7	99.5
Baidu [22]	1.3M	1	-	99.1
L-Softmax [24]	0.5M	1	-	98.7
SphereFace [23]	0.5M	1	95.0	99.4
Ring Loss [46]	3.5M	1	-	99.5
CosFace [41]	5M	1	97.6	99.7
PRN [16]	2.8M	1	95.8	99.7
SphereFace* [23]	6.1M	1	96.1	99.5
UniformFace	6.1M	1	**97.7**	**99.8**

4.5. Experiments on YTF and LFW

In this subsection, we evaluated our UniformFace on the widely-used YTF and LFW datasets. For the YTF dataset, we followed the protocol of unrestricted with labeled outside data, which contained 5,000 video pairs. For the LFW dataset, we also followed the protocol of unrestricted with labeled outside data, where we tested on 6,000 face pairs.



(a) YTF-Loss (b) YTF-Distance (c) LFW-Loss (d) LFW-Distance

Figure 5. Training curves of loss (L_{a-s}) and mean of minimum average inter-class distances on the YTF and LFW datasets.

Table 5 shows the experimental results of UniformFace compared with the state-of-the-art methods on the YTF and LFW datasets, which include DeepFace [37], FaceNet [32], DeepID2+ [35], Range Loss [45], SphereFace [23], Ring Loss [46], CosFace [41] and PRN [16]. From the table, we observe that the usage of uniform loss boosts the performance of 1.7% on YTF and 0.3% on LFW, which decreases the error rates by 41% (from 3.9% to 2.3%) and 60% (from 0.5% to 0.2%), respectively. The main reason is that uniform loss leads to equidistributed representations, which completely exploit the holistic feature space. While DeepFace and DeepID2+ employ multiple models and FaceNet is trained with more than 200M data, UniformFace still outperforms these methods on both YTF and LFW datasets, which demonstrates the effectiveness of the proposed approach.

4.6. Ablation Study

In this subsection, we conducted ablation studies to further demonstrate the effectiveness of UniformFace. Besides quantitative experimental results on benchmark datasets, we first designed apple-to-apple comparisons of training curves and mean of minimum average inter-class distances to compare A-Softmax (under varying m) and our UniformFace. We initialized the network with AlexNet and an additional fully connected layer to reduce dimension to 128, finetuning with the same training data from YTF and LFW. Fig. 5 shows the curves of A-Softmax loss (L_{a-s}) and mean of minimum average inter-class distances. Larger m encourages larger inter-class angular margin, which leads to more discriminativeness and difficulty in learning. However, it does not explicitly reduce the variations of minimum average inter-class distances, while UniformFace has smaller standard deviation (0.02 vs. 0.04 on YTF and 0.05 vs. 0.13 on LFW) as well as larger means for $m = 4$ in A-Softmax.

Then, we tested the effectiveness of the uniform loss (L_u) in learning high-dimensional equidistributed representations. While it is relatively hard to theoretically guarantee uniform distribution, we conducted an experiment to test the uniformity on high-dimensional hypersphere. Given N noise vectors z sampled from the standard normal distribution, we aim to learn a mapping $f(z)$ to the hypersphere space with a 4-layer fully connected network (ReLU for the first three layers), supervised by uniform loss. We set the dimension as 128 and N as 256 for easy design of a ground truth uniform example $[0, \cdots, \pm 1, \cdots, 0]$ (with only one element as 1 or -1 and the others as 0). We compare the mean and standard deviation of minimum distances, which are $\sqrt{2} \pm 0$ for uniform distribution, 1.20 ± 0.02 for the learned mapping, and 0.44 ± 0.04 for random mapping. We observe a similar uniform phenomenon with 2D and 3D cases for high dimensions.

5. Conclusion

In this paper, we have proposed a uniform loss to learn equidistributed representations for face recognition. Unlike existing supervision signals which ignore the distribution of classes and suffer from high locality, the proposed uniform loss considers the class centers as like charges with intra-class repulsion, so that they will be spread uniformly on the hypersphere manifold through potential energy minimization. Under the joint supervision of A-Softmax loss and uniform loss, we maximize the minimum average inter-class distances for all the classes with complete exploitation of the holistic feature space. Extensive experimental results on MegaFace, IJB-A, YTF and LFW demonstrate the effectiveness of the proposed UniformFace. As we set the same quantity of charge for each class center, the inter-class repulsion is only relevant to the distances in this situation. It is an interesting future work to learn data-dependent quantity of charge for varying classes to obtain fine-grained distribution of representations.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, Grant 61672306, Grant 61572271. The authors would like to thank Mr. Cheng Ma and Mr. Haomiao Sun for valuable discussions.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 28(12):2037–2041, 2006. 1, 2

[2] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *TPAMI*, 19(7):711–720, 1997. 1

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017. 1, 5

[4] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *CVPR*, pages 2707–2714, 2010. 2

[5] Jun-Cheng Chen, Vishal M Patel, and Rama Chellappa. Unconstrained face verification using deep CNN features. In *WACV*, pages 1–9, 2016. 7

[6] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. In *NIPSW*, 2015. 4

[7] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 1, 2, 5, 6

[8] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *CVPRW*, 2017. 1, 2, 5

[9] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning rotation-invariant local binary descriptor. *TIP*, 26(8):3636–3651, 2017. 1

[10] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Context-aware local binary feature learning for face recognition. *TPAMI*, 40(5):1139–1153, 2018. 1, 2

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102, 2016. 1, 5

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742, 2006. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 4

[14] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 1

[15] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 2, 4, 5

[16] Bong-Nam Kang, Yonghyun Kim, and Daijin Kim. Pairwise relational networks for face recognition. In *ECCV*, pages 628–645, 2018. 7, 8

[17] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882, 2016. 1, 2, 4, 5

[18] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *CVPR*, pages 1931–1939, 2015. 1, 2, 4, 5

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1

[20] Zhen Lei, Matti Pietikäinen, and Stan Z Li. Learning discriminant face descriptor. *TPAMI*, 36(2):289–302, 2014. 1, 2

[21] Chengjun Liu and Harry Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *TIP*, 11(4):467–476, 2002. 2

[22] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015. 7

[23] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017. 1, 2, 3, 4, 6, 7, 8

[24] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, pages 507–516, 2016. 1, 2, 6, 7

[25] Jiwen Lu, Venice Erin Liong, Xiuzhuang Zhou, and Jie Zhou. Learning compact binary face descriptor for face recognition. *TPAMI*, 37(10):2041–2056, 2015. 1

[26] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *CVPR*, pages 4838–4846, 2016. 7

[27] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, pages 579–596, 2016. 1

[28] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *CVPR*, pages 3406–3415, 2017. 1

[29] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *ICIP*, pages 343–347, 2014. 5

[30] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015. 1, 2, 7

[31] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *BTAS*, pages 1–8, 2016. 7

[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 1, 2, 7, 8

[33] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014. 1, 2

[34] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898, 2014. 1, 2

[35] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, pages 2892–2900, 2015. 1, 2, 7, 8

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 1

[37] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 1, 2, 7, 8

[38] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, pages 1493–1502, 2017. 7

[39] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, pages 1415–1424, 2017. 7

[40] Dayong Wang, Charles Otto, and Anil K Jain. Face search at scale. *TPAMI*, 39(6):1122–1136, 2017. 7

[41] Hao Wang, Yitong Wang, Zheng Zhou, Gong Dihong Ji, Xing, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018. 1, 2, 6, 7, 8

[42] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016. 1, 2, 4, 6, 7

[43] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011. 2, 4, 5

[44] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 23(10):1499–1503, 2016. 5

[45] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *ICCV*, pages 5409–5418, 2017. 1, 2, 7, 8

[46] Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *CVPR*, pages 5089–5097, 2018. 1, 2, 7, 8