SeqFace: Learning discriminative features by using face sequences

Wei Hu | Yangyu Huang | Fan Zhang | Ruirui Li | Hengchao Li

1 College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China
2 Microsoft Research Asia, Beijing, China
3 Sichuan Provincial Key Laboratory of Information Coding and Transmission, Southwest Jiaotong University, Chengdu, China

Correspondence
Fan Zhang, College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China.
Email: zhang.fan.cn@ieee.org

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61871413; Fundamental Research Funds for the Central Universities, Grant/Award Number: XK2020-03

Abstract
Deep convolutional neural networks (CNNs) have greatly improved the Face Recognition (FR) performance in recent years. Almost all CNNs in FR are trained on the carefully labeled datasets containing plenty of identities. However, such high-quality datasets are very expensive to collect, which restricts many researchers to achieve state-of-the-art performance. In this paper, a framework, called SeqFace, for learning discriminative face features is proposed. Besides a traditional identity training dataset, the designed SeqFace can train CNNs by using an additional dataset which includes a large number of face sequences collected from videos. Moreover, the label smoothing regularization (LSR) and a new proposed discriminative sequence agent (DSA) loss are employed to enhance the discrimination power of deep face features via making full use of the sequence data. Only with a single ResNet model, the method achieves very competitive performance on several face recognition benchmarks, including LFW, YTF, CFP, AgeDB, and MegaFace. The code and model are publicly available at the website https://github.com/huangyangyu/SeqFace.

KEYWORDS
CNNs, face recognition, face sequences, training data augmentation

1 | INTRODUCTION

Face recognition is a very important research topic of computer vision and pattern recognition in the past two decades [1]. To advance face recognition, a large number of works have been done on collecting face datasets, designing/learning face descriptors, and training discriminative models. The early works aim to collect facial images under the controlled conditions, represent each facial image using hand-crafted features [2], and evaluate face recognition algorithms on the small-scale datasets. Recently, multiple large-scale face datasets are publicly available, and deep convolutional neural networks (CNNs) are widely used in face recognition (FR), due to their great discriminative feature learning capability. The face feature is mainly trained via two types of methods according to their loss functions in CNN models. One method uses classification loss functions, such as classical softmax loss [3, 4], and angular margin softmax loss functions [5–8]. The other type uses metric learning loss functions, such as contrastive loss and triplet loss [9, 10]. In many recent CNNs for FR, two types of loss functions are usually combined together for learning face features. All these loss functions aim to maximize the inter-identity variations and minimize the intra-identity variations under a certain metric space. No matter which loss functions are applied, these methods share the same type of the training data, called identity data in the paper.

An identity dataset includes $M$ faces of $N$ identities, and each face in the dataset is clearly labeled as the image of the $i^{th}$ ($0 \leq i < N$) identity. Currently most public or private datasets for training deep face features, such as CASIA [11], MS-Celeb-1M [12] and CelebFaces [13], belong to identity datasets. However, a large-scale high-quality identity dataset is very expensive to construct, since it could cost lots of effort and money. Identity data need two kinds of information: face image and identity annotation. Identities in most public and private datasets are celebrities, because celebrity photos are rather easily crawled and annotated from the Internet. However, a celebrity dataset might be not a satisfied training dataset, if there are obvious differences between the evaluated faces and the celebrity faces in age, race, pose, and so on.
Beside facial images collected from the Internet, videos (movies, TVs, surveillance videos, etc.) can also provide large quantities of face images, but few works utilize these face images so far because labelling identities is relatively difficult. Face detection and tracking on videos can automatically generate data with lots of face sequences, and each sequence contains several faces of one identity. We call this type of data as sequence data.

A sequence dataset includes \( M \) faces of \( N \) sequences, and each face is labeled as the image of the \( i^{th} \) (\( 0 \leq i < N \)) sequence. A large-scale high-quality sequence dataset can be efficiently and automatically constructed by using state-of-the-art face detection and tracking methods. Although face sequences are broadly used in video FR applications, previous works have rarely utilized these unlabelled sequence datasets as the training data to learn face features, because identity labels cannot be easily assigned. However, we know that faces in one sequence should belong to one identity, and it is possible to utilize this feature to reduce the intra-identity variations while training face feature CNN models.

In this paper, we propose SeqFace to learn discriminative face features on both identity data and sequence data (see Figure 1). SeqFace aims to make full use of sequence data in training. In SeqFace, a CNN model is jointly supervised by two loss functions. The first one is a chief softmax or angular margin softmax loss with label smoothing regularization (LSR). The second one is a discriminative agent (DSA) loss. These two losses can maximize the inter-identity variations and minimize the intra-identity variations simultaneously. With the help of sequence data, CNN models can be trained with high feature discrimination in SeqFace.

To summarize, our major contributions are as follows:

1) We present SeqFace to learn discriminative face features. Besides the traditional identity data, unlabeled sequence data are used as the training data to enhance the discriminative power of face features for the first time, since faces in a sequence should belong to one identity.

2) To make full use of sequence data, we employed the LSR to help the softmax-like loss to deal with sequence data. A new DSA loss function, which contributes to the intra-class compactness and the inter-class dispersion of the features, is also proposed to train CNNs. Experiments demonstrate that the LSR and the DSA losses both boost the FR performance greatly.

3) We conduct experiments on some popular and challenging FR benchmark datasets with one single ResNet-64, and these experiments demonstrate that the proposed SeqFace obtains state-of-the-art performance on these benchmarks. Section 4 presents some experiments to demonstrate the performance of SeqFace. Finally, conclusion and discussion are provided in Section 5.

The paper is organized as follows. In Section 2, we give a brief introduction of previous works related to the proposed method. Section 3 provides a detailed description of the proposed SeqFace for learning face features on sequence data. In Section 4, we analyze the performance of the SeqFace method on the public datasets. Eventually, Section 5 draws the conclusions.

2 RELATED WORKS

In this section, we briefly review works of deep face recognition, and face sequences related works in FR are also introduced.

2.1 Deep face recognition

Deep face recognition is one of the most active field, and has achieved a series of breakthroughs in recent years thanks to the great success of CNNs [14–17]. Many methods [3, 4, 13, 18–20] have proven that CNNs outperform humans in FR on some benchmark data sets.
Face features are discriminative if their intra-class compactness and inter-class separability are well maximized. The loss function should be carefully designed to accomplish this goal. FR is firstly treated as a multi-class classification problem and CNN models are supervised by the classical softmax loss in many methods [3, 4], but the softmax loss cannot fully guarantee the above goal in theory. Later, some metric learning loss functions, such as contrastive loss [10, 11], triplet loss [9, 18] are applied to boost FR performance greatly. However, well-designed sample mining strategies restrict the application of these losses. Recently, some angular margin softmax [5, 6, 8] and normalization [7, 21, 22] based methods are proposed and achieve outperforming performance, since they encourage larger inter-class and smaller intra-class variance at the same time, and are constructed on the classical softmax loss functions. Other loss functions [23, 24] based on metric loss also demonstrate effective performance on FR. Moreover, some auxiliary loss [25, 26] are employed to train models together with classification loss functions. Recently, some approaches focus on hard examples [27] or AutoML searching [28, 29] to get better loss functions. Though these loss functions achieve better and better performance on FR, but they all depend on full-labeled identity datasets, such as MS-Celeb-1M and Casia, and cannot treat face sequences as the training data to learn discriminative face features.

## 2.2 Sequences in face recognition

In many applications, sequences or image sets are the most natural form of input to the FR system. Video face recognition methods [30–37] based on face sequences, or face sets, also are expected to achieve better performance than ones based on individual images. Most of these studies attempt to utilize redundant information of face sequences/sets to improve recognition performance, but not to learn discriminative features from sequence data. Recently, some approaches [34, 36–39] aim to learn deep video features for video face recognition. In [37], large-scale unlabeled face sequences are employed as the training data, but these sequence data are only utilized to learn transformations between image and video domains. [38] trained a self-supervised Siamese network to obtain the features of a face cluster instead of a single face. [39] also aims to cluster faces in videos more accurately. To conclude, learning discriminative face features still depends on traditional large-scale identity datasets in this deep CNN approaches of video FR.

## 3 THE PROPOSED APPROACH

### 3.1 SeqFace framework

The proposed SeqFace is a framework for learning discriminative face features on identity datasets and sequence datasets simultaneously. In the identity dataset, faces of one identity are labeled as the same ID\textsubscript{identity}. In the sequence dataset, faces in one sequence are labeled as the same ID\textsubscript{sequence}. Two faces with the different ID\textsubscript{identity} must belong to different identities, but two faces with the different ID\textsubscript{sequence} might (or might not) belong to one identity. Considering these two datasets together, there are two circumstances: the identity overlap between these two datasets exists or does not exist.

SeqFace can deal with these two circumstances. However, in order to achieve better performance, we encourage to remove this identity overlap before training, because more constrains can be added in the loss functions as discussed later. Fortunately, removing the identity overlap is not a time-consuming task in many real scenarios. For example, it is almost certain that people in an Asian street surveillance video will not appear in the MS-Celeb-1M [12] dataset.

In SeqFace, a CNN model (ResNet-like models in our implementation) is jointly supervised by one chief classification loss and one auxiliary loss in SeqFace. The final loss can be formulated as

$$L = L_{Chief} + \eta L_{Auxiliary}$$  

(1)

where $\eta$ is a parameter used to balance two loss functions.

Similar with many methods, we also treat the FR problem as a classification task to train CNNs, and CNNs are mainly supervised by a chief classification loss, such as the Softmax loss, the A-Softmax loss in SphereFace [8], and so on. However, although all faces in identity data is labeled as belonging to one class (identity) in the classification loss, all input faces in sequence data cannot belong to any class (identity) in the classification loss. Traditional classification loss cannot deal with sequence data. In SeqFace, we employ the LSR in the chief classification loss to solve this problem as discussed in Section 3.2.

We know that faces in one sequence certainly belong to one identity. Therefore, if a loss encourages the intra-sequence feature compactness, and does not penalize the inter-sequence feature compactness, it could supervise CNNs to learn discriminative face features on sequence data, and it could naturally deal with identity data too. Because this loss mainly affects the intra-sequence and intra-identity compactness, it has to be an auxiliary loss. The center loss [25] is such a loss, but it only concerns the intra-identity and intra-sequence compactness. In order to make full use of sequence information, a DSA loss is presented as an auxiliary loss in the SeqFace as discussed in Section 3.3.

### 3.2 Label smoothing regularization

The softmax loss is applied to supervise CNNs classification, and its simplicity and probabilistic interpretation make the softmax loss widely adopted in FR issues. The softmax loss is the combination of a softmax function and a cross-entropy loss, and the cross-entropy loss is formulated as

$$L_s = - \sum_{i=1}^{C} \log(p(i))q(i),$$  

(2)
where \( C \) is the class number, \( p(i) \in [0, 1] \) is the predicted probability (the output of the softmax function) of the input belonging to class \( i \), and \( q(i) \) is the ground truth distribution defined as

\[
q(i) = \begin{cases} 0 & i \neq y \\ 1 & i = y 
\end{cases}
\]

(3)

where \( y \) is the ground truth class label of the input.

In FR problem, \( q(y) \) should be set to 1 for each face belonging to the \( j \)th identity in identity data. The softmax loss cannot be directly employed to deal with sequence data because of two sequences might belong to the same identity. We employ Label smoothing regularization (LSR) introduced in [40] to deal with sequence faces in Softmax loss. In the LSR, the value of \( q(i) \) can be a float value between 0 and 1 for the input which cannot be clearly labeled as any class.

If the identity overlap does not exists, we define \( q(i) = 1/C \) as [41] for faces in sequence data, in order to keep \( \sum_{i=1}^{C} q(i) = 1 \) (required in Softmax loss). Since \( C \) usually is a large number, \( 1/C \) is close to 0. Furthermore, \( q(i) = 1/C \) means that the feature of a face in sequence data cannot be close to features of all identities in identity data too. Therefore, the cross-entropy loss is rewritten as

\[
L = -(1 - Z) \log(p(y)) - \frac{Z}{C} \sum_{i=1}^{C} \log(p(i)),
\]

(4)

where \( Z = 0 \) for the input face of identity data, and \( Z = 1 \) for the input face of sequence data.

If we are not sure if there is identity overlap, all \( q(i) \) are set to 0 for all input faces in sequence data, then \( L_{\delta} \) is always to be 0, which is equivalent to all faces in sequence data do not participate in training at all.

The LSR can also be integrated into other softmax-like classification loss functions. In our implementation, a feature normalized SphereFace (L2-SphereFace for short in the paper, same with F-Norm SphereFace in [5]) is applied as the chief classification loss. An additional \( L_{\alpha} \)-constraint is added to the regular SphereFace [8], it means the input feature \( \bar{x_k} \) must be firstly normalized and scaled by a scalar parameter \( \delta (\delta = \|\bar{x_k}\|_2) \). Therefore, the decision boundaries of the L2-SphereFace under binary classification is \( \delta(\cos m\theta_1 - \cos \theta_2) = 0 \) for class 1, and \( \delta(\cos \theta_1 - \cos m\theta_2) = 0 \) for class 2. In our implementation, the parameter \( \delta \) and the margin \( m \) are set to 32.0 and 4, respectively. Experiments in Section 4.2 demonstrate the effectiveness of LSR.

### 3.3 DSA loss

In this section, we further propose a new auxiliary loss, namely discriminative sequence agent loss (DSA Loss), which concerns the intra-class compactness and the inter-class dispersion, and deals with sequence data simultaneously.

First, considering the traditional classification problem with an identity dataset, we define

\[
d_{k,a} = \frac{(\bar{x}_k - \bar{z}_a)^2}{4},
\]

(5)

as the distance between the feature \( \bar{x}_k \) of the \( k \)th training sample and the feature center \( \bar{z}_a \) of the \( n \)th class(identity), \( d_{k,a} \) is actually equivalent to the Euclidean distance. Note that if \( \bar{x}_k \) and \( \bar{z}_a \) are normalized, \( d_{k,a} \) can be re-formulated as

\[
d_{k,a} = \frac{(1 - \cos \theta_{k,a})}{2},
\]

(6)

where \( \theta_{k,a} \) denotes the angle between \( \bar{x}_k \) and \( \bar{z}_a \), and \( d_{k,a} \) can be regarded as the angular distance. Since our target is to reduce the distance between \( \bar{x}_k \) and \( \bar{z}_a \) and enlarge other distances between \( \bar{x}_k \) and \( \bar{z}_a \) for all \( n \neq y_k \), where \( y_k \) is the label of the \( k \)th training sample, a discriminative loss can be formulated as

\[
L_{k,a} = \begin{cases} d_{k,a} & n = y_k \\ \max(\alpha \cdot d_{k,a} + \beta, 0) & n \neq y_k 
\end{cases},
\]

(7)

where \( \alpha \in [1, +\infty) \) and \( \beta \in [0, +\infty) \) are two parameters to adjust the discriminative power of the learned features. Therefore, the final loss function is

\[
L_D = \frac{1}{K} \sum_{k=1}^{K} \left[ \lambda L_{k,y_k} + (1 - \lambda) \frac{1}{(N - 1)p} \sum_{n=1}^{N} b(1, p)L_{k,a} \right],
\]

(8)

where the parameter \( \lambda \) is applied to balance the intra-class compactness and the inter-class dispersion, \( N \) is the number of identity (class) of the identity dataset. We introduce another parameter \( p \) as the probability that the \( n \)th center is employed in computing the final loss, because \( N \) might be a huge number and it will be time-consuming if all \( L_{k,a} \) are computed in each iteration. \( b(1, p) \) means the Bernoulli distribution with the probability \( p \).

The gradients of \( L_D \) with respect to \( \bar{z}_a \) and the update equation of \( \bar{z}_a \), similar with that in the center loss, are computed as:

\[
\frac{\partial L_D}{\partial \bar{z}_a} = \frac{1}{K} \left[ \lambda \frac{\bar{x}_k - \bar{z}_a}{2} + (1 - \lambda) \frac{1}{(N - 1)p} \right] \sum_{n=1}^{N} b(1, p) \delta(L_{k,a} > 0)(\alpha \frac{\bar{z}_k - \bar{z}_a}{2} - \frac{\bar{x}_k - \bar{z}_a}{2})],
\]

(9)

and

\[
\Delta \bar{z}_a = -\frac{\sum_{n=1}^{N} \delta(y_k = n) \cdot \frac{\bar{x}_k - \bar{z}_a}{2}}{1 + \sum_{n=1}^{N} \delta(y_k = n)},
\]

(10)
where $\delta_{\text{condition}} = 1$ if the condition is satisfied, and $\delta_{\text{condition}} = 0$ if not.

According to Equation (8), the feature $\tilde{x}_k$ is pulled towards the feature center $\tilde{c}_iy_i$ of its identity, and is pushed away from feature centers of other identities randomly selected in each training iteration.

### 3.3.1 DSA for sequence data

Taking into account sequence data, there is a slight modification in Equation (8) to compute the final DSA loss. We assume that there are $C$ identities (we label them 0 to $C-1$) in the identity dataset and $H$ sequences (we label them $C$ to $C+H-1$) in the sequence dataset. The discriminative loss (Equation 7) is reformulated as

$$L_{k,a} = \begin{cases} d_{k,a} \gamma \cdot \max(\alpha \cdot d_{k,a} - d_{k,a} + \beta, 0) & n = y_k \ n \neq y_k \end{cases},$$

where $\gamma$ indicates whether the distance between the $n^{th}$ and the $j_{k}^{th}$ features should be enlarged. If the identity overlap exists between the identity and sequence datasets, $\gamma$ can be set to 1 only when the $k^{th}$ sample is an identity sample and the $n < C$. Otherwise, $\gamma$ is set to 1.

Moreover, the final loss function (Equation 8) is rewritten as

$$L_{D} = \frac{1}{K} \sum_{k=1}^{K} \left[ \lambda L_{k,j_k} + (1-\lambda) \left( \frac{\delta}{(N-1)\rho} \sum_{n=1}^{N} b(1, \rho) L_{k,n} \\ + \frac{\eta}{(H-1)\rho} \sum_{n \neq j_k}^{C+H} b(1, \rho) L_{k,n} \right) \right],$$

where $\delta$ and $\eta$ are also Boolean values. If the $k^{th}$ sample is selected from the identity dataset, $\delta$ and $\eta$ are set to 1. If the $k^{th}$ sample is selected from the sequence dataset, only $\delta$ is set to 1. That is to say, if the $k^{th}$ sample is in the identity dataset, $\tilde{x}_k$ should be pushed away from feature centers of other identities and all sequences, or $\tilde{x}_k$ is only pushed away from feature centers of identities. Figure 2 illustrates two examples.

There are four parameters ($\lambda$, $\alpha$, $\beta$, and $\rho$) in the DSA loss function. The parameter $\lambda$ can be set to 0.5 since we concern both the intra-class compactness and the inter-class dispersion. The parameters $\alpha$ and $\beta$ are used to adjust the discriminative power of features. Using larger values is preferred, but it will increase the difficulty of convergence in training. According to our experiments, $\alpha = 2.0$ and $\beta = 1.0$ can be applied in most applications. The parameter $\rho$ is applied to select part of identities/sequences while computing $L_{k,n}$, in order to reduce the computing cost. The value of the parameter $\rho$ can be set flexibly based on computing resources in real applications.

### 3.3.2 MNIST example

We perform a toy example on the MNIST dataset [42] with our DSA loss. LeNet++ [25], a deeper and wider version of LeNet, is employed. The last hidden layer output of the model is restricted to 2-dimensions for easy visualization (see Figure 3). For comparison, we train 4 models supervised by a softmax loss, a softmax loss and a center loss, a softmax loss and a DSA loss, a softmax loss and a DSA loss (with normalized $\tilde{x}_k$ and $\tilde{c}_i$), respectively. We set $\lambda = 0.5$, $\alpha = 2.0$, $\beta = 1.0$ and $\rho = 1.0$ in the DSA loss. The loss weight values of the center/DSA loss are set to 0.04. All models are trained with the batch size of 32. The learning rate begins with 0.01, and is divided by 10 at 14K iterations. The training process is finished at 20K iterations. As shown in Figure 3, the features learned with the DSA loss are more discriminative. The feature dispersion in Figure 3(b) and Figure 3(c) demonstrates that the DSA loss can enlarge inter-class distances, and the feature centers of different classes are pushed away from each other. Table 1 lists the classification accuracies of 4 models on MNIST test set. From the results, we can get the following observations: (1) the center loss and the DSA loss both improve the classification performance; (2) as an auxiliary loss, the proposed DSA loss outperforms the center loss.
4 | EXPERIMENT

4.1 | Implementation details

In our experiments, all the face images and their landmarks are detected by MTCNN \[43\]. The faces are aligned by similar transformation as \[44\], and are cropped to 144 x 144 RGB images (randomly cropped to 128 x 128 in training). Each pixel in RGB images is normalized by subtracting 127.5 then divided by 128.

4.1.1 | Training and testing

Caffe \[45\] is used to implement CNN models. Different CNN models are employed in the experiments, which will be further introduced. All weights of the auxiliary losses (\(\eta\) in Equation 1) are set to 0.04 in the experiments. Euclidean distances (do not normalize \(\vec{x}_k\) and \(\vec{c}_n\) in Equation 5) are applied in the DSA loss functions used in these section. At the testing stage, only features of the original image are directly extracted from the last full connected layer of CNNs, and the cosine similarity is used to measure the feature distance in the experiments. More details are presented in the corresponding sections. The code and model are publicly available at the website\(^1\). All verification accuracies of SeqFace are calculated by averaging the results of 5 trained models.

4.2 | Exploration experiment

In this section, the employed CNN is a ResNet-20 network which is similar to \[8\], and it is trained on the publicly available CASIA-WebFace dataset \[11\] containing about 0.5M faces from 10,575 identities. All models are trained with the batch size of 32 on one TitanX GPU. The learning rate begins with 0.01, and is divided by 10 at 200K iterations. The training process is finished at 300K iterations.

Because there is no public FR training dataset containing sequence data, we employ CASIA-WebFace to generate one in order to facilitate to reproduction and comparison. To evaluate the effectiveness of sequence data, 10,575 identities in the CASIA-WebFace dataset are randomly divided into two parts:

---

\(^1\) [https://github.com/huangyangyu/SeqFace](https://github.com/huangyangyu/SeqFace)
TABLE 2  Face verification accuracy on LFW dataset

| Model | Loss          | Training Dataset | Accuracy |
|-------|---------------|------------------|----------|
| I     | Softmax       | Dataset A        | 95.12%   |
| II    | L2-SphereFace | Dataset A        | 97.35%   |
| III   | L2-SphereFace + Center loss | Dataset A | 97.45%   |
| IV    | L2-SphereFace + DSA loss | Dataset A | 98.25%   |
| V     | LSR-L2-SphereFace | Dataset A + B | 97.57%   |
| VI    | LSR-L2-SphereFace + DSA loss | Dataset A + B | 98.85%   |
| VII   | L2-SphereFace | CASIA-WebFace    | 99.03%   |

the dataset A (5,000 identities) and the dataset B (5,575 identities). Faces in the dataset B is then randomly split into 32,996 sequences. The dataset A and B are treated as the identity dataset and the sequence dataset respectively. It is clear that the CASIA-WebFace dataset contains full annotations and the dataset A+B only contains part annotations of 10,575 identities. To demonstrate the effectiveness of LSR and DSA loss on the dataset with part annotations, we train 7 models (see Table 2) for comparison.

First, we use a regular softmax loss (Model I), a L2-SphereFace loss (Model II) to train 2 CNN models respectively. Only the dataset A is used as the training dataset. Verification accuracies demonstrate that the L2-SphereFace greatly boosts the performance.

Then, the center loss and our DSA loss are applied as the auxiliary loss to jointly supervise the CNN models (Model III and Model IV, respectively) with a L2-Sphere loss. The reported results demonstrate that two auxiliary loss functions have positive effect on the FR performance, and our DSA loss outperforms the center loss.

Moreover, sequence data (the dataset B) are added to train the CNN model supervised by a LSR-based L2-SphereFace (Model V). According to results, we can conclude that LSR also plays a positive role in training. LSR and DSA loss then work together to train a CNN model (Model VI), and we can have following observation: our SeqFace greatly enhances the discriminative power of learned features.

Last, we also train a model (Model VII) on total CASIA-WebFace with a L2-SphereFace loss. Comparing accuracies between the Model V and VI, we can conclude that complete identity annotation is naturally preferred in training datasets, but the little gap shows that competitive performance also can be achieved by making full use of sequence information.

4.3  Evaluation on public benchmarks

In this section, we evaluate the proposed SeqFace on several public benchmarks, including Labelled Faces in the Wild (LFW) [46], YouTube Faces (YTF) [47], Celebrities in Frontal Profile (CFP) [48], Age Database (AgeDB) [49], and MegaFace Challenge [50].

A ResNet-27 model\(^2\) (the architecture is shown in Figure 4) and a ResNet-64 [8] are employed for evaluation. To accelerate the training process, we first train a baseline model under the supervision of the regular L2-SphereFace on the identity dataset only, and then fine-tune the baseline model by using the SeqFace. Our models are trained with batch size of 128 on 4 Titanx GPU. The learning rate begins with 0.01, and is divided by 10 at 300K and 600K iterations. The training is finished at 800K iterations. The models are jointly supervised by a LSR-L2-SphereFace loss and a DSA loss, and are learned on the MS-Celeb-1M and our Celeb-Seq datasets described below. In the DSA loss, \(\lambda = 0.5, \alpha = 2.0, \beta = 1.0\). The parameter \(p\) is set to 0.001 because of the large number of sequences in the Celeb-Seq dataset.

4.3.1  Training datasets

A refined MS-Celeb-1M (4M images and 79K identities) provided by [44] is used as the identity dataset. Since there is no public sequence datasets for training deep CNNs, we first construct a sequence dataset Celeb-Seq-Overlap, which includes about 2.5M face images of 550K face sequences. We firstly extract about 800K face sequences by using MTCNN [43] and Kalman-Consensus Filter (KCF) [51] to detect and track video faces from 32 online TV Channels, then compute image features with the model provided by SphereFace [8]. Noisy faces, and nearly duplicate faces in one sequence are discarded from the dataset automatically. In Celeb-Seq-Overlap, some celebrities can be found in MS-Celeb-1M. Faces of overlap identities with MS-Celeb-1M are then removed to generate another dataset Celeb-Seq. We also remove face images belong to identities that appear in the LFW and YTF test sets. Some face sequences in the Celeb-Seq dataset are shown in Figure 5.

4.3.2  Evaluation on LFW and YTF

LFW [46] and YTF [47] are challenging testing benchmarks released for face verification. LFW dataset contains 13,233 faces of 5749 different identities, with large variations in pose, expression and illuminations. YTF dataset includes 3425 videos of 1595 identities. We follow the unrestricted with labeled outside data protocol. To evaluate performance on YTF, the simple average feature of all faces in a video is applied to compute the final score.

Table 3 reports the verification performance of several methods and some commercial systems. To demonstrate effectiveness of the SeqFace, the performance of our baseline ResNet-27 is also reported in the table. The SeqFace achieves the highest accuracies on these two benchmarks, even compared with other commercial systems. Note that the ArcFace employs the improved ResNets [52]. It is reported in the ArcFace that a regular 50-layer ResNet achieves a 99.71% accuracy on LFW. Moreover, our ResNet-27 and ResNet-64 models achieve 99.50% and

\(^2\)https://github.com/ydwen/caffe-face
*FIGURE 4* The ResNet-27 architecture for the experiments. The CNN is jointly supervised by the LSR-L2-SphereFace and the DSA loss. **ID** denotes the input identity data, **SEQ** denotes the input sequence data, **C** denotes the convolution layer, **P** denotes the max-pooling layer, and **FC** denotes the fully connected layer.

*FIGURE 5* Some face sequences in the Celeb-Seq dataset. All faces are aligned and cropped to $144 \times 144$. Some sequences belong to the same identity (two sequences at top-right corner). Note that the numbers of faces in the sequences are different from each other.

*TABLE 3* Verification accuracies(%) of different methods on LFW and YTF. Accuracies of some commercial systems are cited from the LFW website. Note that the ResNet models in the ArcFace use the improved residual units [52], and the training MS-Celeb-1M dataset used in the ArcFace contains 3.8M images and 85K identities.
99.67% at VR@FAR=0 on LFW. The models will be publicly accessed online in near future. Figure 6 gives an overview of all failure cases of our ResNet-64 model. The results also prove that identity overlap between sequence data and identity data should be removed to achieve better performance.

The SeqFace is only a framework to make use of sequence data. We believe that new loss functions and modern networks (such as a deeper ResNet with improved residual units [52]) can be employed in the SeqFace to further improve the performance.

4.3.3 Evaluation on CFP and AgeDB

There are 500 identities, each with 10 frontal and 4 profile images, in CFP dataset [48]. In this paper, we use the most challenging subset CFP-FP, which includes 10 folders with 350 same-person pairs and 350 different-person pairs, to demonstrate our performance. AgeDB dataset [49] contains 12,240 images of 440 identities. There are four groups of test data with different year gaps, and we also use the most challenging subset AgeDB-30, which includes ten split of face images and each split contains 300 positive examples and 300 negative examples, to demonstrate our performance.

Table 4 reports the verification performance of several recent methods. The SeqFace also achieves the highest accuracies on these two benchmarks.

4.3.4 Evaluation on megaface

MegaFace [50] is a challenging testing benchmark to evaluate the performance of FR methods at the million scale. It includes a probe set and a gallery set. The gallery set consists of more than 1 million images from 690K different identities. The probe sets consists of two datasets: Facescrub and FGNet. We evaluate the model on one of the three gallery set (set 1) with the provided code³ for both face identification and verification protocols. Table 5 shows that our models achieve competitive performance on MegaFace benchmark. For fair comparison with other methods, the features of FaceScrub are extracted directly from the faces with landmarks provided by MegaFace. Moreover, since removing noisy images in the MegaFace can dramatically improve the identification performance (from 82% to 98% in [5]), we report the performance of our models on the original and cleaned versions of MegaFace dataset. The results in Table 5 demonstrate that the SeqFace is a competitive method.

5 Conclusion

A large-scale high-quality dataset for training CNNs in FR is very expensive to construct. Face features learned on publicly available datasets for researchers might not achieve satisfied performance in some circumstances, for example, evaluating Asian people in surveillance videos. Though large amount of face images in the real situation can be collected, assigning labels to these images is still time-consuming. Fortunately, a dataset containing large amount of face sequences can be efficiently constructed by using face detection and tracking methods.

In this paper, we proposed a framework named SeqFace, which can utilize identity and sequence data together to learn highly discriminative face features. A chief classification loss and

³ http://megaface.cs.washington.edu/participate/challenge.html
another auxiliary loss are combined to learn features of these two types of datasets. The LSR is employed to help the chief loss to deal with sequence input. The DSA loss was also proposed to supervise CNNs as an auxiliary loss. We achieved good results on several popular face benchmarks only with a simple ResNet model. We also believe that higher performance can be obtained, if more advanced loss functions ([5, 6]) and CNN architectures [52] are employed.

Compared with other face feature learning approaches, besides the traditional identity data, SeqFace can additionally employ face sequences to learn discriminative face features. As far as we know, SeqFace is the first framework to employ face sequences as training data to learn face features. Although identity-overlap between identity and sequence data is prohibited in SeqFace, avoiding such overlap is not a time-consuming task in many real scenarios as discussed in Section 4. Solving this limitation is also our future work. Moreover, it is obvious that SeqFace also has great potential to be applied in other similar fields, such as Person-reidentification.

ACKNOWLEDGEMENTS
This work was supported in part by the National Natural Science Foundation of China under Grant No. 61871413, and in part by the Fundamental Research Funds for the Central Universities under Grant XK2020-03.

ORCID
Fan Zhang https://orcid.org/0000-0002-2058-2373

REFERENCES
1. Zhao, W.-Y., et al.: Face recognition: A literature survey. ACM Comput. Surv. 35(4), 399–458 (2003)
2. Ahonen, T., et al.: Face description with local binary patterns: Application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 28(12), 2037–2041 (2006)
3. Taigman, Y., et al.: Deepface: Closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708. IEEE Computer Society, Los Alamitos (2014)
4. Sun, Y., et al.: Deep learning face representation from predicting 10,000 classes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898. IEEE Computer Society, Los Alamitos (2014)
5. Deng, J., et al.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4690–4699. IEEE, Piscataway (2019)
6. Wang, H., et al.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274. IEEE, Piscataway (2018)
7. Wang, F., et al.: Normface: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM international conference on Multimedia, pp. 1041–1049. ACM, New York (2017)
8. Liu, W., et al.: Sphereface: Deep hypersphere embedding for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6738–6746. IEEE Computer Society, Los Alamitos (2017)
9. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: Ferragen, A., Pélillo, M., Loog, M. (eds.) SIMBAD. Lecture Notes in Computer Science, vol. 9370, pp. 84–92. Springer, (2015)
10. Chopra, S., et al.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 539–546. IEEE Computer Society, Los Alamitos (2005)
11. Yi, D., et al.: Learning face representation from scratch. arXiv:1411.7923, 2014
12. Guo, Y., et al.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision, pp. 87–102. Springer, Cham (2016)
13. Chen, Y., et al.: Deep learning face representation by joint identification-verification. In: International Conference on Neural Information Processing Systems, pp. 1988–1996. ACM, New York (2014)
14. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Piscataway (2016)
15. Krizhevsky, A., et al.: Imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, pp. 1097–1105. ACM, New York (2012)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint, arXiv:1409.1556, 2014
17. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9. IEEE Computer Society, Los Alamitos (2015)
18. Schroff, F., et al.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823. IEEE Computer Society, Los Alamitos (2015)
19. Sun, Y., et al.: DeepId3: Face recognition with very deep neural networks. arXiv preprint, arXiv:1502.00873, 2015
20. Sun, Y., et al.: Deeply learned face representations are sparse, selective, and robust. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2892–2900. IEEE, Piscataway (2015)
21. Liu, W., et al.: Large-margin softmax loss for convolutional neural networks. In: International Conference on Machine Learning, pp. 507–516. IEEE Computer Society, Los Alamitos (2016)
22. Zheng, Y., et al.: Ring loss: Convex feature normalization for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5089–5097. IEEE, Piscataway (2018)
23. Deng, J., et al.: Marginal loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2006–2014. IEEE, Piscataway (2017)
24. Zhang, X., et al.: Range loss for deep face recognition with long-tailed training data. In: IEEE International Conference on Computer Vision, pp. 5419–5428. IEEE, Piscataway (2017)
25. Wen, Y., et al.: A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision, pp. 499–515. Springer, Cham (2016)
26. Qi, C., Su, F.: Contrastive-center loss for deep neural networks. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2851–2855. IEEE, Piscataway (2017)
27. Wang, X., et al.: Mis-classified vector guided softmax loss for face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence vol. 34, pp. 12241–12248. AAAI Press, Palo Alto (2020)
28. Wang, X., et al.: Loss function search for face recognition. Proc. Mach. Learn. Res. 119, 10029–10038 (2020)
29. Li, C., et al.: AM-LFS: Automl for loss function search. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8410–8419. IEEE, Piscataway (2019)
30. Yamaguchi, O., et al.: Face recognition using temporal image sequence. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, pp. 318–323. IEEE, Piscataway (1998)
31. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2567–2573. IEEE, Piscataway (2010)
32. Wang, W., et al.: Discriminative covariance oriented representation learning for face recognition with image sets. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5749–5758. IEEE, Piscataway (2017)
33. Bashbaghi, S., et al.: Dynamic ensembles of exemplar-svm for still-video face recognition. Pattern Recognit. 69, 61–81 (2017)
34. Ding, C., Tao, D.: Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 1002–1014 (2017)
35. Huang, Z., et al.: A benchmark and comparative study of video-based face recognition on cox face database. IEEE Trans. Image Process. 24(12), 5967–5981 (2015)
36. Parchami, M., et al.: Video-based face recognition using ensemble of haar-like deep convolutional neural networks. In: International Joint Conference on Neural Networks, pp. 4625–4632. IEEE, Piscataway (2017)
37. Sohn, K., et al.: Unsupervised domain adaptation for face recognition in unlabeled videos. In: IEEE International Conference on Computer Vision, pp. 5917–5925. IEEE, Piscataway (2017)
38. Sharma, V., et al.: Self-supervised learning of face representations for video face clustering. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1–8. IEEE, Piscataway (2019)
39. Tapaswi, M., et al.: Video face clustering with unknown number of clusters. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5027–5036 (2019)
40. Szegedy, C., et al.: Rethinking the inception architecture for computer vision. In: Computer Vision and Pattern Recognition, pp. 2818–2826. IEEE, Piscataway (2016)
41. Zheng, Z., et al.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: IEEE International Conference on Computer Vision, pp. 3774–3782. IEEE, Piscataway (2017)
42. LeCun, Y.: The MNIST database of handwritten digits. (1998). http://yann.lecun.com/exdb/mnist/. Accessed January 2019
43. Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett. 23(10), 1499–1503 (2016)
44. Wu, X., et al.: A light CNN for deep face representation with noisy labels. IEEE Trans. Inf. Forensics Secur. 13(11), 2884–2896 (2018)
45. Jia, Y., et al.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia, pp. 675–678. ACM, New York (2014)
46. Huang, G.B., et al.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report, University of Massachusetts (2007)
47. Wolf, I., et al.: Face recognition in unconstrained videos with matched background similarity. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR ’11, pp. 529–534. IEEE Computer Society, Washington, DC (2011)
48. Sengupta, S., et al.: Frontal to profile face verification in the wild. In: IEEE Winter Conference on Applications of Computer Vision, pp. 1–9. IEEE, Piscataway (2016)
49. Moschoglou, S., et al.: AgeDB: The first manually collected, in-the-wild age database. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1997–2005. IEEE, Piscataway (2017)
50. Kemelmachershlizerman, I., et al.: The megaface benchmark: 1 million faces for recognition at scale. In: Computer Vision and Pattern Recognition, pp. 4873–4882. IEEE, Piscataway (2016)
51. Olfati-Saber, R.: Kalman-consensus filter : Optimality, stability, and performance. In: Proceedings of the IEEE Conference Decision and Control, 2009 Held Jointly with the 2009 Chinese Control Conference, pp. 7036–7042. IEEE, Piscataway (2010)
52. Han, D., et al.: Deep pyramidal residual networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6307–6315. IEEE, Piscataway (2017)

How to cite this article: Hu, W., et al.: SeqFace: Learning discriminative features by using face sequences. IET Image Process. 1–11 (2021). https://doi.org/10.1049/ipr2.12243