Deep Coupled ResNet for Low-Resolution Face Recognition

Ze Lu, Xudong Jiang, *Senior Member, IEEE*, and Alex Kot, *Fellow, IEEE*

***Abstract*—Face images captured by surveillance cameras are often of low resolution (LR), which adversely affects the performance of their matching with high-resolution (HR) gallery images. Existing methods including super resolution, coupled mappings (CMs), multidimensional scaling, and convolutional neural network yield only modest performance. In this letter, we propose the deep coupled ResNet (DCR) model. It consists of one trunk network and two branch networks. The trunk network, trained by face images of three significantly different resolutions, is used to extract discriminative features robust to the resolution change. Two branch networks, trained by HR images and images of the targeted LR, work as resolution-specific CMs to transform HR and corresponding LR features to a space where their difference is minimized. Model parameters of branch networks are optimized using our proposed CM loss function, which considers not only the discriminability of HR and LR features, but also the similarity between them. In order to deal with various possible resolutions of probe images, we train multiple pairs of small branch networks while using the same trunk network. Thorough evaluation on LFW and SCface databases shows that the proposed DCR model achieves consistently and considerably better performance than the state of the arts.**

***Index Terms*—Convolutional neural network (CNN), coupled mappings (CMs), feature extraction.**

I. Introduction

FACE recognition (FR) has been a very active research area due to increasing security demands, commercial applications and law enforcement applications [1]–[3]. Promising results have been achieved under challenging conditions, such as occlusion [4], variations in pose, and illumination [5]. While many FR approaches have been developed for recognizing high-resolution (HR) face images [6]–[8], there are few studies focused on FR in surveillance systems, where HR cameras are not available or there is a long distance between the camera and the subject. Under the condition of low-resolution (LR) images, FR approaches developed for HR images usually decline [2], [9]. It is still a challenge to recognize faces when only LR probe images are available.

Here, we focus on the LR face recognition (LRFR) problem of matching LR probe face images with HR gallery images. Most of the approaches proposed for this task can be generally divided into two categories. One is to reconstruct the HR probe image from the LR one by super-resolution (SR) techniques and use it for classification. Although SR-based methods, such as [10]–[13], can generate visually appealing HR images, they are computationally expensive and not optimized for recognition purposes; thus, the results can be further improved [2], [14]–[16].

The other category is to simultaneously transform the LR probe and corresponding HR gallery images into a common feature subspace where the distance between them is minimized. Li *et al.* [2] propose to learn two matrices that project the face images with different resolutions into a unified feature space, where the difference between the LR image and its HR counterpart is minimized. Based on the idea of linear discriminant analysis, several discriminant subspace methods are proposed in [17]–[19]. Instead of using linear methods, Ren *et al.* [20] project the LR and HR face images into an infinite common subspace by minimizing the dissimilarities captured by kernel Gram matrices. Multidimensional scaling (MDS) is employed in [14] to simultaneously transform the features from the poor quality probe images and the high-quality gallery images in the manner that their distances approximate those between gallery images. The same authors propose a reference-based approach for reducing the computational cost in [21]. Two discriminative MDS methods are proposed in [16] to make full use of identity information, including both interclass and intraclass distances. Their new objective function is claimed to enlarge the interclass distances to ensure discriminability.

In general, subspace-based methods achieve better recognition performance than SR-based methods. However, subspace-based methods usually extract pixel values or scale-invariant feature transform from images as feature representations. Their performance can be boosted by using feature representations that are robust to the resolution change. Motivated by the superior performance of convolutional neural networks (CNN) [22], Zeng *et al.* [15] train a deep convolutional network, resolution-invariance CNN (RICNN), to learn resolution invariant features in a supervised way by mixing the real HR images with the upsampled LR ones. Although RICNN improves the performance of LRFR, it is sensitive to resolution change of probe images as indicated in [16].

In this letter, we propose a CNN-based approach, the deep coupled ResNet (DCR) model, to solve above-mentioned

Manuscript received January 15, 2018; revised February 22, 2018; accepted February 22, 2018. Date of publication February 27, 2018; date of current version March 12, 2018. This work was carried out at the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore, which was supported by the Infocomm Media Development Authority, Singapore. This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 1 RG 123/15. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastien Marcel. *(Corresponding author: Ze Lu.)*

The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: zlu008@ntu.edu.sg; exdjiang@ntu.edu.sg; eackot@ntu.edu.sg).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2810121

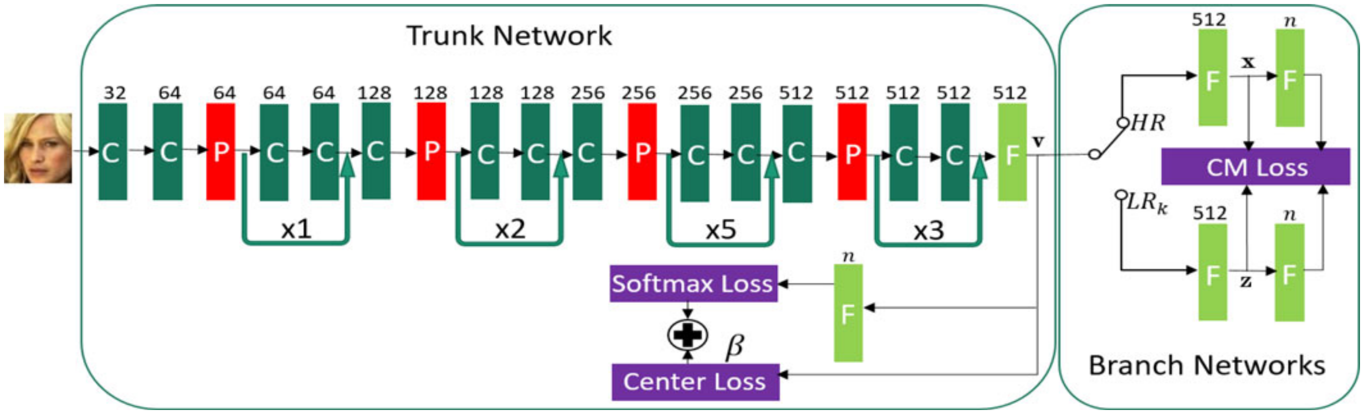

Fig. 1. Architecture of the proposed DCR model. The trunk network learns discriminant features (indicated by $\mathbf{v}$) shared by different resolutions of images, and the branch networks are trained as CMs (indicated by $\mathbf{x}$ for HR features and $\mathbf{z}$ for LR features, respectively). C, P, and F indicate convolutional layer, max-pooling layer, and fully-connected layer, respectively. The number of output feature maps in convolutional layers and the number of outputs in fully connected layers are indicated by those on top of each layer. "$\times h$" represents a residual module that repeats for h times. k indicates the resolution of LR training images and β is a scaling parameter for center loss.

problems in LRFR. The novelty and contribution of the DCR model come from following four aspects.

1) The DCR model consists of one big trunk network and two small branch networks.
2) The trunk network is trained only once to learn discriminant features shared by face images of different resolutions. It is constructed based on recently proposed residential modules [23].
3) Two branch networks are trained to learn resolution-specific coupled-mappings (CMs) so that HR gallery images and LR probe images are projected to a space where their distances are minimized. A CM loss is proposed to optimize model parameters of branch networks.
4) In reality, there can be various resolutions of probe images to be matched with HR gallery images, the proposed DCR model solves this problem by training multiple pairs of small branch networks (2 MB for each pair) while using the same big trunk network (105 MB). Resolution indicator methods [24] can be employed to determine the resolution of probe images and which branch network to be used.

II. DCR Model

A. Architecture of DCR

The key of LRFR is to extract the feature that is robust to resolution change and to measure the similarity between the HR gallery image and the LR probe image. Recently, CNN has been widely used for feature learning in FR and excellent performances have been achieved as in [23] and [25]. This motivates us to employ CNN in learning discriminative features shared by different image resolutions and in deriving CMs to minimize the distance between the HR gallery image and the LR counterpart of a specific resolution. The architecture of the proposed DCR model is shown on Fig. 1. It consists of one trunk network and two branch networks. The trunk network is designed to extract discriminant face features that are robust to the resolution degradation of face images. Two branch networks are trained to learn CMs, which minimize the distance between feature vectors extracted by the trunk branch from an HR image and its corresponding LR image of a specific resolution k.

The trunk network of the proposed DCR model is constructed based on the CNN model in [8] and the ResNet model proposed in [23]. Different from previous CNN architectures such as VGG, ResNet in [23] consists of residual modules, which conduct additive merging of signals. He *et al.* [23] argue that residual connections are inherently important for training very deep architectures. ResNet has become a seminal work, demonstrating that the degradation problem of deep networks can be solved through the use of residual modules. As shown in Fig. 1, the trunk network takes raw pixels of face images as input. C, P, and F indicate convolutional layer, max-pooling layer, and fully-connected layer, respectively. The kernel size of convolution layers is 3×3 with stride 1 and the kernel size of max-pooling layers is 2×2 with stride 2. Each convolutional layer is followed by a PReLU [26] nonlinear unit. The number of output feature maps in convolutional layers and the number of outputs in fully connected layers are indicated by those on top of each layer. "$\times h$" represents a residual module that repeats for h times.

The output feature vector $\mathbf{v}$ of the second last fully-connected layer in the trunk network forms input of the branch networks. The HR feature vector and its LR counterpart are fed into the branch networks in pair. Each branch network consists of two fully-connected layers and one PReLU unit in the middle. Thus, the CMs are formed as two nonlinear mappings. Output feature vectors of the first fully-connected layers ($\mathbf{x}$ and $\mathbf{z}$) are taken as the feature representations for HR and LR images, respectively. The trunk network outputs feature representations that are robust to the resolution change of face images. Its model size is around 105 MB. The branch network reduces the distance between the HR feature and its LR counterpart of a specific resolution. Its model size is around 2 MB, 1.9% of the whole model.

B. Training Strategy and CM Loss

The recently released CASIA-WebFace [27] database is used to train the proposed DCR model. The 434 793 images of 9067

Fig. 2. Example face images from CASIA-WebFace.

subjects, which contain at least 14 images per subject, compose the training set. Face images are normalized to 112 × 96 pixels with an affine transformation according to the coordinates of five sparse facial points, i.e., both eye centers, the nose tip, and both mouth corners. We employ an off-the-shelf face alignment tool [28] for facial point detection. Some example face images are shown in Fig. 2. A two-step training strategy is adopted to effectively optimize parameters of the DCR model. In the first step, the trunk network is trained using face images of three significantly different image resolutions, $112 \times 96, 40 \times 40$, and 6×6, so that it can be applied to the feature extraction of images whose resolution varies from 112×96 to 6×6. Similar to [15], images are first down-sampled to LR images. After that, LR images are rescaled to the required input size of the network using bicubic interpolation before being fed to the network for training. In our case, the required size of input images is 112×96. This process effectively makes the LR images the blurred HR images. The joint supervision of softmax loss and center loss [8] L_t is used to train the trunk network as

$$L_t = -\Sigma_{i=1}^{3m}\log\frac{e^{W_{y_i}^T\mathbf{v}_i+d_{y_i}}}{\Sigma_{j=1}^{n}e^{W_j^T\mathbf{v}_i+d_j}} + \beta\Sigma_{i=1}^{3m}||\mathbf{v}_i-\mathbf{c}_{y_i}^v||_2^2 \qquad (1)$$

where m is the number of training samples of the same resolution and n indicates the number of subjects in the training data. $\mathbf{v}_i$ indicates the feature vector extracted by trunk network from the ith training image. W_j denotes the jth column of the weights W in the last fully-connected layers of trunk networks, and d is the bias item. y_i is the class label for the ith sample and $\mathbf{c}_{y_i}^v$ denotes the y_ith class center of deep features $\mathbf{v}$. β is a scaling factor.

In the second step, model parameters of the trunk network remain unchanged and two branch networks are trained using HR images (112 × 96) and LR images of similar resolution to that of the targeted LR probe images. The CM loss function is proposed to supervise the training of branch networks.

In the training process of branch networks, we first expect to maximize the distances between each face image and its neighbors from different classes within HR or LR features. The softmax loss defined in the following equation facilitates interclass separation:

$$L_s = -\Sigma_{i=1}^{m}\log\frac{e^{U_{y_i}^T\mathbf{x}_i+b_{y_i}}}{\Sigma_{j=1}^{n}e^{U_j^T\mathbf{x}_i+b_j}} - \Sigma_{i=1}^{m}\log\frac{e^{V_{y_i}^T\mathbf{z}_i+a_{y_i}}}{\Sigma_{j=1}^{n}e^{V_j^T\mathbf{z}_i+a_j}} \qquad (2)$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ indicate the feature vectors extracted by branch networks from the ith HR and LR images, respectively. U_j, V_j denote the jth column of the weights U, V in the last fully-connected layers of branch networks, and a, b are the bias items, for HR feature and its LR counterpart, respectively.

In order to preserve intraclass compactness, we minimize the distances between each face image and its neighbors from the same class within HR or LR features. The center loss defined in the following equation minimizes the intraclass variations:

$$L_c = \Sigma_{i=1}^{m}||\mathbf{x}_i - \mathbf{c}_{y_i}^x||_2^2 + \Sigma_{i=1}^{m}||\mathbf{z}_i - \mathbf{c}_{y_i}^z||_2^2 \qquad (3)$$

where $\mathbf{c}_{y_i}^x$ and $\mathbf{c}_{y_i}^z$ denote the y_ith class centers of HR features $\mathbf{x}$ and LR features $\mathbf{z}$, respectively.

For the task of LRFR, it is important to ensure the consistency between the LR feature and corresponding HR feature. Specifically, the final LR and corresponding HR feature vectors should be as close as possible. This can be achieved by minimizing the Euclidean loss defined in

$$L_e = \Sigma_{i=1}^{m}||\mathbf{x}_i - \mathbf{z}_i||_2^2. \qquad (4)$$

Considering the above three criteria, the CM loss used for optimizing parameters of branch networks is constructed as:

$$L_{CM} = L_s + \lambda L_c + \alpha L_e \qquad (5)$$

where λ and α are two scaling factors used for balancing three loss functions. In this way, not only the discriminability of HR and LR features are taken into account, but also the relationship between HR and corresponding LR features.

C. Matching Face Images

Once the training is finished, both HR gallery images and LR probe images can be fed into DCR to obtain their feature representations. Face verification or identification can be performed by computing their similarities. The big trunk network contains 98.1% of parameters of the whole DCR model, whereas small branch networks contain 1.9% of all DCR parameters. In real applications, probe images captured by surveillance cameras can be of many different LRs. Thus, multiple pairs of branch networks are trained in DCR to deal with different resolutions of LR probe images. Hence, multiple branch networks greatly increase the effectiveness and efficiency of the DCR model to match HR images to different resolutions of probe face images.

III. Experiments

Extensive experiments are conducted on LFW [29] and SCface [30] databases to evaluate the proposed DCR model for matching LR probe images with HR gallery face images. Both LFW and SCface databases are widely-used benchmarks for FR in unconstrained environments. On LFW, the face verification performance of DCR is compared with VGGFace [25], LightCNN [31], ResNet [8], and their fine-tuned versions, VGGFace-FT, LightCNN-FT, and ResNet-FT, respectively. We fine-tune the pre-trained models by the same LR training data for the same number of epochs as in the training process of the DCR model. The open-source deep learning toolkit Caffe [32] is utilized to fine-tune the deep models. During fine-tuning, the batch size is set to 128. The models are fine-tuned with descending learning ratios and the fine-tuning stops when the loss does not decrease any more. Moreover, to evaluate the effectiveness of the proposed branch network for other networks, we replace the trunk network with LightCNN and VGGFace, and name the obtained models as Coupled-LightCNN and Coupled-VGGFace, respectively. Furthermore, the performance of the trunk network trained by face images of three different resolutions is also reported for comparison. Different sizes of LR probe images are used for testing. Note that different sizes of

TABLE I
FACE VERIFICATION ACCURACY OF DIFFERENT APPROACHES USING DIFFERENT PROBE SIZES ON LFW

Probe size	8 × 8	12 × 12	16 × 16	20 × 20	112 × 96
LightCNN [31]	67.7	78.3	86.9	92.7	98.9
LightCNN-FT	70.3	79.3	88.9	92.9	98.8
Coupled-LightCNN	80.0	85.1	90.2	93.5	99.0
VGGFace [25]	75.0	82.6	89.3	93.4	97.7
VGGFace-FT	82.3	88.6	92.7	94.8	98.2
Coupled-VGGFace	83.7	88.9	93.1	95.2	98.3
ResNet [8]	72.7	84.1	92.3	95.4	98.7
ResNet-FT	88.9	93.8	95.9	96.8	98.8
Trunk network	92.2	93.6	95.5	96.8	98.4
DCR (coupled-trunk)	93.6	95.3	96.6	97.3	98.7

face images are rescaled to the required input size of the network using bicubic interpolation before being fed to the network for feature extraction. On SCface, besides the above-mentioned nine CNN models, four state-of-the-art LRFR approaches, MDS [14], [21], DMDS, LDMDS [16], and RICNN [15], are also used for comparison with the DCR model. The values of β in (1), λ and α in (5) are set to 0.008 in all experiments.

A. Experiments on LFW

The LFW database contains 13 233 images of 5749 subjects. Images in this database exhibit rich intrapersonal variations of pose, illumination, and expression. LFW has been extensively studied for the research of unconstrained FR in recent years. We follow the "Unrestricted, Labeled Outside Data Results" protocol in [29] and compute the mean verification accuracy by the tenfold cross-validation scheme on the View 2 data. Face images are normalized and aligned using same methods as on CASIA-WebFace images. For two images in the face verification paradigm, we take the first one as HR (112 × 96) gallery image and down-sample the second one to 8 × 8, 12 × 12, 16 × 16, or 20 × 20 as the LR probe image. Same sizes of CASIA-WebFace images are used for training of corresponding branch networks. The same face images are used for the fine-tuning of LightCNN, VGGFace, and ResNet models. PCA and cosine distance are used to calculate the similarity between two features. The total scatter matrix of PCA is computed using the nine training folds of LFW data in the tenfold cross validation. Face verification accuracies of VGGFace [25], VGGFace-FT, Coupled-VGGFace, LightCNN [31], LightCNN-FT, Coupled-LightCNN, ResNet [8], ResNet-FT, the trunk network and the proposed DCR model on LFW are shown in Table I. In the last column, the accuracies for HR probe images of the same resolution as gallery images are also presented.

B. Experiments on SCface

The SCface database contains images of 130 subjects taken in uncontrolled indoor environment using five video surveillance cameras of various qualities. For each subject, there are 15 images taken at three distances (five images at each distance), 4.20 m (*d*1), 2.60 m (*d*2), and 1.00 m (*d*3), by surveillance cameras, and one frontal mugshot image taken by a digital camera. Following the experimental settings in [16], frontal mugshot images are employed as gallery images and images taken by surveillance cameras at distance $di, i = 1, 2, 3$ are used as probe images. We take CASIA-WebFace images of size 112 × 96 as HR images and those of 112 × 96, 30 × 30, and 20 × 20 as LR images for training of branch networks at distance of *d*3, *d*2, and *d*1, respectively. Same as in [16], 50 out of 130 subjects in the SCface database are randomly chosen for fine-tuning of the branch networks and training of PCA. Rest of the subjects are for testing. Thus, there is no identity overlap between the training and test sets. The same face images from CasiaWebface and SCface datasets are used for the fine-tuning of LightCNN, VGGFace, and ResNet models. The nearest-neighbor classifier is used to classify all probe images. We report the FR rates of MDS [14], [21], DMDS, LDMDS [16], RICNN [15], VGGFace [25], VGGFace-FT, Coupled-VGGFace, LightCNN [31], LightCNN-FT, Coupled-LightCNN, ResNet [8], ResNet-FT, the trunk network and the proposed DCR model on Table II.

TABLE II
FR RATES OF DIFFERENT APPROACHES AT DIFFERENT DISTANCES ON SCFACE

Distance	*d*1	*d*2	*d*3
MDS [14], [21]	60.3	66.0	69.5
DMDS [16]	61.5	67.2	62.9
LDMDS [16]	62.7	70.7	65.5
RICNN [15]	23.0	66.0	74.0
LightCNN [31]	35.8	79.0	93.8
LightCNN-FT	49.0	83.8	93.5
Coupled LightCNN	50.5	85.0	94.0
VGGFace [25]	41.3	75.5	88.8
VGGFace-FT	46.3	78.5	91.5
Coupled-VGGFace	62.3	91.0	94.8
ResNet [8]	36.3	81.8	94.3
ResNet-FT	54.8	86.3	95.8
Trunk network	52.0	89.5	96.3
DCR (Coupled-ResNet)	73.3	93.5	98.0

We can observe from Tables I and II that: First, the branch network greatly increases the performance of LightCNN, VGGFace, and the trunk (ResNet) networks; second, the proposed DCR model achieves much higher FR accuracy than the state-of-the-art methods consistently for different resolutions of probe images on LFW and SCface datasets. The performance gain is significant for very LR probe images.

IV. CONCLUSION

In this letter, we propose a novel CNN-based approach named as the DCRN model for the task of LRFR. It first extracts discriminative features shared by face images of different resolutions by a ResNet-like network, the trunk network. After that, CMs are learned by branch networks to project features of HR images and corresponding LR images of a specific resolution into a common subspace where their distance is minimized. Experiments on LFW and SCface datasets show that the proposed DCR model achieves consistently and considerably better performance than the state of the arts.

References

[1] W. Zhang, S. Shan, X. Chen, and W. Gao, “Local Gabor binary patterns based on Kullback–Leibler divergence for partially occluded face recognition,” *IEEE Signal Process. Lett.*, vol. 14, no. 11, pp. 875–878, Nov. 2007.

[2] B. Li, H. Chang, S. Shan, and X. Chen, “Low-resolution face recognition via coupled locality preserving mappings,” *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 20–23, Jan. 2010.

[3] Z. Lu, X. Jiang, and A. C. Kot, “A color channel fusion approach for face recognition,” *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 1839–1843, Nov. 2015.

[4] J. Lai and X. Jiang, “Modular weighted global sparse representation for robust face recognition,” *IEEE Signal Process. Lett.*, vol. 19, no. 9, pp. 571–574, Sep. 2012.

[5] B. Wang, W. Li, W. Yang, and Q. Liao, “Illumination normalization based on Weber’s law with application to face recognition,” *IEEE Signal Process. Lett.*, vol. 18, no. 8, pp. 462–465, Aug. 2011.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[7] C. Ding and D. Tao, “Robust face recognition via multimodal deep face representation,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 499–515.

[9] B. Boom, G. Beumer, L. J. Spreeuwers, and R. N. Veldhuis, “The effect of image resolution on the performance of a face recognition system,” in *Proc. Int. Conf. Control, Autom., Robot., Vis.*, IEEE, 2006, pp. 1–6.

[10] W. W. Zou and P. C. Yuen, “Very low resolution face recognition problem,” *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 327–340, Jan. 2012.

[11] P. H. Hennings, Y. S. Baker, and B. V. Kumar, “Simultaneous super-resolution and feature extraction for recognition of low-resolution faces,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

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

[13] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, “Studying very low resolution recognition using deep networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4792–4800.

[14] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, “Pose-robust recognition of low-resolution face images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 3037–3049, Dec. 2013.

[15] D. Zeng, H. Chen, and Q. Zhao, “Towards resolution invariant face recognition in uncontrolled scenarios,” in *Proc. Int. Conf. Biometrics*, IEEE, 2016, pp. 1–8.

[16] F. Yang, W. Yang, R. Gao, and Q. Liao, “Discriminative multidimensional scaling for low-resolution face recognition,” *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 388–392, Mar. 2018.

[17] C. Zhou, Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, “Low-resolution face recognition via simultaneous discriminant analysis,” in *Proc. Int. Joint Conf. Biometrics*. IEEE, 2011, pp. 1–6.

[18] S. Siena, V. N. Boddeti, and B. V. Kumar, “Coupled marginal fisher analysis for low-resolution face recognition,” in *Proc. Eur. Conf. Computer Vis.*, Springer, 2012, pp. 240–249.

[19] J. Shi and C. Qi, “From local geometry to global structure: Learning latent subspace for low-resolution face image recognition,” *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 554–558, May 2015.

[20] C. Ren, D. Dai, and H. Yan, “Coupled kernel embedding for low-resolution face image recognition,” *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3770–3783, Aug. 2012.

[21] S. P. Mudunuri and S. Biswas, “Low resolution face recognition across variations in pose and illumination,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 1034–1040, May 2016.

[22] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, “Context contrasted feature and gated multi-scale aggregation for scene segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[24] Y. Wong, C. Sanderson, S. Mau, and B. C. Lovell, “Dynamic amelioration of resolution mismatches for local feature based identity inference,” in *Proc. Int. Conf. Pattern Recognit*. IEEE, 2010, pp. 1200–1203.

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proc. Br. Mach. Vis. Conf.*, vol. 1, no. 3, 2015, p. 6.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[27] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” arXiv:1411.7923.

[28] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in *Proc. Eur. Conf. Comput. Vis*. Springer, 2014, pp. 94–108.

[29] 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,” Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[30] M. Grgic, K. Delac, and S. Grgic, “SCface—Surveillance cameras face database,” *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, 2011.

[31] X. Wu, R. He, Z. Sun, and T. Tan, “A light CNN for deep face representation with noisy labels,” arXiv:1511.02683.

[32] Y. Jia *et al.*, “Caffe: Convolutional architecture for fast feature embedding,” in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.