Classifier shared deep network with multi-hierarchy loss for low resolution face recognition

Jingna Sun, Wenming Yang, Yehu Shen, Qingmin Liao

PII: S0923-5965(19)30221-8
DOI: https://doi.org/10.1016/j.image.2019.115766
Reference: IMAGE 115766

To appear in: *Signal Processing: Image Communication*

Received date : 11 March 2019
Revised date : 21 December 2019
Accepted date : 25 December 2019

Please cite this article as: J. Sun, W. Yang, Y. Shen et al., Classifier shared deep network with multi-hierarchy loss for low resolution face recognition, *Signal Processing: Image Communication* (2019), doi: https://doi.org/10.1016/j.image.2019.115766.

Highlights

- We propose a novel class center-based method to narrow the discriminations of HR domain and LR domain by sharing the classifier, instead of the sample-based method.
- We enforce the losses between HR and LR intermediate layers to reduce their domain gap and adopt max-pooling operation on intermediate features to enhance the performance.
- Experiments on LFW and SCface validate the effectiveness of our method.

Classifier Shared Deep Network with Multi-Hierarchy Loss for Low Resolution Face Recognition

Jingna Sun[a], Wenming Yang[a,*], Yehu Shen[b], Qingmin Liao[a]

[a] *Shenzhen Key Lab. of Info. Sci&Tech/Shenzhen Engineering Lab. of IS&DCP, Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China*

[b] *College of Mechanical Engineering, Suzhou University of Science and Technology, Suzhou, China.*

Abstract

Face images in real Closed-Circuit Television (CCTV) are usually with low resolution, which remarkably deteriorates the performance of existing face recognition algorithms and hinders the application of face recognition. The main technical focus of this issue, matching between high-resolution (HR) and low-resolution (LR) face images has attracted significant attention. In order to better address this problem, we propose a Classifier Shared Deep Network with Multi-Hierarchy Loss (CS-MHL-Net) for low-resolution face recognition (LRFR) in this paper. Firstly, considering that contrastive loss and its variants are not conducive to the convergence of network and the reduction of discrepancy, a shared classifier between HR and LR face images is proposed to further narrow the domain gap between HR and LR by sharing the corresponding weights which can be seen as the class center. Secondly, to fully exploit intermediate features and loss constraints, we embed multi-hierarchy loss into intermediate layers, with the target of reducing the distances between HR and LR intermediate features after max pooling and avoiding the decreasing of accuracy caused by over-utilization of intermediate features. Experimental results on LFW and

☆Fully documented templates are available in the elsarticle package on CTAN.

*Corresponding author

Email addresses: sunjn17@mails.tsinghua.edu.cn (Jingna Sun), yang.wenming@sz.tsinghua.edu.cn (Wenming Yang), yehushen@mail.usts.edu.cn (Yehu Shen), liaoqm@sz.tsinghua.edu.cn (Qingmin Liao)

SCface demonstrate the effectiveness and superiority of the proposed method.

Keywords: low-resolution face recognition; classifier shared deep network; multi-hierarchy loss; intermediate features

1. Introduction

Convolutional neural networks (CNNs) have achieved great success in many fields such as object classification [1, 2], scene understanding [3, 4], and action recognition [5]. Most importantly, CNNs have greatly improved the performance of face recognition [6, 7, 8, 9] in recent years, which laid the foundation for face recognition in real applications. Current accuracy of the-state-of-the-art face recognition algorithms has achieved more than 99% on the LFW database [10]. However, in reality, the qualities of images captured by surveillance videos are severely affected by different image resolutions. The recognition accuracy dropped severely when identifying extremely low-resolution images. In this paper, we will focus on improving the performance of low-resolution face recognition (LRFR) which has made progress and many more [11, 12, 13, 14, 15, 16, 17, 18].

This paper focuses on the matching problem between low-resolution (LR) face images and high-resolution (HR) face images. How to make the network extracting discriminative features of LR face images and narrowing the domain gap between HR and LR are the main directions to improve the performance of LRFR. There are many traditional works [14, 19, 20, 13, 21, 22, 23, 24, 18, 15, 25, 26] making contributions to the improvement of LRFR. Some of these works [14, 13, 21, 19] focus on transforming the LR images to HR images and promoting the recognition accuracy through the reconstructed LR images. The other works [19, 20, 22, 23, 24, 18, 15, 25, 26] pay more attention to the process of extracting the LR features and narrow the distances between LR features and HR features.

In recent years, deep learning has become the mainstream to improve the performance of low-resolution face recognition. There are also many works

optimizing the features of LR images [27, 28, 11, 12, 29, 30, 31, 32]. These works adopt many measures to adjust the LR networks approaching the HR network to promote the recognition accuracy between HR images and LR images. Many classic super-resolution methods [33, 34, 35, 36, 37] focus on transforming LR images to HR images, which can also be applied to improve the quality of LR images. For very low resolution face images, super-resolution methods can not reconstruct the LR images well [18, 34], and the highest downsampling multiple is usually eight. In spite of the limit of super-resolution methods, the ideas of super-resolution methods can be learned to improve the LRFR.

It is worthing to mention that all above deep learning methods that adjusting LR network to approach HR network are inspired by sample-based approaches which are not naturally lead to the reduction of the differences between HR and LR domain. Sample-based method narrows the gap between HR features and LR features intuitively, but the final recognition accuracy of LR images is highly related to the selected strategies of samples. In addition, the sample-based method is local and does not consider the whole class compactness. To improve the defects in sample-based method, in this paper, we propose a class center-based method and make full use of the intermediate feature maps of the network to further narrow the domain gap between HR and LR, which are implemented by classifier shared network and multi-hierarchy loss functions, respectively. Center-based method fixes the centers of LR features and HR features to reduce the differences, which is not dependent on the selecting of samples and can consider the whole optimization of classes. In subsection 3.2.1, we adopt a toy experiment to compare the sample-based method and center-based method intuitively, and in section 4, we present the superiority of center-based method. To further improve the recognition accuracy of LR images, the intermediate features of LR network are applied to narrow the gap between LR features and HR features further.

The overview of CS-MHL-Net proposed in this paper is shown in Fig.1. The main contributions of this paper are as follows:

1) We propose a novel class center-based method to narrow the discrimi-

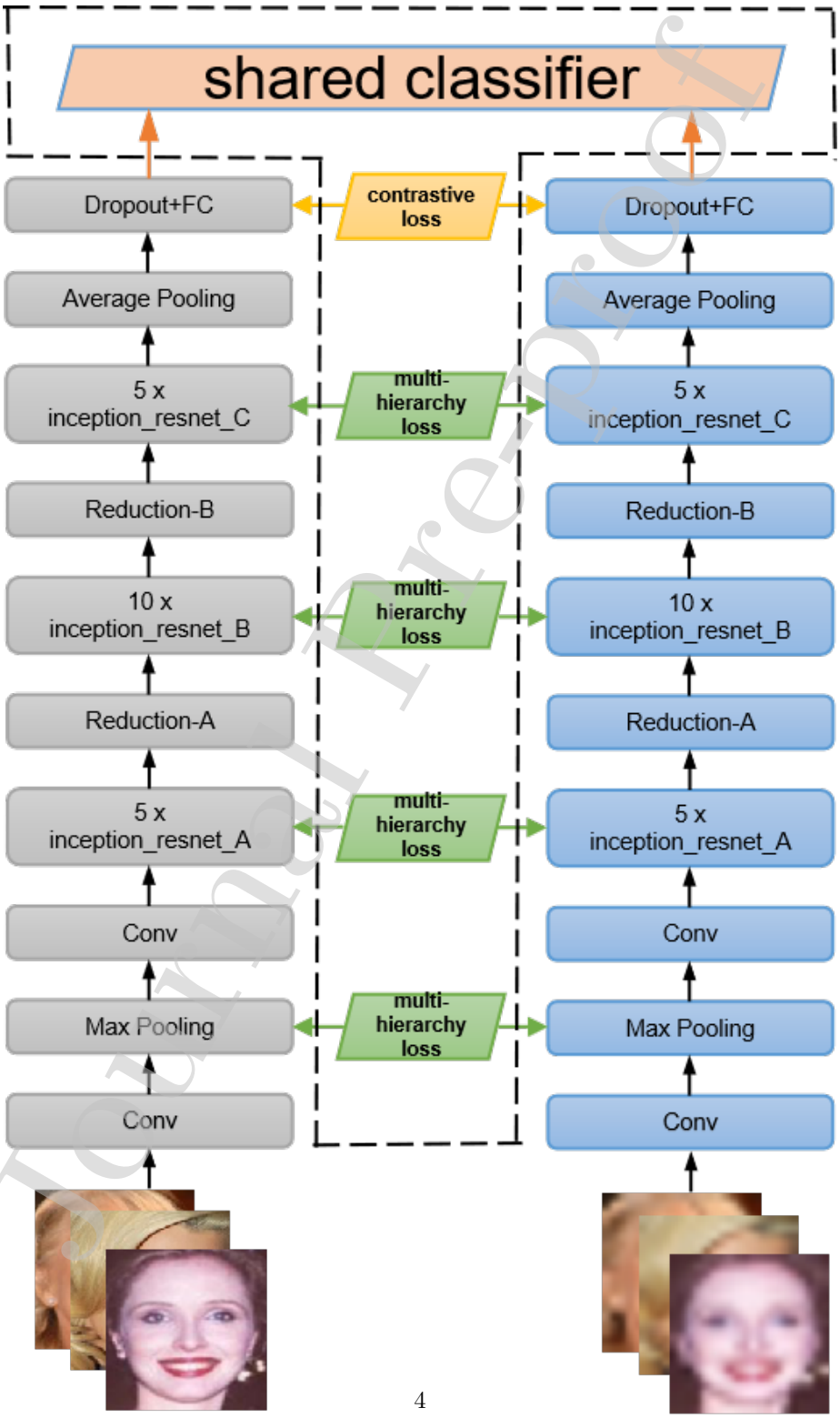


Figure 1: The overview of CS-MHL-Net. The gray color represents fixed parameters, and the blue and light orange color represent what we will fine-tune. HR network and LR network share the same classifier, which is the light orange color. The green and yellow parallelograms represent the intermediate layer which we will apply multi-hierarchy loss functions and the features applied contrastive loss respectively.

nations of HR domain and LR domain by sharing the classifier, instead of the sample-based method, such as contrastive loss [38, 39]. Additionally, we also apply the contrastive loss to the input map of the classifier to further improve the LRFR performance.

2) We enforce the losses between HR and LR intermediate layers to reduce their domain gap and adopt max-pooling operation on intermediate features to enhance the performance. Experimental results show that the optimization of intermediate layers can further improve the accuracy of low-resolution face recognition.

3) Experiments on LFW [10] and SCface [40] validate the effectiveness of our method, and show that our approach is superior to sample-based methods and comparable to state-of-the-art results.

2. Related Work

In this section, we mainly introduce the development of LRFR from three aspects: SR-based methods, Common subspace-based methods and Combined methods.

2.1. SR-based methods

This kind of methods performs super-resolution reconstruction on low-resolution face images and recognizes the reconstructed images [14, 13, 33, 34, 35, 36, 37]. Many traditional and deep learning methods can be applied to improve the quality of LR images.

Traditional methods; [14] proposed to apply the super-resolution reconstruction on all local face blocks to obtain a more detailed super-resolution face image. Wu et al. [13] obtained HR images by the linear reconstruction of low-resolution face images. Deep learning methods; [33] stacked small filters and adopted residual learning to promote the performance of reconstructed images. EDSR [34] improved the residual blocks to better deal with the low-level problems. [35] fuse the attention block to the residual blocks. [36] applied GAN to generate

the HR images and retained identity information at the same time. [37] used face prior geometric knowledge to improve the LR images.

Although SR-based methods improve the visual effects of low-resolution face images, they are not optimal for identification purposes, which limit further improvements. Especially for the very low-resolution images, SR-based methods can not obtain well-reconstructed HR images. The SR-based methods will be better to be auxiliary measures to improve the LRFR.

2.2. Common Subspace-based Methods

This kind of methods maps HR face images and LR face images into a common subspace and reduces the distances between the HR and LR features. We also introduce the works from traditional methods and deep learning methods.

Traditional methods; Zhou et al. [22] mapped HR and LR face images into the same space using a linear mapping matrix. CKE [23] applied a kernel function to the HR and LR face images to obtain the corresponding features. [24] used two linear matrices to map HR and LR images into a common space respectively, and proposed a novel loss function to reduce the distance between high-resolution and low-resolution images belonging to the same category while enlarging the gap between different classes. [18, 15] learned a common space by applying the same mapping matrix on HR and LR face images, and designed a loss function that narrowed the intra-class distance and expanded the inter-class variances skillfully. Deep learning methods; [32] use the heterogeneous joint Bayesian (HJB) classifier to narrow the distance between HR features and LR features. [27] applied the Siamese network to low-resolution face recognition problems. [28] and [11] extracted features by training the HR and LR face images together. [11] additionally enhanced the network's expression ability. Wang et al. [29] adopted a unique parameter sharing method to make high and low-resolution space as close as possible. [31] adopted the improved triplet loss to promote the accuracy further . [30] applied center loss on the LR network and HR network respectively, and then reduce the distance between corresponding HR images and LR images.

Common subspace-based methods usually see the HR network as the teacher and then adjust the LR network. Most of the common subspace-based methods optimize the LR network based on the samples of HR images and LR images, which is local. In the follows, we will introduce a center-based method and optimize the HR images and LR images belonging to the same class globally.

2.3. Combined Methods

There are also many works combine above two methods to further promote the accuracy of LRFR. [20] and [25] proposed a method which first reconstructed the LR face images, and then mapped the reconstructed images into a subspace. GenLR-Net [12] fixed the high-resolution network and fine-tuned the network for LRFR, which was optimized by using the loss function of multiple high-level features. The super-resolution method was also applied to reconstruct the low-level features by calculating the mean squared error (MSE) loss.

In this paper, we focus on the common subspace-based methods and proposed a method to further narrow the distances between HR features and LR features.

3. The Proposed Method

In this section, we will introduce the structure of our networks and demonstrate the motivation of our networks.

3.1. The Pipeline of CS-MHL-Net

The overall network structure is shown in Fig.1, which is called CS-MHL-Net. The left gray network is the HR network with fixed parameters, which is served to extract the features of HR images. The right blue network is called the LR network, which is fine-tuned in the training process and is used to obtain the features of LR images. The network of the two branches is similar to [7] except that we change the stride in block $'MaxPooling'$ from 2 to 1 and is initialized with the same parameters of the pre-trained HR network. In the figure, HR and LR network share the same classifier, and the loss function is Cosface [8], and the

details are shown in subsection 3.2. To further narrow the differences between HR images and LR images, we add multi-hierarchy loss in the intermediate layers, which is colored in green. In subsection 3.3, we represent the form of multi-hierarchy loss. In the last fully connected layer, we add a contrastive loss such as [18, 23] to reduce the distance between HR features and LR features.

3.2. Details on Shared Classifier

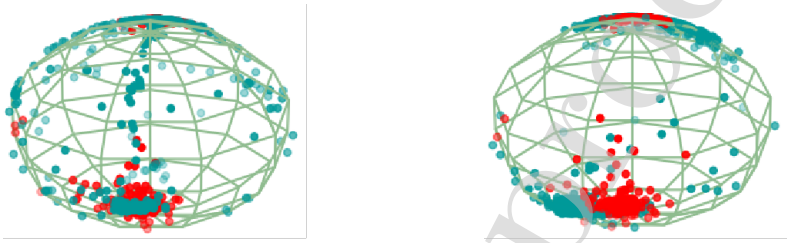
3.2.1. The Motivation of Shared Classifier

In this paper, we apply two identical network structures to extract the features from HR and LR face images, respectively. To narrow the difference between HR and LR features, the methods such as [23, 24, 18, 22, 41, 15, 27] designed the loss functions which reduce the distance between HR and LR face images belonging to the same class and expand the inter-class variances. These approaches are sample-based methods and have a limited effect on reducing the discriminations between HR and LR domain. We propose a class center-based method, to say, fixing the class centers of the HR and LR images so that HR and LR domain gap is further narrowed and the overlapping between HR and LR space is enhanced. The class centers are fixed through sharing the classifier of the two branch network.

(a) HR (b) LR (c) The combine of HR and LR

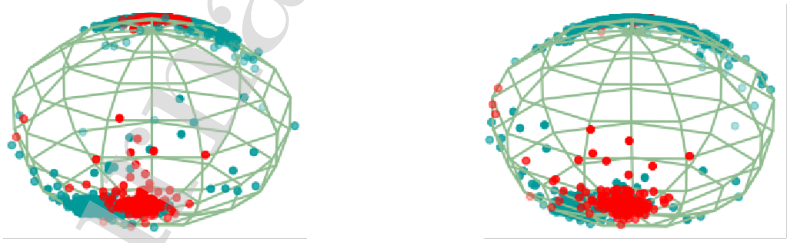
Figure 2: The distribution of HR and LR digital features. We only show the features of digits '0' and '1' to represent clearly.

To verify the superiority of the center-based method, we implement a toy experiment on MNIST [42], which is a handwritten digital dataset and contains



(a) Sample based method (b) Center based method

Figure 3: The training result of sample based method and center based method



(a) Select a batch of samples randomly (b) Select the same batch of LR and HR samples

Figure 4: The impact of selecting the batch of training samples on center based method

Table 1: The digital recognition accuracy of Eq.3 and Eq.2. The accuracy is obtained by testing 10000 images which contain ten classes.

*Eq.*3	*Eq.*2
66.92%	75.77%

Table 2: The accuracy with different training batch selecting methods

*Fig.*4(*a*)	*Fig.*4(*b*)
75.77%	75.88%

60,000 training samples. The size of the original images, 28×28, is taken as the HR digital images and we downsample the HR digital images to 3×3 and then upsample the 3×3 images to 28×28 as the LR digital images. The HR network is trained using Cosface [8] and we set $s = 15$ and $m = 0.1$.

In the following, we only show the features of digits 0 and 1 to better demonstrate the difference of center-based method and sample-based method. In Fig.2, the red dots of (a) shows the distribution of HR digital features. The green dots of (b) represents the LR digital features output from the HR network, which demonstrates the HR network is not suitable to recognize very low-resolution digital images. Fig.2 (c) combines the distribution of HR and LR digital features. Fig.3 shows the HR and LR digital features after training the LR network with the sample-based method and center-based method, respectively. Fig.3(a) represents the LR digital features which is trained with the contrastive loss in Eq.3. In the experiment, we find that selecting the batch of HR and LR digital images randomly leads to the failure of the convergence of LR network, so we train the LR network with the same batch of HR digital images. Fig.3(b) shows the LR digital features trained with Eq.2. From the two subfigures, it can be seen that the LR digital features on the left side are more scattered compared to the right side. In addition, we further list the test accuracy in Table.1. In the experiments, 10000 images which contain ten classes are tested to compare the recognition accuracy, and the results demonstrate the effectiveness of our

center-based method.

In the experiments, we find that the selection of samples has a great impact on the sample-based method. We also experimentally analyze whether the same impact exists in the center-based method. Fig.4 (a) shows the result of using training batch selected randomly, while (b) shows the result of applying the same batch with HR network. Fig.4 (a) and Fig.4 (b) almost show the same results. Table.2 illustrates the accuracy with different training batch selecting methods. The results verify how to select samples in the training process almost does not affect center-based samples.

Based on the above analysis of the sample-based method and center-based method, the center-based method is more easy to carry out and can obtain better recognition accuracy on LR images.

3.2.2. Center based method

In section 3.2.1, we have represented the priority of sharing classifier, which we also called center-based method. In this section, we demonstrate why sharing classifier can be seen as the center-based method and explain the loss function in the shared classifier.

In the experiments, we normalize the classifier's weights and the features outputting from the two branch network. Then the corresponding normalized weight of each class can be approximated as the class center, which can be explained by Eq.1. In Eq.1, it is the derivative of Softmax loss, where C and N represent the number of categories and samples of a batch respectively. $\boldsymbol{w}$ is the weight of the classifier. $\boldsymbol{w}_{y_i}$ and $\boldsymbol{w}_j$ are the $y_i th$ and jth columns. $\boldsymbol{x}_i$ represents the input feature of the classifier and the label of $\boldsymbol{x}_i$ is y_i. s is the scale to make the network easier to converge. $\mathbb{I}(label_(x_i) = y_i)$ is the indicator function, and only when the condition is satisfied, its value is 1. With the form of $\frac{\partial L}{\partial \widetilde{\boldsymbol{w}}_{y_i}}$, we can find that the normalized weight $\widetilde{\boldsymbol{w}}_{y_i}$ is updated by the weighted average of $\boldsymbol{x}_i$. When the cosine distance between $\widetilde{\boldsymbol{w}}_{y_i}$ and $\boldsymbol{x}_i$ is larger, the loss is smaller, so $\widetilde{\boldsymbol{w}}_{y_i}$ can be seen as the ideal target. Because the shared classifier makes the two branch network share the same $\widetilde{\boldsymbol{w}}$, we can fix the same center between LR

images and the corresponding HR images.

$$\frac{\partial L}{\partial \widetilde{\boldsymbol{w}}_{y_i}} = -\frac{1}{N}\sum_{i=1}^{N} s\widetilde{\boldsymbol{x}}_i(1-p_i)\mathbb{I}(label(x_i)=y_i) \tag{1}$$

where

$$p_i = \frac{e^{s\widetilde{\boldsymbol{w}}_{y_i}^T\widetilde{\boldsymbol{x}}_i}}{\sum_{j=1}^{C} e^{s\widetilde{\boldsymbol{w}}_j{}^T\widetilde{\boldsymbol{x}}_i}},\ \widetilde{\boldsymbol{w}}_j = \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|},\ \widetilde{\boldsymbol{x}}_i = \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}\ .$$

In this paper, we feed the features from HR and LR network into the same classifier to make sure that the HR and LR images belonging to the same class share the same class center. The loss function used in the classifier is cosface [8] which is shown in Eq.2. m is the margin parameter and $\widetilde{\boldsymbol{w}}_{y_i}$ can be approximated as the center of the $y_i th$ class.

$$L_{classifier} = -\frac{1}{2N}\sum_{i=1}^{2N} log\frac{e^{s*(\widetilde{\boldsymbol{w}}_{y_i}\widetilde{\boldsymbol{x}}_i - m)}}{e^{s*(\widetilde{\boldsymbol{w}}_{y_i}\widetilde{\boldsymbol{x}}_i - m)} + \sum_{j=1,j\neq y_i}^{C} e^{s*\widetilde{\boldsymbol{w}}_j\widetilde{\boldsymbol{x}}_i}} \tag{2}$$

Sample-based methods such as contrastive loss optimize the LR face features to approach the HR face features with the same class label, which demonstrates that selecting HR and LR samples randomly in the training process is not proper. The form of the training batch decides whether the LR network can converge and how fast the convergence is. Our center-based method means HR and LR face features share the same class center. Therefore, the convergence of LR face features is based on the class center instead of the HR face features with the same label, which is more conducive to the clustering of the HR and LR face features. In Section 4, we will verify the superiority of the center-based method through the experiments on LFW [10].

3.3. Multi-Hierarchy Loss

Based on the above loss function, we add contrastive loss [38, 39] to the features $\boldsymbol{x}_i$ to further optimize the network. Because we use cosine distance to measure the difference between HR and LR features, we normalize the feature $\boldsymbol{x}_i$ to obtain better performance. The equation is shown in Eq.3. $\boldsymbol{x}_i^{HR}$, $\boldsymbol{x}_i^{LR}$ represent the features extracted from HR network and LR network respectively.

α is the parameter determining the minimum distance of the samples belonging to different classes, which is estimated through experiments. When $\boldsymbol{x}_i^{HR}$ and $\boldsymbol{x}_i^{LR}$ belong to the same categories, y_i is 1, otherwise, $y_i = 0$.

$$L_{contrastive} = \frac{1}{N}\sum_{i=1}^{N}(y_i * D_i^2 + (1 - y_i) * max(0, \alpha - D_i^2)) \tag{3}$$

where

$$D_i = \left\| \widetilde{\boldsymbol{x}}_i^{HR} - \widetilde{\boldsymbol{x}}_i^{LR} \right\|_2$$

As is shown in [43] and [44], enforcing the losses between intermediate layers can boost the performance. In our paper, we make full use of HR and LR intermediate features to optimize the network further. The intuitive idea is to use the contrastive loss to optimize the intermediate features. Since the optimization of the intermediate features is not a classification problem, we focus on the distance of HR and LR intermediate features in the same class. The loss function is as follows.

$$L_{intermediate} = \frac{1}{N_{same} * h * w * c}\sum_{i=1}^{N_{same}}(y_i * ||\boldsymbol{f}_i^{HR} - \boldsymbol{f}_i^{LR}||^2) \tag{4}$$

Where $\boldsymbol{f}_i^{HR}$ and $\boldsymbol{f}_i^{LR}$ are the intermediate features of HR and LR networks. N_{same} stands for the number of image pairs from the same categories in a batch. h, w, c are the height, width, and channels of intermediate features.

However, in the experiments, we observe that the recognition accuracy decreases when using Eq.4. The reason may be that the loss function Eq.4 is not suitable for the optimization of intermediate features. We conjecture that intermediate features contain task-specific information, but Eq.4 is a pixel-wise matching, which is not suitable for the network to extract more discriminative features. In this paper, we lose this constraint. First, we assume that the influence of different resolutions is secondary to the identity information in intermediate features. Therefore, we apply max pooling layers to extract the main information of intermediate features, which intend to extract identity information and to ignore the other interference factors. This max-pooling operation

does not increase extra parameters. On the contrary, it reduces the amount of calculation compared to the original operation. It also improves the recognition accuracy in the experimentswhich will be verified in Sec.4. With max-pooling operation, Eq.4 is transformed to Eq.5 which is called multi-hierarchy loss and P represents the operation of max pooling.

$$L_{MHL} = \frac{1}{N_{same} * h * w * c} \sum_{i=1}^{N_{same}} (y_i * ||P(\boldsymbol{f}_i^{HR}) - P(\boldsymbol{f}_i^{LR})||^2) \tag{5}$$

In summary, the loss function we applied in the experiments is a combination of the above three loss functions. The total loss function is shown in Eq.6. λ and β are the parameters obtained from experiments.

$$L = L_{classifier} + \lambda L_{contrastive} + \beta L_{MHL} \tag{6}$$

4. Experiments

The experiments in this paper are conducted on LFW [10] and SCface [40]. CASIA-Webface [45] is the training dataset, which contains about 0.49M face images from 10,575 subjects. Before training, the images are pre-whitened and aligned by MTCNN [46]. The size of the images is 112*96. The corresponding LR images are obtained by downsampling the HR images to the corresponding resolution and then are resized to 112*96. In this paper, the network structure is *inception_resnet_v1* [7, 6, 47]. In order to better adapt to the size 112*96, we set the stride of the *MaxPooling* layer shown in Fig.1 to 1. Eq.2 is the base loss function to train an HR network with the training set CASIA-Webface, and the parameters of pre-trained HR network are loaded into the HR, LR networks and classifier as shown in Fig.1. In the training process of LR network, the batch_size in the experiments is 90, and the initial value of the learning rate is 0.1 which will be reduced to 0.01, 0.001, 0.0001 in step 30k, 70k and 120k respectively.

Figure 5: Examples in SCface. Each row shows the images taken from five cameras.

Table 3: Different Intermediate Features Performance on LFW

MaxPooling	5 x inception_resnet_A	10 x inception_resnet_B	5 x inception_resnet_C
94.583%	94.633%	94.8%	**94.817**%

Table 4: The Accuracy On Four Different Resolutions

	8*8HL	8*8LL	12*12HL	12*12LL	16*16HL	16*16LL	20*20HL	20*20LL	112*96
HR SqueezeNet [48]	51.050%	62.100%	53.933%	62.700%	60.117%	67.367%	70.733%	73.017%	95.500%
LR SqueezeNet [48]+Eq.2	79.450%	86.667%	86.333%	**90.933**%	89.033%	92.800%	90.383%	93.400%	95.500%
LR SqueezeNet [48]+Eq.2+Eq.3	81.117%	86.000%	87.317%	90.600%	90.767%	92.717%	91.067%	**93.767**%	95.500%
LR SqueezeNet [48]+Eq.2+Eq.5	81.017%	**88.117**%	86.717%	90.600%	89.500%	92.717%	90.550%	93.033%	95.500%
LR SqueezeNet [48]+Eq.2+Eq.3+Eq.5	**82.383**%	87.817%	**87.733**%	90.467%	**90.883**%	**92.867**%	**92.000**%	93.500%	95.500%
HR LightCNN [49]	56.833%	67.800%	62.833%	68.300%	73.717%	66.150%	83.333%	72.783%	97.917%
LR LightCNN [49]+Eq.2	84.117%	86.167%	90.350%	90.600%	93.433%	92.800%	94.283%	93.500%	97.917%
LR LightCNN [49]+Eq.2+Eq.3	84.767%	**88.467**%	90.517%	90.883%	93.500%	93.350%	94.950%	94.417%	97.917%
LR LightCNN [49]+Eq.2+Eq.5	84.750%	87.950%	90.417%	91.017%	93.633%	93.283%	95.250%	94.417%	97.917%
LR LightCNN [49]+Eq.2+Eq.3+Eq.5	**84.933**%	88.300%	**90.617**%	**91.167**%	**93.833**%	**94.033**%	**95.350**%	**94.667**%	97.917%
HR ResNet [3]	64.883%	68.500%	73.883%	75.417%	81.983%	80.500%	88.950%	85.850%	98.533%
LR ResNet [3]+Eq.2	85.667%	85.583%	89.800%	88.383%	93.933%	92.133%	96.550%	95.167%	98.533%
LR ResNet [3]+Eq.2+Eq.3	85.750%	85.667%	91.817%	91.083%	95.333%	94.550%	96.717%	95.600%	98.533%
LR ResNet [3]+Eq.2+Eq.5	85.933%	86.133%	92.633%	**92.367**%	95.417%	94.500%	96.700%	**95.867**%	98.533%
LR ResNet [3]+Eq.2+Eq.3+Eq.5	**86.650**%	**86.350**%	**92.867**%	91.967%	**95.433**%	**95.200**%	**96.733**%	95.817%	98.533%
our HR network	54.683%	65.800%	63.467%	67.567%	79.750%	75.083%	90.433%	85.350%	99.100%
our LR network+Eq.3	87.317%	87.433%	93.733%	92.217%	96.317%	94.500%	97.583%	96.183%	99.100%
our LR network+Eq.2	89.017%	90.903%	94.617%	93.733%	96.917%	96.050%	97.933%	97.400%	99.100%
our LR network+Eq.2+Eq.3	89.117%	**93.350**%	94.767%	**94.017**%	97.033%	95.850%	97.983%	97.283%	99.100%
our LR network+Eq.2+Eq.5	89.300%	90.417%	94.817%	93.350%	97.000%	**96.350**%	98.000%	97.300%	99.100%
our LR network+Eq.2+Eq.3+Eq.5	**90.033**%	90.617%	**94.900**%	93.617%	**97.183**%	96.083%	**98.150**%	**97.467**%	99.100%

4.1. Experiments on LFW

LFW [10] contains 13233 images of 5749 subjects. The size of the test set is 6000. In our experiments, we downsample the dataset to four resolutions: 8*8, 12*12, 16*16, 20*20, then resize them to 112*96 using a bicubic interpolation method. We will test the HR vs. LR (HL) face recognition accuracy and LR vs. LR (LL) face recognition accuracy. When testing HL recognition accuracy, we fix one image of the pair and downsample the other. To verify the effectiveness of our proposed loss function for the intermediate features, we use Eq.2 as the loss function of the classifier and the proposed Eq.5 for the intermediate features to test LFW. We take the 12*12 low-resolution face recognition as an example. We use different intermediate layer features applied to Eq.5, and show the experimental results in Table.3. The first row represents the corresponding middle layers. From Table.3, we can observe that the higher-level features can achieve better accuracy. As the amount of loss function of the intermediate features increases, it is difficult to choose reasonable hyperparameters. Therefore, in our experiments, the multi-hierarchy loss is only applied to the last output of block $'5 \times inception\_resnet\_C'$.

In order to verify the effectiveness of our proposed method, we conduct experiments on different networks, such as LightCNN [49], ResNet [3] and SqueezeNet [48]. We also show the effect of the different combination of our proposed loss functions. The specific experimental results are shown in Table.4. HR network represents the high-resolution network which is used for extracting the features of HR images. LR network is the obtained training network and is served for the LR images. $8*8$, $12*12$, $16*16$, $20*20$ and $112*96$ are the corresponding downsampled resolutions. From the table, we can see that a higher recognition rate is achieved by combining the three loss functions. It can also show that the class center-based method is better than the sample-based method, namely, the contrastive loss, especially on lower resolution images (eg.$8*8$). Table.4 verifies the contributions of Eq.5 applied to intermediate features, which boost the performance on the HL recognition. In this table, the LL recognition accuracies also perform better, which shows that

Table 5: Performance of Different Method At SCface

	$d1$	$d2$	$d3$
MDS [50], [51]	60.3%	66.0%	69.5%
RIDN-HJB [32]	57.5%	–	–
DMDS [18]	61.5%	67.2%	62.9%
LDMDS [18]	62.7%	70.7%	65.5%
RICNN [28]	23.0%	66.0%	74.0%
our LR network+Eq.2+Eq.3+Eq.5	**65.5%**	**87.2%**	**98.7%**

our method can make the LR face images of the same category more compact.

4.2. Experiments On SCface

SCface [40] contains 130 subjects, and the images are obtained by five cameras with different qualities. The examples are shown in Fig.5. Each person has three images which are taken at different distances for each camera, namely $d1$ (4.20m), $d2$ (2.60m), and $d3$ (1.00m). These three distances represent different image resolutions. SCface also offers high-resolution mugshot images as gallery images. In the experiments, we use the HR network trained by CASIA-Webface and the 30*30, 20*20 LR network trained by our proposed method to test the images from distances $d3$, $d2$, and $d1$ respectively. Same as [18], we divide the dataset in SCface into two parts, 50 of the 130 subjects as the training set, and the remaining 80 subjects as the test set. Additionally, the training set is selected randomly, and Eq.3 is used as a loss function in the process of fine-tuning. The experimental results are shown in Table.5. We compare the accuracy of our method with other methods, such as MDS [50], [51], DMDS, LDMDS [18]. RICNN [28]. From Table.5, we can see our method achieves the highest accuracy in SCface, which shows the superiority of our method.

5. Conclusions

In this letter, we propose a deep network named CS-MHL-Net to improve the performance of LRFR. The shared classifier is more potent for the cluster of HR and LR face images compared to contrastive loss which we also apply to optimize the network further. The multi-hierarchy loss of intermediate features after max-pooling forces the network to extract more discriminative LR features. Experiments on LFW and SCface demonstrate the effectiveness of our approach.

Acknowledgment

This work was partly supported by the Natural Science Foundation of China (No.61471216, No.61771276 and No.61501451), the National Key Research and Development Program of China (No.2016YFB0101001), and the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (No.JCYJ20170307153940960 and No.JCYJ20170817161845824)

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[2] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

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

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene cnns, arXiv preprint arXiv:1412.6856.

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

[6] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[7] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[8] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, W. Liu, Cosface: Large margin cosine loss for deep face recognition, arXiv preprint arXiv:1801.09414.

[9] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Advances in neural information processing systems, 2014, pp. 1988–1996.

[10] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, in: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition, 2008.

[11] M. Parchami, S. Bashbaghi, E. Granger, Video-based face recognition using ensemble of haar-like deep convolutional neural networks, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE, 2017, pp. 4625–4632.

[12] S. Prasad Mudunuri, S. Sanyal, S. Biswas, Genlr-net: Deep framework for very low resolution face and object recognition with generalization to unseen categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 489–498.

[13] J. Wu, S. Ding, W. Xu, H. Chao, Deep joint face hallucination and recognition, arXiv preprint arXiv:1611.08091.

[14] R. A. Farrugia, C. Guillemot, Face hallucination using linear models of coupled sparse support, IEEE Transactions on Image Processing 26 (9) (2017) 4562–4577.

[15] J. Shi, C. Qi, From local geometry to global structure: Learning latent subspace for low-resolution face image recognition, IEEE Signal Processing Letters 22 (5) (2015) 554–558.

[16] J. Jiang, R. Hu, Z. Wang, Z. Cai, Cdmma: Coupled discriminant multimanifold analysis for matching low-resolution face images, Signal Processing 124 (2016) 162–172.

[17] Y. Chu, T. Ahmad, G. Bebis, L. Zhao, Low-resolution face recognition with single sample per person, Signal Processing 141 (2017) 144–157.

[18] F. Yang, W. Yang, R. Gao, Q. Liao, Discriminative multidimensional scaling for low-resolution face recognition, IEEE Signal Processing Letters 25 (3) (2018) 388–392.

[19] W. W. Zou, P. C. Yuen, Very low resolution face recognition problem, IEEE Transactions on Image Processing 21 (1) (2012) 327–340.

[20] P. H. Hennings-Yeomans, S. Baker, B. V. Kumar, Simultaneous super-resolution and feature extraction for recognition of low-resolution faces, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[21] D. Zhang, J. He, Face super-resolution reconstruction and recognition from low-resolution image sequences, in: Computer Engineering and Technology (ICCET), 2010 2nd International Conference on, Vol. 2, IEEE, 2010, pp. V2–620.

[22] C. Zhou, Z. Zhang, D. Yi, Z. Lei, S. Z. Li, Low-resolution face recognition via simultaneous discriminant analysis.

[23] C.-X. Ren, D.-Q. Dai, H. Yan, Coupled kernel embedding for low-resolution face image recognition, IEEE Transactions on Image Processing 21 (8) (2012) 3770–3783.

[24] S. Siena, V. N. Boddeti, B. V. Kumar, Coupled marginal fisher analysis for low-resolution face recognition, in: European Conference on Computer Vision, Springer, 2012, pp. 240–249.

[25] M. Jian, K.-M. Lam, Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition, IEEE Transactions on Circuits and Systems for Video Technology 25 (11) (2015) 1761–1772.

[26] Y. Peng, L. Spreeuwers, R. Veldhuis, Low-resolution face alignment and recognition using mixed-resolution classifiers, IET biometrics 6 (6) (2017) 418–428.

[27] C. Herrmann, D. Willersinn, J. Beyerer, Low-resolution convolutional neural networks for video face recognition, in: Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on, IEEE, 2016, pp. 221–227.

[28] D. Zeng, H. Chen, Q. Zhao, Towards resolution invariant face recognition in uncontrolled scenarios, in: Biometrics (ICB), 2016 International Conference on, IEEE, 2016, pp. 1–8.

[29] Z. Wang, S. Chang, Y. Yang, D. Liu, T. S. Huang, Studying very low resolution recognition using deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4792–4800.

[30] Z. Lu, X. Jiang, A. Kot, Deep coupled resnet for low-resolution face recognition, IEEE Signal Processing Letters 25 (4) (2018) 526–530.

[31] J. Zha, H. Chao, Tcn: Transferable coupled network for cross-resolution face recognition, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 3302–3306.

[32] D. Zeng, L. Spreeuwers, R. Veldhuis, Q. Zhao, Combined training strategy for low-resolution face recognition with limited application-specific data, IET Image Processing 13 (10) (2019) 1790–1796.

[33] J. Kim, J. Kwon Lee, K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

[34] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136–144.

[35] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 286–301.

[36] C.-C. Hsu, C.-W. Lin, W.-T. Su, G. Cheung, Sigan: Siamese generative adversarial network for identity-preserving face hallucination, arXiv preprint arXiv:1807.08370.

[37] Y. Chen, Y. Tai, X. Liu, C. Shen, J. Yang, Fsrnet: End-to-end learning face super-resolution with facial priors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2492–2501.

[38] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: European Conference on Computer Vision, Springer, 2016, pp. 135–153.

[39] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: null, IEEE, 2006, pp. 1735–1742.

[40] M. Grgic, K. Delac, S. Grgic, Scface–surveillance cameras face database, Multimedia tools and applications 51 (3) (2011) 863–879.

[41] B. Li, H. Chang, S. Shan, X. Chen, Low-resolution face recognition via coupled locality preserving mappings, IEEE Signal processing letters 17 (1) (2010) 20–23.

[42] Y. LeCun, The MNIST database of handwritten digits, http://yann. lecun. com/exdb/mnist/.

[43] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Artificial Intelligence and Statistics, 2015, pp. 562–570.

[44] X. Peng, J. Hoffman, S. X. Yu, K. Saenko, Fine-to-coarse knowledge transfer for low-res image classification, arXiv preprint arXiv:1605.06695.

[45] D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, arXiv preprint arXiv:1411.7923.

[46] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (10) (2016) 1499–1503.

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE, 2015, pp. 1–9.

[48] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer, Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size, arXiv preprint arXiv:1602.07360.

[49] X. Wu, R. He, Z. Sun, T. Tan, A light cnn for deep face representation with noisy labels, IEEE Transactions on Information Forensics and Security 13 (11) (2018) 2884–2896.

[50] S. Biswas, G. Aggarwal, P. J. Flynn, K. W. Bowyer, Pose-robust recognition of low-resolution face images, IEEE transactions on pattern analysis and machine intelligence 35 (12) (2013) 3037–3049.

[51] S. P. Mudunuri, S. Biswas, Low resolution face recognition across variations in pose and illumination, IEEE transactions on pattern analysis and machine intelligence 38 (5) (2016) 1034–1040.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jingna Sun: Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft

Wenming Yang: Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition

Yehu Shen: Writing - Review & Editing, Visualization, Funding acquisition

Qingmin Liao: Writing - Review & Editing