Deep Siamese network for low-resolution face recognition

Shun-Cheung Lai and Kin-Man Lam
Department of Electronic and Information Engineering,
The Hong Kong Polytechnic Univeristy, Hong Kong
E-mail: shun-cheung.lai@connect.polyu.hk, enkmlam@polyu.edu.hk

***Abstract*—Real-world surveillance face images are usually of low-resolution (LR) because the faces are captured at a distance. Matching the LR query faces with high-resolution (HR) gallery faces is still challenging and remains an open problem. The existing face recognition networks fail to extract discriminative features from the LR face images as they never encounter any LR face images during training. One intuitive way to solve the problem is to randomly downsample the training face images to different resolutions for training. This implicitly makes the face recognition network invariant to the resolution change. To better address this problem, we propose to train a face recognition network using a deep Siamese network, which is simple yet effective. Firstly, a shared classifier is used to classify the deep features extracted from HR and LR facial image pairs, explicitly narrowing the domain gap between the HR and LR deep features. Secondly, on top of the deep Siamese network, a new loss function, namely the cross-resolution triplet loss, is used to pull the matching pairs further while pushing the non-matching pairs in the learned feature space. Therefore, the trained network can extract discriminative features across different resolutions. Experiments demonstrate the superiority of our proposed method on a synthetic LR face dataset, LFW, and two real-world LR face datasets, SCface and QMUL-SurvFace.**

I. Introduction

Face recognition (FR) is a well-studied topic. Because of the development of deep learning and the availability of large-scale labeled face datasets [1], [2], [3], [4] and the novel loss functions [5], [6], [7], [8], [9], the state-of-the-art face recognition models have made tremendous improvements on public benchmarks, achieving accuracies of over 99% on LFW [10]. However, those face images used for training and evaluating deep FR models are high-resolution (HR), high-quality web-based face images. The off-the-shelf deep FR models suffer from noticeable performance degradation when applied to real-world face images, e.g., surveillance face images. This is due to the domain discrepancy between the source domain, i.e., those web face images used to train the models and the target domain, i.e., those surveillance face images. Compared to the web face images, real-world face images are contaminated by complex nuisance factors, such as sensor noise, motion blur, bad illumination, etc. Moreover, the real-world surveillance face images are usually of low-resolution (LR) because the faces are captured at a distance. Matching LR query images with a HR gallery set, or ever with a LR gallery set, is known as low-resolution face recognition (LRFR). Due to the increasing popularity of surveillance systems, LRFR in the wild has a wide range of applications and is an urgent issue.

The approaches proposed for LRFR can be generally divided into two categories. One is to use super-resolution (SR) techniques to reconstruct the HR faces from LR query faces by enhancing the image resolution and quality. Then, the reconstructed faces are used for recognition. Face SR was first proposed in [11], which employs Bayesian formulation to estimate the gradient prior from the Gaussian and Laplacian pyramids of the HR training images to reconstruct faces. Another early work [12] uses Principal Component Analysis (PCA) to reconstruct LR faces by the weighted sum of the face images in the training set. Recent work [13] employs Generative Adversarial Networks (GANs) [14] to perform face SR so that it can generate visually appealing super-resolved images for very low-resolution face images. However, these methods are vision-oriented and are not optimized for recognition purposes. Alternatively, identity-preserved face SR should be considered. [15] proposed a framework based on singular value decomposition (SVD) to perform face SR and recognition simultaneously. In [16], [17], [18], different kinds of identity-preserved loss are combined with the pixel-wise loss to perform face SR while preserving the identities of the LR face images so that the super-resolved face images are beneficial for recognition. However, these methods are not feasible for real-world LR face images as they require *paired* LR and HR images for training, where the LR and HR images have pixel-to-pixel correspondence, which is unavailable for real-world LR face images.

The other category is to project the features of LR faces and their HR counterparts into a common subspace, where the feature distance in the common subspace is minimized. Li et al. [19] proposed a method based on coupled mappings, which projects face images of the same person at different resolutions into a unified feature space, where the difference between the LR and HR features is minimized. Lu et al. [20] extended it into a deep learning framework. They proposed a deep coupled ResNet (DCR), whose trunk network is trained by face images of different resolutions, and branch networks are used to transform HR and LR features into a resolution-specific common subspace. In [21], a Resolution-Invariant Deep Network is proposed to learn resolution-invariant features, which can preserve the discriminative information among the face images of different resolutions. Yang et al. [22]

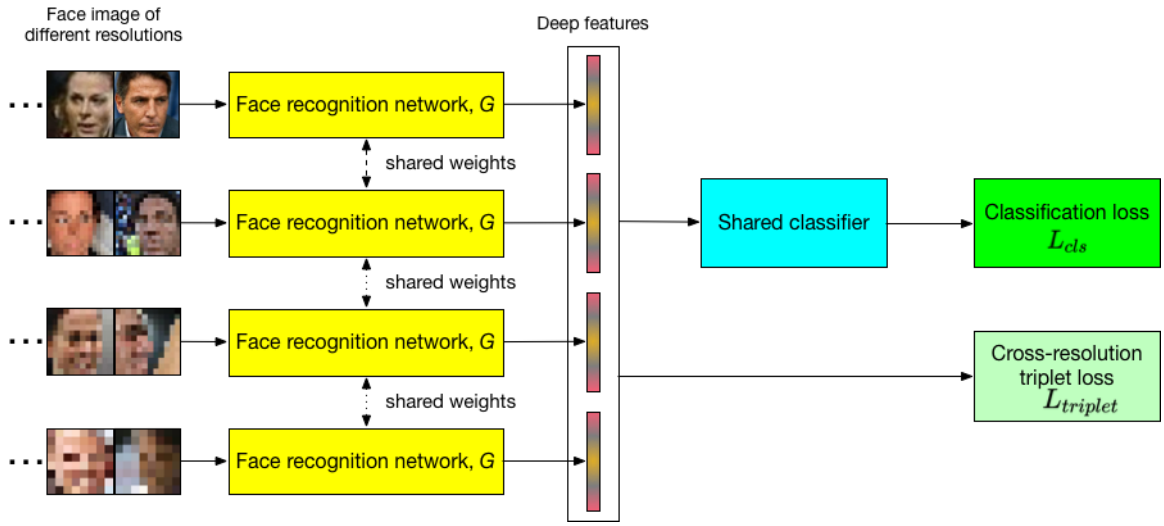

Figure 1. The overview of the proposed deep Siamese network structure. A 4-way face recognition network is used to extract the deep features of face images of sizes 128×128, 16×16, 12×12 and 8×8 pixels. The face recognition networks share the same classifier.

employed a discriminative multidimensional scaling (MDS) method to learn a mapping matrix, which projects LR and HR features into a common subspace. Recently, Zha et al. [23] proposed an online triplet selection method to address the resolution-mismatch problem, which uses a transferrable triple loss to pull the cross-resolution matching pairs and push the non-matching pairs. In general, the subspace-based methods achieve better recognition rates than the SR-based methods because they consider feature extraction and recognition in a unified way.

In this paper, we propose a deep Siamese network to address the LRFR problem. During training, a face recognition network is used to extract the deep features from face images of the same person across different resolutions, and a shared classifier is used to classify the deep features. This can explicitly narrow the domain gap between the LR and HR face images. By using the classification loss L_{cls}, it will minimize intra-class variations while maximizing the inter-class variations across face images of different resolutions. Additionally, a cross-resolution triplet loss $L_{triplet}$ is proposed to effectively pull the matching pairs and push the non-matching pairs across different resolutions. The details of the proposed methods will be presented in the next section. Experiment results show that our proposed method achieves state-of-the-art performance on the LFW [10], SCface [24], and QMUL-SurvFace [25] benchmarks.

II. Proposed Method

To learn discriminative features across different resolutions, the proposed method adopts a deep Siamese network, as shown in Figure 1. It is composed of a K-way face recognition network G for extracting the deep features from HR face images and their LR counterparts. The HR and LR deep features are then fed to a shared classifier, where the additive margin softmax loss (AM-softmax) [7], [8] is employed to obtain the classification loss L_{cls}. Moreover, to further reduce the domain gap between the HR and LR deep features across different resolutions, we propose to train the network with a new loss function, namely the cross-resolution triplet loss $L_{trsiplet}$. The details of the loss functions will be discussed in Section II-B.

A. Network architecture

The face recognition network is a convolutional neural network for learning the deep representations. In our experiments, we use the ResNet architecture adopted from [6], which is shown in Figure 2. It consists of 20 convolutional layers. The kernel size used in the convolutional layer is 3×3, with stride 1 (s1). Downsampling is performed by the 3×3 convolutional layers with stride 2 (s2). Each convolutional layer is followed by the PReLU nonlinear unit [26]. F denotes a fully connected layer. The number of feature maps is indicated on top of each layer, and $\times n$ means a residual connection that repeats n times. This network only accepts images of 128×128 pixels, so all the face images fed to the network are resized to 128×128 pixels.

B. Loss functions

Mathematically, given a set of training face images $\mathbf{X} = \left\{\mathbf{x}_i^{(0)}, \mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(K-1)}, y_i\right\}_{i=1}^{N}$, where $\mathbf{x}_i^{(0)}$ is the i-th HR face image of size 128×128 pixels, $\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(K-1)}$ are the LR face images of the same subject, y_i is the class label of the i-th subject, and N is the mini-batch size. In our experiment, due to the limitation of the GPU memory, we set $K = 4$. This means that one HR branch and three LR branches are used such that $\mathbf{x}_i^{(0)}, \mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \mathbf{x}_i^{(3)}$ represent the training faces of the i-th subject of resolution 128×128, 16×16, 12×12 and 8×8 pixels, respectively. The deep features $\boldsymbol{f}_i^{(k)}$ are obtained

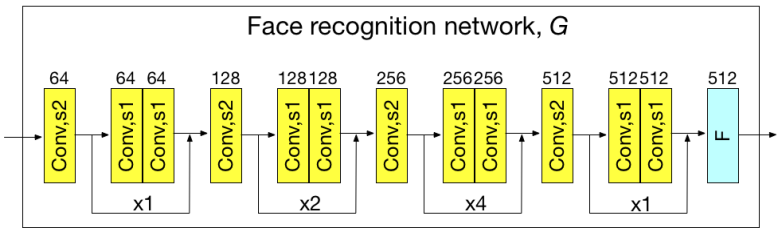

Figure 2. The network architecture of the face recognition network.

by feeding the face images to the face recognition network, such that $\boldsymbol{f}_i^{(k)} = G(\mathbf{x}_i^{(k)})$, where $k = 0, 1, 2, 3$.

To effectively train the face recognition network, cosine-based softmax losses have been widely used as the loss function for the classifier. Here, we consider the additive margin softmax loss (AM-softmax) [7], [8], as the loss function of our classifier, which is shown as follows:

$$L_{AMS}(\boldsymbol{f_i}) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\left(\boldsymbol{w}_{y_i}^T\boldsymbol{f}_i - m\right)}}{e^{s\left(\boldsymbol{w}_{y_i}^T\boldsymbol{f}_i - m\right)} + \sum_{j=1,j\neq y_i}^{C} e^{s\boldsymbol{w}_j^T\boldsymbol{f}_i}} \tag{1}$$

where $\boldsymbol{w_j}$ is the j-th class center vector, C is the total number of classes, and both $\boldsymbol{w_j}$ and $\boldsymbol{f}_i$ are ℓ_2-normalized. s is the scaling factor and m is the margin penalty. The overall classification loss L_{cls} is the average of the AM-softmax loss of the HR and LR features, as follows:

$$L_{cls} = \frac{1}{K}\sum_{k=1}^{K} L_{AMS}(\boldsymbol{f}_i^{(k)}) \tag{2}$$

We empirically found that $L_{AMS}(\boldsymbol{f}_i^{(0)}) < L_{AMS}(\boldsymbol{f}_i^{(1)}) < L_{AMS}(\boldsymbol{f}_i^{(2)}) < L_{AMS}(\boldsymbol{f}_i^{(3)})$. This reflects that the deep feature becomes more discriminative when the resolution is increasing, and there is a domain gap between the HR and LR deep features. To further reduce the domain gap and make the LR deep features more discriminative, we propose the cross-resolution triplet loss $L_{triplet}$, so that the deep features of different resolutions can be matched during training. This loss function is shown as follows:

$$L_{triplet} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\max\left(0, 1 + p_i^{(k)} - n_i^{(k)}\right) \tag{3}$$

where

$$p_i^{(k)} = \max_{l=1\ldots K, l\neq k} d\left(\boldsymbol{f}_i^{(k)}, \boldsymbol{f}_i^{(l)}\right) \tag{4}$$

$$n_i^{(k)} = \min_{l=1\ldots K, j=1\ldots N, j\neq i} d\left(\boldsymbol{f}_i^{(k)}, \boldsymbol{f}_j^{(l)}\right) \tag{5}$$

where $d(\cdot)$ is the cosine distance, which is expressed as follows:

$$d\left(\boldsymbol{f}_i^{(k)}, \boldsymbol{f}_j^{(l)}\right) = 1 - (\boldsymbol{f}_i^{(k)})^T\boldsymbol{f}_j^{(l)} \tag{6}$$

where both $\boldsymbol{f}_i^{(k)}$ and $\boldsymbol{f}_j^{(l)}$ are ℓ_2-normalized. Therefore, $p_i^{(k)}$ are the farthest matching pairs and $n_i^{(k)}$ are the closest non-matching pairs within a mini-batch. Using the cross-resolution triplet loss $L_{triplet}$, we can pull the farthest matching pairs while pushing the closest non-matching pairs across different resolutions. The overall loss function L is the sum of these two losses, as follows:

$$L = L_{cls} + \lambda L_{triplet} \tag{7}$$

where λ is used to balance two losses. We empirically set it to 1.

III. Experiments

The VGGFace2 [4] dataset is used as the training set to train the proposed deep Siamese network. It contains about 3.31M images from 9,131 subjects. All the face images are cropped and aligned based on 5 facial landmarks detected by MTCNN [27]. The scaling factor s and margin penalty m of the classification loss L_{cls} are set to 30 and 0.35, respectively. During training, a HR face of 128×128 pixels is downsampled to 16×16, 12×12, 8×8 pixels, and then upsampled to 128×128 pixels to form the LR inputs. The bilinear kernel is used as the upsampling and downsampling operator. The aligned face images are normalized to $[-1, 1]$, and they are augmented by flipping horizontally with a 50% probability.

The face recognition network and the classifier are trained from scratch. Stochastic gradient descent (SGD) optimizer is used with the weight decay parameter of 5×10^{-4} and momentum of 0.9. N is 128 (i.e., 512 images from 128 subjects in a mini-batch). The learning rate is initialized at 0.1, and it is divided by 10 at 40K, 60K, 80K iterations. Training is finished at 100K iterations. We train the models with the PyTorch [28] library using two GTX 1080TI GPUs. During inference, a single face recognition network is used to extract the deep features of the face images. Cosine similarity is used to measure the similarity of two deep features.

For a fair comparison, we use the same 20-layer ResNet to construct a baseline model, where a 1-way face recognition network is used. Therefore, only the classification loss is used as the loss function, which is the AM-softmax loss. The training face images are randomly downsampled between 8×8 and 128×128 pixels and then upsampled to 128×128 pixels. The rest of the training pipeline is the same as our proposed deep Siamese network. For comparison, we also trained a model, denoted as Ours (w/o $L_{triplet}$), which does not use the cross-resolution triplet loss $L_{triplet}$ in training, i.e.λ is set to 0.

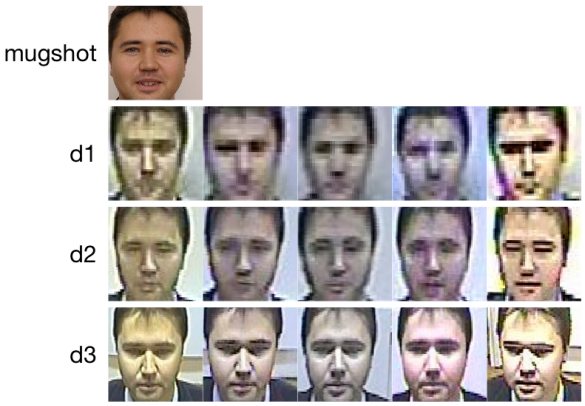

Figure 3. Examples of SCface. It consists of a HR mugshot and LR images taken from 3 distances by 5 cameras.

A. Experiments on LFW

The LFW [10] dataset contains 13,233 images from 5,749 subjects. The face images were captured in uncontrolled environments with variations, such as pose, illumination, and the aging of persons. We follow the unrestricted protocol to report the mean accuracy of 10-fold cross-validation on 6,000 face pairs, where half of the matches are positive while the other half are negative. Same as in [20], [23], we take the first one as HR (128×128 pixels) gallery and the second one as LR query. The query image is synthetically downsampled to 8×8, 12×12, 16×16 and 20×20 pixels, and then upsampled to 128×128 pixels. The face images are cropped and aligned based on 5 facial landmarks, same as the training pipeline. We compare our proposed method to our baseline models, Sun et al. [29], TCN [23], and DCR [20]. The experiment results are shown in Table I. We also tabulate the face-verification results on HR query images (128×128 or 112×96 pixels) as reference. The results demonstrate the effectiveness of the deep Siamese network, as the Ours (w/o $L_{triplet}$) model is a strong baseline. Our proposed method achieves an increase in accuracy by 1.63% on average, compared to our baseline model, and achieves an increase in accuracy by 0.45% on average, compared to the Ours (w/o $L_{triplet}$) model. Furthermore, our proposed method outperforms the best state-of-the-art method, DCR, by 1.48% on average. This illustrates the power of the combination of the deep Siamese network and the cross-resolution triplet loss.

Moreover, we conducted an additional experiment on the LFW benchmark, where the gallery face images and the query face images are downsampled to the same size. This demonstrates a LR-to-LR face recognition problem, which is more challenging. The experiment results are shown in Table II. We can see that our proposed method outperforms the baseline models.

B. Experiments on SCface

The SCface [24] is a real-world surveillance dataset, which contains face images from 130 subjects captured by surveillance cameras at different distances under uncontrolled indoor environments as shown in Figure 3. Same as [22], we consider the daytime data only. For each subject, a digital camera capture a HR mugshot image, and five surveillance cameras capture 15 LR images with with various quality at 3 distances, i.e., 5 images at each of the distances: 4.20m (d1), 2.60m(d2), and 1.00m(d3). Following [22], the HR mugshot images are considered as gallery images, while all the LR images, captured at d1, d2, and d3, are considered as query images. All the images are cropped and aligned based on 5 facial landmarks. Additionally, same as [22], we fine-tune our models by randomly selecting face images from 50 subjects, while the remaining 80 subjects are used for testing. Thus, there is no identity overlap between the training set and the test set. The fine-tuned models are denoted with '-FT'. We also provide the results of the models without fine-tuning, where all the 130 subjects are used for testing. During fine-tuning, the scaling factor s is set at 5 as the number of classes decreases to 50, and the learning rate is set at 1×10^{-5}. Fine-tuning is finished after 10,000 iterations. For our baseline

Table I
VERIFICATION RATES (%) OF DIFFERENT METHODS BASED ON LFW 6,000 PAIRS, FOLLOWING THE LFW UNRESTRICTED SETTING.

Query size→	8×8	12×12	16×16	20×20	128×128 (112×96)
Sun et al. [29]	90.0	94.9	97.2	98.2	(99.1)
DCR [20]	93.6	95.3	96.6	97.3	(98.7)
TCN [23]	90.5	94.7	97.2	97.8	(98.8)
Baseline	90.8	95.4	97.5	**98.5**	**99.4**
Ours (w/o $L_{triplet}$)	94.3	96.9	97.9	97.8	98.9
Ours	**94.8**	**97.6**	**98.2**	98.1	99.1

Table II
VERIFICATION RATES (%) OF DIFFERENT METHODS BASED ON LFW 6,000 PAIRS, FOLLOWING THE LFW UNRESTRICTED SETTING. GALLERY SET AND QUERY SET ARE OF THE SAME RESOLUTION.

Size→	8×8	12×12	16×16	20×20
Baseline	86.9	93.5	95.7	**97.5**
Ours (w/o $L_{triplet}$)	90.3	95.4	97.0	97.1
Ours	**90.8**	**95.9**	**97.4**	**97.5**

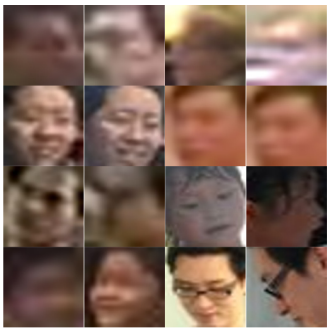

Figure 4. Examples of QMUL-SurvFace.

Table III
RECOGNITION RATE (%) OF DIFFERENT METHODS AT 3 DISTANCES ON THE SCFACE BENCHMARK. '-FT' MEANS PERFORMING FINE-TUNING WITH THE SCFACE TRAINING SET.

Distance→	d1	d2	d3	avg.
MDS [30], [31]	60.3	66.0	69.5	65.3
DMDS [22]	61.5	67.2	62.9	63.9
LDMDS [22]	62.7	70.7	65.5	66.3
RICNN [21]	23.0	66.0	74.0	54.3
FAN [32]	62.0	90.0	94.8	82.3
Baseline	76.3	98.6	99.4	91.4
Ours (w/o $L_{triplet}$)	77.8	96.9	98.5	91.1
Ours	79.7	95.7	98.2	91.2
Sun et al.-FT[29]	65.6	87.2	98.7	83.8
DCR-FT [20]	73.3	93.5	98.0	88.3
TCN-FT [23]	74.6	94.9	98.6	89.4
FAN-FT [32]	77.5	95.0	98.3	90.3
Khali et al.-FT [33]	88.3	98.3	98.6	95.0
Baseline-FT	82.5	98.0	**99.5**	93.3
Ours (w/o $L_{triplet}$)-FT	86.0	97.0	98.0	93.7
Ours-FT	**93.0**	**98.5**	98.5	**96.7**

model, the input faces are randomly selected from the HR mugshot and the LR surveillance images. For the Ours (w/o $L_{triplet}$)-FT and Ours-FT models, $\mathbf{x}_i^{(0)}$ is the HR mugshot, while $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \mathbf{x}_i^{(3)}$ are the LR surveillance images randomly selected at d1, d2, and d3, respectively, of the i-th subject. We compare our results with MDS [30], [31], DMDS [22], LDMDS [22], RICNN [21], Sun et al. [29], DCR [20], and TCN [23]. We also compare them with a face SR-based method FAN [32], and a recently proposed distillation-based method [33].

The face recognition rates are tabulated in Table III. From the results, first, we can see that the Ours (w/o $L_{triplet}$) model is a relatively strong baseline, as it outperforms most other methods, even without performing fine-tuning. This illustrates the effectiveness of the deep Siamese network. With the cross-resolution triplet loss, the performance can be further improved. After fine-tuning, our proposed method outperforms the state-of-the-art methods.

C. Experiments on QMUL-SurvFace

QMUL-SurvFace [25] is a challenging real-world surveillance dataset, as the gallery and query face images are barely visible as shown in Figure 4. It contains 463,507 face images from 15,573 unique identities captured from real-world surveillance videos. We consider the face verification protocol, which contains 10,638 pairs, with half of the matches being positive, and the other half being negative. We compare our method with the state-of-the-art method FAN [32], on the database. As shown in Table IV, our proposed method achieves better True Accept Rates (TARs) at different False Accept Rates (FARs), and also larger Area Under the ROC Curve (AUC). Compared to the Ours (w/o $L_{triplet}$) model, our proposed method improves the performance by a large margin. This can illustrate the effectiveness of the cross-resolution triplet loss, which reduces the domain gap across different resolutions.

Table IV
VERIFICATION RATES (%) OF DIFFERENT METHODS ON QMUL-SURVFACE BENCHMARK.

Method	TAR(%)@FAR				AUC
	30%	10%	1%	0.1%	
FAN [32]	71.30	44.59	12.94	2.75	76.94
Baseline	66.13	37.76	11.64	3.53	74.17
Ours (w/o $L_{triplet}$)	68.23	42.86	14.44	6.28	75.71
Ours	**75.09**	**52.74**	**21.41**	**11.02**	**80.03**

IV. CONCLUSIONS

In this paper, we propose a deep Siamese network to address the low-resolution face recognition (LRFR) problem. Our method uses the Siamese network to extract deep features from face images across different resolutions, and a shared classifier is used to make the deep features of different resolutions compare with the same class center vectors. Additionally, we have proposed a cross-resolution triplet loss to further narrow the domain gap between deep features across different resolutions, which can pull the farthest matching pairs closer and push the closest non-matching pairs farther away. Experiments on the LFW, SCface, and QMUL-SurvFace databases have demonstrated the superiority of our proposed LRFR method, which achieves better performance than the state-of-the-art methods.

REFERENCES

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC)*, 2015.

[2] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv: 1411.7923*, 2014.

[3] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," 2016.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising face across pose and age," 2018.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision (ECCV)*, 2016.

[6] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, July 2018.

[8] H. Wang et al., "Cosface: Large margin cosine loss for deep face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, Oct 2007.

[11] S. Baker and T. Kanade, "Hallucinating faces," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2000.

[12] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35, no. 3, pp. 425–434, Aug 2005.

[13] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] I. Goodfellow et al., "Generative adversarial nets," in *NIPS*, 2014.

[15] M. Jian and 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*, vol. 25, no. 11, pp. 1761–1772, Nov 2015.

[16] K. Zhang et al., "Super-identity convolutional neural network for face hallucination," in *European Conference on Computer Vision (ECCV)*, 2018.

[17] S. C. Lai, C. H. He, and K. M. Lam, "Low-resolution face recognition based on identity-preserved face hallucination," in *IEEE International Conference on Image Processing (ICIP)*, 2019.

[18] C. C. Hsu, C. W. Lin, W. T. Su, and G. Cheung, "Sigan: Siamese generative adversarial network for identity-preserving face hallucination," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6225–6236, Dec 2019.

[19] B. Li, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 20–23, Jan 2010.

[20] Z. Lu, X. Jiang, and A. Kot, "Deep coupled resnet for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 526–530, Feb 2018.

[21] D. Zeng and Q. Zhao, "Towards resolution invariant face recognition in uncontrolled scenarios," in *International Conference on Biometrics (ICB)*, 2016.

[22] F. Yang, W. Yang, R. Gao, and Q. Liao, "Discriminative multidimensional scaling for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 388–392, Mar 2018.

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

[24] M. Grgic, K. Delac, and S. Grgic, "Scface - surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 863–879, Feb 2011.

[25] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," *arXiv preprint arXiv: 1804.09691*, 2018.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[28] A. Paszke, , et al., "Automatic differentiation in pytorch," in *Conference on Neural Information Processing Systems (NIPS) Workshop*, 2017.

[29] J. Sun, Y. Shen, W. Yang, and q. Liao, "Classifier shared deep network with multi-hierarchy loss for low resolution face recognition," *Signal Processing: Image Communication*, vol. 82, Mar 2020.

[30] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 1034–1040, May 2016.

[31] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, "Pose-robust recognition of low-resolution face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 3037–3049, Dec 2013.

[32] X. Yin, Y. Tai, Y. Huang, and X. Liu, "Fan: Feature adaptation network for surveillance face recognition and normalization," *arXiv preprint arXiv: 1911.11680*, 2019.

[33] S. S. Khalid, M. Awais, Z. H. Feng, C. H. Chan, A. Farooq, A. Akbari, and J. Kittler, "Resolution invariant face recognition using a distillation approach," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 410–420, Oct 2020.