Face verification algorithm based on Angular Margin triplet

Ning Wang\textsuperscript{1}, Tao Meng\textsuperscript{2}, xiaoyong Chen\textsuperscript{3}, Dunlei Yuan\textsuperscript{2}, Qihong Yue\textsuperscript{1}, Xiangqian Liu\textsuperscript{1a}\textsuperscript{*}

\textsuperscript{1}School Institute of Information Science, Beijing Jiaotong University, Beijing Jiaotong University, Haidian, Beijing, China
\textsuperscript{2}China Railway Jinan Bureau Group Co., Ltd, Jinan, Shandong, China
\textsuperscript{a}email: xqliu@bjtu.edu.cn

Abstract. In recent years, the face recognition algorithm based on CNN mainly focuses on the improvement of loss function. This paper improves the softmax loss function from two aspects: one is to normalize the weight and features of the loss function, the other is to introduce in class cosine similarity judgment on the basis of the original loss function. These improvements eventually make the distribution of features within the class closer and the distance between classes as far as possible. The recognition accuracy and generalization ability of the network are enhanced.

1. Introduction
Face recognition is one of the key research topics in the field of computer vision\textsuperscript{[1-24]}. Early recognition methods are mainly based on hand-designed feature descriptors, which are highly targeted and lack of generalization ability. The extensive use of convolutional neural network makes it possible for the network to automatically extract image features, which promotes the performance of face recognition algorithms.

In recent years, CNN based face recognition algorithms mainly focus on the improvement of loss function\textsuperscript{[18-24]}, and the general direction is based on metric learning or classification loss. With the help of large-scale training data and 3D face alignment model, the feature extracted from twin network is used as the similarity judgment of two input faces by end-to-end learning method, which is close to human recognition level. The network trained by 10000 different identities is extracted from different regions for training and then fused. Sparse automatic encoder combines convolution and pooling to construct network to extract features, and uses softmax regression model to classify features, which performs better on small-scale data sets.

However, the depth of the network is shallow, the abstraction of the features is low, and it does not perform well on large-scale complex scene data sets. Softmax loss function is not discriminative, so people begin to solve this problem from the perspective of metric learning. Contrast loss makes similar features more compact by constructing sample pairs. In order to measure the similarity of human faces, the method of constructing positive and negative pairs by triple loss is used to map human faces into Euclidean space. Well designed binary or triple samples greatly increase the sample size, which is sensitive to performance and time-consuming. Although some methods add regularization term to the loss of classification, so that the function can achieve convex optimization, but the generalization ability is insufficient.
In order to solve the above problems, this paper improves the softmax loss function mainly in the following two parts: one is to normalize the weight and characteristics of the loss function, the other is to introduce in class cosine similarity judgment based on the original loss function. Finally, the distribution of features within the class is more close, and the distance between classes is separated as much as possible. The recognition accuracy and generalization ability of the network are enhanced.

The whole face recognition process includes face position detection, face key point alignment, face feature extraction and similarity calculation and retrieval. This section will briefly introduce the detection algorithm and basic network structure used in the proposed algorithm.

Multi task convolutional neural network mtcnn \cite{19} is composed of p-net which can quickly generate candidate regions, R-Net which can filter high-precision candidate regions, final bounding box and key o-net. Candidate box plus classifier, non maximum suppression, image pyramid, boundary box regression and other techniques are adopted.

Firstly, the input image is scaled several times to form image pyramid with different scales, which makes the algorithm not easily affected by the target scale. The network structure is shown in Figure 1. The regional recommendation network p-net (proposal network) preliminarily extracts the feature calibration frame. Then, the candidate frames are corrected by the bounding box regression variables. Finally, the highly overlapped candidate frames are merged by NMS method. A simple CNN is used to get face candidate frames with certain credibility. Then, the candidate box is input into R-Net (refine network) network to refine the selection, and the candidate box with poor filtering effect is regressed to NMS optimized prediction result through bounding box. Finally, through the o-net (output network) network, which is more complex than R-Net, more monitoring signals are used to distinguish the target area and the regional border regression, and the coordinates of the upper left and lower right corner of the border box and the five face feature points are output. Increase the diversity of convolution layer, as a secondary classification task, reduce the number and size of convolution kernel.

\begin{equation}
L_i^\text{det} = -(y_i^\text{det} \log(p_i) + (1 - y_i^\text{det})(1 - \log(p_i)))
\end{equation}

The learning objective of the bounding box regression is the regression problem, which measures the offset between the abscissa and ordinate of each prediction candidate window $\hat{y}_i^{\text{box}}$ and the top left corner of the nearest real window $y_i^{\text{box}}$, the height and width of the bounding box, and selects the square loss function as the optimization objective.

Face key point detection includes the coordinates of eyes, nose and mouth corner, and also selects a similar square loss function. Do the last three training samples of N sources, $\gamma$ represents the importance of tasks respectively, and $y_i^\text{det} \in \{0,1\}$ is the real tag value. It synthesizes the loss function constructed by all tasks.
The depth of the network has a great influence on the recognition accuracy. Shallow network can't extract face features effectively, too deep network training process is prone to gradient explosion or dispersion, also prone to network performance degradation problem, resulting in greater error. The deep residual network RESNET solves the problem that the error rate increases instead of decreasing when the network depth increases. Therefore, the basic model of the algorithm in this section selects the deep residual network. In Table 1, conv1/2/3/4 represents a convolution unit composed of convolution layer and residual unit. \([3 \times 3, 64] \times 1, S2\) represents a convolution layer composed of 2, 64 convolution kernels. Each convolution layer is followed by a residual cell. The input is the feature mapping of different scales, FC is the fully connected layer, and the output is 512 dimensional facial features.

| Layer type | Conv1.x | Conv2.x | Conv3.x | Conv4.x | FC |
|------------|---------|---------|---------|---------|----|
| Layer      | \([3 \times 3, 64]\) | \([3 \times 3, 128]\) | \([3 \times 3, 256]\) | \([3 \times 3, 64]\) |    |
| structure  | \(\times 1, S2\) | \(\times 1, S2\) | \(\times 1, S2\) | \(\times 1, S2\) |    |
| output     | \(56 \times 48\) | \(28 \times 24\) | \(14 \times 12\) | \(7 \times 6\) | \(512\) |

2. Design of face verification model based on Angular Margin triple

After in-depth comparative analysis of the advantages and disadvantages of various loss functions, we can see that compared with the original classification loss function, the complexity of softmax loss function based on Angular Margin is reduced, and the performance of the model is improved, realizing the effect of reducing the distance within the class and increasing the distance between classes. However, the distance between difficult samples is too small to satisfy this strong constraint.

In this paper, based on the Angular Margin softmax, the ternary loss group trimming is added. Firstly, the structure of residual network is deeply studied, and then the pre training model is obtained by adding the \(1 \times 1\) convolution layer of pooling sum. Then, the advantages and disadvantages of the Angular Margin loss function are analyzed, and the form of triple function fine-tuning is introduced to transform the optimization process based on Euclidean space to the unit circle. A series of extended methods are used to get face representation by feature fusion.

2.1. Analysis and Discussion on the selection of parameter m

The widely used mathematical expression of softmax loss is as follows:

\[
L = \frac{1}{n} \sum_j L_j = \frac{1}{n} \sum_i -\log \left( \frac{e^{f_{y_i,j}}}{\sum_{y} e^{f_{y,j}}} \right)
\]

(2)

\(f_j\) is the number of training class elements, where \(f\) is the number of training class elements

\[
f_j = W_j^T x_i + b_j
\]

(3)

\(f_{y_i,j} = W_{y_i,j}^T x_i + b_{y_i,j}\)

(4)

Where \(x_i\) and \(y_i\) and \(W_{y_i}\) are the weights of the \(j\) and \(y_i\) columns of the \(i\)-th training sample respectively, which can be obtained by substituting formula (3) (4) into formula (2)

\[
l_i = -\log \left( \frac{e^{f_{y_i,j}}}{\sum_{y} e^{f_{y,j}}} \right)
\]

(5)

The angle between the weight vector \(W_j\) and the feature vector \(X_i\) is \(\theta_{y_i,j}\). By weakening the constraint of value range \(\theta_{y_i,j}\), the definition range of cosine function is generalized to the monotone decreasing angle function as shown below

\[
\mu(\theta_{y_i,j}) = (-1)^k \cos(m \theta_{y_i,j}) - 2k
\]

(6)

When \(m \geq 2\) is applied to the feature manifold classification, there is a higher requirement for feature manifold classification. Assuming the sample feature \(x\) of category 1 and the two column weights \(W_1, W_2\), the classification rule of the loss function is
According to the properties of cosine function in monotone interval, the above formula is equivalent to \( m\theta_1 < \theta_2 \). The size of \( \theta_1, \theta_2 \) on the unit hypersphere is equal to the corresponding arc length \( \omega_1, \omega_2 \). Then the decision boundary of correctly classifying \( x \) into category one is:

\[
\omega_1 < \omega_2
\]

The above formula can be regarded as a circular region on a hyperspherical manifold. The larger the value is, the smaller the circle area is, and the more accurate the distinguishing constraint is. Optimizing the Angular Margin between feature vectors makes the learned features more distinguishable on the hypersphere. With the increase of the value, the angular margin increases, the constraint region on the manifold decreases, and the angular margin constrained by the loss function increases.

The minimum value of \( m \) in ideal feature distribution is that when \( m \) is greater than the minimum value, the maximum angle feature distance within class is limited to be less than the minimum angle feature distance between classes. In the case of binary classification, the weight vectors \( W_1 \) and \( W_2 \) are considered. When \( m \) is greater than or equal to 2, the maximum angle of the first decision boundary is

\[
\frac{\theta_{12}}{m-1} + \frac{\theta_{12}}{m+1} \leq \frac{(m-1)\theta_{12}}{m+1}, \quad \theta_{12} \geq \frac{m-1}{m} \pi
\]

Where \( \theta_{12} \) is the angle between the category weights \( W_1 \) and \( W_2 \). The following constraints are needed to make the maximum within class feature angle distance less than the minimum between class feature angle distance:

\[
\frac{\theta_{12}}{m+1} + \frac{\theta_{12}}{m-1} \leq \frac{(m-1)\theta_{12}}{m+1}, \quad \theta_{12} > \frac{m-1}{m} \pi
\]

By solving the above inequality, the lower bound of \( m \)-value is \( 2 + \sqrt{3} \). In the multi class case, suppose that any class weight \( W_i \) is uniformly distributed in Euclidean space, so there is \( \theta_i^{i+1} = 2\pi/k \), where \( \theta_i^{i+1} \) is the angle between \( W_i \) and \( W_{i+1} \). For type \( i \) \( W_i \), it needs to be restricted

\[
\frac{\theta_{i}^{i+1}}{m+1} + \frac{\theta_{i}^{i+1}}{m-1} \leq \min \left[ \frac{(m-1)\theta_{i}^{i+1}}{m+1}, \frac{(m-1)\theta_{i}^{i+1}}{m+1} \right]
\]

The lower bound of \( m \)-value is 3. Therefore, through the derivation of this algorithm, \( m \) value is approximately selected as the required feature distribution standard.

### 2.2. Data enhancement and network structure design

In the training and testing phase of the network model, a series of data enhancement operations are taken on the image. During training, the face in each training set is translated, zoomed and flipped horizontally to enhance the data set, prevent over fitting and improve the robustness of the algorithm. In the test phase, in order to maintain the consistency with the training phase, the original features and the features obtained by horizontal flipping are fused to improve the expression ability of face features.

Convolution in CNN is based on the operation between convolution kernel and multi-channel feature graph. When we observe an object, we can neither observe a pixel nor see all the whole at once. Convolution can be regarded as the perception of a part, that is, the weighted sum of parts. In this paper, a 1 x 1 convolution kernel is added to the residual network to transfer the pixel information to the next layer, and a larger convolution kernel or pooling method is used to downsampling the feature map. By controlling the convolution kernel, we can reduce or increase the dimension, reduce the model parameters, normalize the features of different sizes, and integrate the cross channel information. MsrA also uses 3x3 convolution xkernel before and after 1 x 1 convolution. Convolution with step size of 2 and 2x2 pooling layer with step size of 2 are added between the residual units of the deep residual network to enhance the robustness of the model, reduce the parameters and prevent certain over fitting. The detailed composition of the final network model is shown in Table 2.
2.3. Loss function design

In the analysis of Section 2.1, this algorithm selects the parameter value of approximately 4 as the required feature distribution standard. However, compared with the decision boundary of the original loss function, the constraint of classification decision is more difficult, and it is difficult to satisfy some difficult samples. In this paper, we use the triple optimization algorithm to solve this problem.

![Figure 2. Network structure design](image)

Table 2. The detailed structure of the network model is given in this paper

| Layer type | Conv1.x | Conv2.x | Conv3.x | Conv4.x | FC |
|------------|---------|---------|---------|---------|----|
| Layer structure | [3×3,64] × 1,52 | 1X1/1 | 1X1/1 | 1X1/1 |
| output | 56×47 | 27×23 | 13×11 | 6×5 | 512 |

The loss function reduces the distance between the anchor points belonging to the same identity and the positive samples, and increases the distance between the anchor points belonging to different identities and the negative samples

$$L_{original} = \sum_{i=1}^{n} [\|x_i^a - x_i^p\|_2^2 - \|x_i^a - x_i^n\|_2^2 + \alpha],$$  \hspace{1cm} (13)

Where $x_i^a, x_i^p, x_i^n$ is the anchor point and positive and negative samples respectively, $\alpha$ is the distance between positive and negative samples. Because the angle of space is based on the angle of space triples, we need to transform the space triples into the angle of space triples. The weights of positive and negative examples are normalized, while the vectors represented by anchor points are not normalized. The positive and negative samples of softmax with Angular Margin are distributed on the unit circle, so the optimization process of this method is shown in Figure 3:

![Figure 3. Improved angle triple loss function optimization](image)

By enhancing the original data, the loss function is used to extract the face features of the original image and the horizontally flipped image, and the corresponding features are fused to get the final face representation. Finally, the asoftmax triplet loss (as-tl) function based on Angular Margin triplet is proposed

$$L_{AS-TL} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{\|x_i^a - x_i^p\|}{\sum_{j=1}^{n} \|x_i^a - x_i^j\|}$$  \hspace{1cm} (14)
Among them, the category weight and bias are normalized, and the category of different sample feature vectors is directly measured by the **Angular Margin** between features. When using triple to fine tune the network, the positive and negative samples do not update the category weight, that is, the partial derivative of the loss function with respect to the category weight is zero. The **Angular Margin** of difficult samples in the feature space distribution is small, and the selection of sample pairs needs some well-designed data mining.

3. Analysis of experimental results
In traditional face recognition based on closed set learning, all the identities in the test set will appear in the training set. In practical application scenarios, the more common situation is the open set recognition in which the identity of the test set is not completely or does not appear in the training set. In this section, LFW and YTF data sets are used as the test set, and CASIA webface published by Chinese Academy of Sciences is used as the training set. At the same time, the overlapped face images in the test and training sets are removed. Firstly, the pre training model is obtained by the improved residual network, then the angle triplet of the model is fine tuned, and finally the face representation is obtained by feature fusion. Compared with other algorithms on LFW and YTF datasets, the advantages and disadvantages of the proposed algorithm are evaluated.

3.1. Preprocessing and network parameter setting
Firstly, mtcnn, the face detection algorithm mentioned in Section 2.2, is used to detect face regions and feature points, and then the cropped face image is obtained by similarity transformation. Then each pixel value ([0,255]) of the image is normalized to the interval of [0,1]. The image of CASIA webface training set with relatively small amount of data is flipped horizontally to expand the data set.

This experiment is carried out in the Caffe framework of GPU acceleration environment. In deep learning, learning rate is a very important super parameter, which controls the update speed of network weights. This section uses the Adam optimizer. Adam optimizer is a gradient descent algorithm, which is used to find the best parameters of the global region. Adam is different from the traditional random gradient descent algorithm. Adam adaptively sets the learning rate to solve the problem of high noise and sparse gradient. Random gradient descent calculates the loss for each data, and then solves the gradient update parameter, which has a fixed learning rate. Finally, Adam with momentum of 0.9 and weight decay rate of 5e-4 is selected for the optimization algorithm in this section.

In the network structure, pooling layer and $1 \times 1$ convolution layer are added to the 20 layer deep residual network. Among them, the initial learning rate of network selection is set to 0.1, and the learning rate decreases with the iterative training. When the epoch reaches 16000 iterations, it is reduced to one tenth of the initial learning rate, and then it is reduced to one tenth of the previous one every 4000 iterations. Finally, it reaches the end of convergence training at 28000 iterations.

| Methods      | training data set size | model number | accuracy /% |
|--------------|------------------------|--------------|-------------|
| DeepFace     | 4M                     | 3            | 97.35       |
| DeepFR       | 2.6M                   | 1            | 98.95       |
| FaceNe       | 200M                   | 1            | 99.65       |
| CenterLoss   | WebFace                | 1            | 99.05       |
| A-Softmax    | WebFace                | 1            | 99.10       |
| AS-TL        | WebFace                | 1            | **99.22**   |

3.2. Accuracy evaluation on data sets
LFW is released by the University of Marseilles, Amsterdam, USA. It is a commonly used test set in the field of face recognition. The pictures mainly come from the natural scenes of life, and there are great differences within individuals due to lighting, posture, expression, occlusion, age and other factors. It is mainly used in the research of face recognition in unrestricted environment. In this dataset, 6000 pairs of faces are randomly selected to form recognition pairs, of which 3000 pairs belong to the same
identity and the remaining 300 pairs belong to different identities. The 10% average accuracy is selected as the model performance evaluation index. Table 3 shows the accuracy comparison of the proposed algorithm (as-tl) on LFW dataset.

The standard deviation of the final algorithm is 0.0054 and the threshold is 0.3050. It can be seen from table 3 that the Angular Margin triple loss function (as-tl) algorithm proposed in this section compresses the feature space, increases the inter class distance and decreases the intra class distance. Compared with the large model small-scale training set algorithm, the recognition rate is 99.22%. Compared with the softmax loss function, the angle triplet loss trimming improves the performance of the algorithm. But the algorithm is not as good as deepid2 + with multi model fusion and facenet with difficult sample mining. In addition to the standard evaluation protocol, there is a more difficult unconstrained large-scale identification benchmark blurfr on LFW dataset. In this protocol, we construct a 10 fold experiment consisting of 156915 positive pairs and 46960863 negative pairs.

The blurfr protocol with more positive and negative pairs is more challenging to the algorithm. There are two evaluation indexes, DIR@FAR=1% It indicates the recognition rate with the false recognition rate far of 1%, VR@FAR=0.1% It indicates the verification rate when the error recognition rate is 0.1%. Because blurfr uses all the face images in the dataset, the evaluation protocol is more strict.

Table 4. accuracy evaluation of using as-tl under YTF dataset standard protocol

| Methods      | training data set size | model number | accuracy /% |
|--------------|------------------------|--------------|-------------|
| DeepFace     | 4M                     | 3            | 91.4        |
| DeepID2+     | 300K                   | 25           | 93.2        |
| FaceNet      | 200M                   | 1            | 95.1        |
| CenterLoss   | WebFace                | 1            | 94.4        |
| A-Softmax    | WebFace                | 1            | 94.26       |
| AS-TL        | WebFace                | 1            | 94.43       |

YTF (YouTube faces) is a video face data set. The shortest video clip is 48 frames, the longest is 6070 frames, and the average length is 181.3 frames. Test with 5000 video pairs on YTF dataset. Table 3 shows the model performance comparison of each algorithm. Because the image quality on YTF dataset is not as good as LFW, the overall recognition rate is lower. However, the performance of this algorithm is better than that of the similar single model small-scale training set algorithm.

4. Conclusion
To solve the problem that difficult samples are difficult to converge in the training process of angle softmax algorithm, in order to reduce the strong constraint of angle softmax on feature space, this paper proposes a recognition algorithm based on angle metric triple loss function (as-tl). In the network structure of the original algorithm, a 1 × 1 convolution layer and a pooling layer are added to normalize different size features, interact and integrate cross channel information. The proposed algorithm is used to fine tune the pre training model. At the same time, the data set is expanded by using data enhancement technology. The original image and horizontal flipped image are sent into the fine tuning network to extract the corresponding features. Compared with similar algorithms, it achieves better performance in large-scale face recognition data set.

Acknowledgments
This paper was supported by the joint development project of China Railway Jinan Bureau Group Co., Ltd. and Beijing Jiaotong University(Research on standardized management of technical data of train diagram(2018Y01)).

References
[1] Anil, K., Arun, A., Karthik. N. (2011) Introduction to biometrics. Springer, New York, USA.
[2] C. Dragoi, T. A. Henzinger, et al. NormFace: L2 Hypersphere Embedding for Face Verification[J]. arXiv preprint arXiv:1704.06369v1, 2017
[3] Q. Cao, L. Shen, W. Xie, et al. Vggface2: a dataset for recognising faces across pose and age. IEEE International Conference on Automatic Face & Gesture Recognition, 2018, 67-74.

[4] Ho, H., Gopalan, R. (2014) Model-driven domain adaptation on product manifolds for unconstrained face recognition. Int. J. Comput. Vis., 109: 110–125.

[5] Hadid, A. (2014) Face biometrics under spoofing attacks: vulnerabilities, countermeasures, open issues, and research directions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 113–118.

[6] Ching, I., Rabello, A., Marcel. S. (2014) Biometrics evaluation under spoofing attacks. IEEE Trans. Inf. Forensics Sec., 9: 2264–2276.

[7] M. Kanti, P. Saha, A. Singha, et al. Enhancement of robustness of face recognition system through reduced gaussianity in Log-ICA. Expert Systems with Applications, Volume 116, 2019, Pages 96-107.

[8] S. Yadav, P. Vishw. Extended interval type-II and kernel based sparse representation method for face recognition. Expert Systems with Applications, Volume 116, 2019, Pages 265-274.

[9] Y. Wu, H. Liu, J. Li, et al. Deep Face Recognition with Center Invariant Loss. Proceedings of the Thematic Workshops of ACM Multimedia 2017, p 408–414.

[10] Cox, I. J., Ghосn, J., Yianilos. P. N. (1996) Feature-based face recognition using mixture-distance. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 209-216.

[11] Turk, M. A., Pentland. A. P. (1991) Face recognition using eigenfaces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 586-591.

[12] Belhumeur, P. N., Hespanha, J. P. Kriegman, D. J. (1996) Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In: European Conference on Computer Vision, pp. 43-58.

[13] Lowe. D. G. (1999) Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision, pp. 1150-1157.

[14] Liu, C., Wechsler. H. (2002) Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image processing, 11: 467-476.

[15] Dalal, N., Triggs. B., (2005) Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 886-893.

[16] Ahonen, T., Hadid, A., Pietikainen. M. (2006) Face description with local binary patterns: application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2037-2041.

[17] Li, Z., Imai, J., Kaneko. M. (2010) Robust face recognition using block-based bag of words. In: International Conference on Pattern Recognition, pp. 1285-1288.

[18] Y. Wen, K. Zhang, Z. Li, et al. A discriminative feature learning approach for deep face recognition. European Conference on Computer Vision, 2016, 499-515.

[19] W. Liu, Y. Wen, Z. Yu, et al. Large-Margin Softmax Loss for Convolutional Neural Networks. Proceedings of The 33rd International Conference on Machine Learning, 2016: 507-516.

[20] Y. Guo, L Zhang. One-shot Face Recognition by Promoting Underrepresented Classes. arXiv preprint arXiv:1707.05574, 2017

[21] Huang, G. B., Ramesh, M., Berg, T., et al. (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report, University of Massachusetts, Amherst, pp. 07-49.

[22] Taigman, Y., Yang, M., Ranzato, M., et al. (2014) Deepface: Closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701-1708.

[23] Sun, Y., Chen, Y., Wang, X., et al. (2014) Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, pp. 1988-1996.

[24] Y. Wen, K. Zhang, Z. Li, et al. A discriminative feature learning approach for deep face recognition. European Conference on Computer Vision, 2016, 499-515.