Generate to Adapt: Resolution Adaption Network for Surveillance Face Recognition

Han Fang, Weihong Deng*, Yaoyao Zhong, and Jiani Hu

Beijing University of Posts and Telecommunications
{fanghan, whdeng, zhongyaoyao, jnhu}@bupt.edu.cn

Abstract. Although deep learning techniques have largely improved face recognition, unconstrained surveillance face recognition is still an unsolved challenge, due to the limited training data and the gap of domain distribution. Previous methods mostly match low-resolution and high-resolution faces in different domains, which tend to deteriorate the original feature space in the common recognition scenarios. To avoid this problem, we propose resolution adaption network (RAN) which contains Multi-Resolution Generative Adversarial Networks (MR-GAN) followed by a feature adaption network. MR-GAN learns multi-resolution representations and randomly selects one resolution to generate realistic low-resolution (LR) faces that can avoid the artifacts of down-sampled faces. A novel feature adaption network with translation gate is developed to fuse the discriminative information of LR faces into backbone network, while preserving the discrimination ability of original face representations. The experimental results on IJB-C TinyFace, SCface, QMUL-SurvFace datasets have demonstrated the superiority of our method compared with state-of-the-art surveillance face recognition methods, while showing stable performance on the common recognition scenarios.

Keywords: Surveillance Face Recognition, Generative Adversarial Networks, Feature Adaption

1 Introduction

Surveillance face recognition is an important problem, which is widely existed in the real-world scenarios, *e.g.*, low-quality faces captured from surveillance cameras are used to match low-resolution (LR) faces or high-resolution (HR) faces. The performance on high-resolution testing sets such as LFW [21] has been greatly improved by SOTA face recognition methods [10, 25, 39] and the large-scale datasets [4, 16, 42]. However, due to the large distribution discrepancy between HR and LR faces, the performance of common recognition methods will deteriorate in surveillance face recognition significantly.

Since most of faces in existing datasets [4, 16, 42] are in high-quality, network will focus on learning more informative representation of high resolution such as eyebrows [5], while ignore the information of low-resolution, such as facial

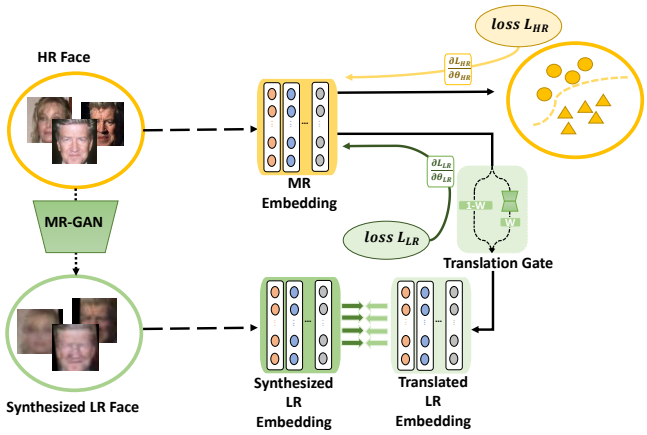


Fig. 1. In RAN, we follow the concept of ***"generate to adapt"***. Unlike [32], MR-GAN is utilized to synthesize realistic resolution-degraded distribution as anchors. Then feature adaption network adopts translation gate to determine the source of translated LR features and minimizes the distance between translated and synthesized LR distribution. The embedding space is directly supervised by HR faces and indirectly supervised by synthesized LR faces, aiming to obtain robust multi-resolution embedding.

contour. When test in the surveillance face, the informative embedding can not catch the lost detail. One intuitive method is to employ face super-resolution [2, 8, 43], and then apply synthesized faces for face recognition. Due to the inevitably introduced noise, the performance will be degraded with this method. The other approach translates the embedding of HR faces and down-sampled LR faces into a unified space to minimize the distance of same identity [18, 44, 54]. However, recent works [3, 23] show down-sampling is not good for scale degradation. In this work, we aim to adopt MR-GAN based data argumentation and propose the progressive training procedure to fuse multi-resolution representations.

We propose a novel resolution adaption network (RAN) which includes multi-resolution generative adversarial networks (MR-GAN) to synthesize realistic LR faces. Feature adaption network is then included to progressively learn the multi-resolution (MR) knowledge. The framework is depicted in Figure 1. Different from [3], which adopted GANs to generate LR images as an intermediate step to achieve image super-resolution, our MR-GAN aims to directly generate realistic LR faces that can be augmented in large-scale datasets and provide prior multi-resolution representations. The global and local mechanism is adopted in generator to pay attention to different areas. In the global stream of generator, input faces are down-sampled into three scales and passed to extract specific knowledge. Then multi-resolution representations are gradually combined and converged into stream of the lowest-resolution to obtain the refined global face by spatial attention. Multi-resolution fusion is conducted by connecting information from sub-encoders of higher-resolution repeatedly and one resolution can

be selected randomly to refine realistic LR faces. Meanwhile, the local regions of lowest-scale face are employed to obtain the refined regions, aggregated with global face to generate realistic LR faces. So the coarse, but still discriminative faces can be employed to provide the low-resolution representations.

Following the concept of generating to adapt, we propose a novel feature adaption network to guide the HR model to fuse the discriminative information of the generated LR faces and maintain steady discrimination ability of the HR faces. So, the problem of domain shift by pulling features of different domains close to each other compulsively can be prevented. Specifically, translation gate is proposed to balance the source of translated embedding and preserve LR representations progressively. To minimize the distance between translated LR embedding and realistic LR embedding extracted by synthesized LR faces, HR model can be guaranteed to contain enough LR information and construct MR embedding, retaining both the information of facial details and contours.

In summary, this paper makes the following contributions:

- We propose multi-resolution GAN to synthesize realistic LR faces, which can avoid the artifacts of down-sampled faces. The representations of different resolutions are combined and injected into stream of the lowest resolution to refine LR faces. And the global and local architectures are both employed into generator and discriminator to reinforce the realism of generated faces.
- We propose feature adaption network to redirect HR model to focus on fusing LR information while preserving HR representations. This network employs translation gate to progressively extract LR knowledge from HR embedding, aiming to ensure that HR model contains enough LR information.
- We select small face from IJB-C [29] and construct testing set named IJB-C TinyFace to exploit unconstrained surveillance face recognition. Our method achieves state-of-the-art performance on surveillance datasets: SCface [15], QMUL-SurvFace [9], and IJB-C TinyFace and shows the stable performance on LFW [21], CALFW [53], CPLFW [52], AgeDB-30 [30] and CFP-FP [34].

2 Related Work

The method we proposed aims to learn and adapt embedding both in HR and LR domains. Therefore, we briefly review previous works from two aspects: common face recognition and surveillance face recognition.

Common Face Recognition. Face recognition [40] is a popular issue in computer vision. The performance has been greatly improved due to the development of discriminative loss functions [10, 25, 33, 39, 49, 51] and deep architectures [17, 19, 20, 35, 36]. And the availability of large-scale datasets, such as CASIA-Webface [42], MS-Celeb-1M [16] and VGGFace2 [4] also contribute to the development of large-scale common face recognition. However, since most of faces in existing datasets are in high-quality, network will focus on learning more informative representations of high resolution, which fails to achieve satisfactory performance on low-resolution face recognition due to the large resolution gap.

Surveillance Face Recognition. There are two categories of method to resolve mismatch between HR and LR faces in surveillance face recognition. The most common studies have concentrated on face super-resolution. These hallucination based methods aim to obtain an identity preserved HR faces from the LR input and use synthesized HR faces for recognition. Bulat et al. proposed Super-FAN [2] to integrate a sub-network for facial landmark localization into a GAN-based super-resolution network. Chen et al. [8] suggested to employ facial prior knowledge, including facial landmark heatmaps and parsing maps to super-resolve LR faces. Zhang et al. [47] proposed a super-identity loss and presented domain integrated training approach to construct robust identity metric. Ataer-Cansizeoglu [1] proposed a framework which contains a super-resolution network and a feature extraction network for low-resolution face verification. The other category of works is to learn projection into a unified space and minimize the distances between LR and HR embedding. Zeng et al. [45] proposed to learn resolution-invariant features to preserve multi-resolution information and classify the identity. Lu et al. [27] proposed the deep coupled ResNet (DCR) model, consisting of one trunk network and two branch networks to extract discriminative features robust to the resolution. Yang et al. employed [41] multi-dimensional scaling method to learn a mapping matrix, projecting the HR and LR images into common space. Ge et al. proposed [13] selective knowledge distillation to selectively distill the most informative facial features from the teacher stream.

3 Methodology

3.1 Framework Overview

Instead of employing down-sampling and bicubic linear interpolation to obtain LR faces [5, 27], our MR-GAN can generate LR faces to avoid artifacts, allowing us to leverage unpaired HR faces, which is crucial for tackling large-scale datasets where paired faces are unavailable. The proposed adaption network is adopted to improve performance on LR faces while still preserve the discrimination ability on HR faces. As shown in Figure 1, our method consists of three steps: (i) Synthesize realistic LR faces; (ii) Employ HR faces and synthesized LR faces as training dataset to train HR and LR model respectively; (iii) Using feature adaption network to re-guide HR model to learn resolution-robust distribution.

3.2 Low-Resolution Face Synthesis

Resolution-aggregated Generator. To minimize the distance between HR and LR domains, we first adopt simple down-sampling to obtain three inputs in three degrees of blur: x_{r_1}, x_{r_2} and x_{r_3}, where x_{r_1} maintains in the highest resolution and x_{r_3} is the face of the lowest resolution. Then we use generator to further refine the global and local information based on down-sampling. Inspired by HRNet [36], we introduce parallel sub-networks to repeatedly receive the information from sub-networks of higher-resolution and integrate the feature

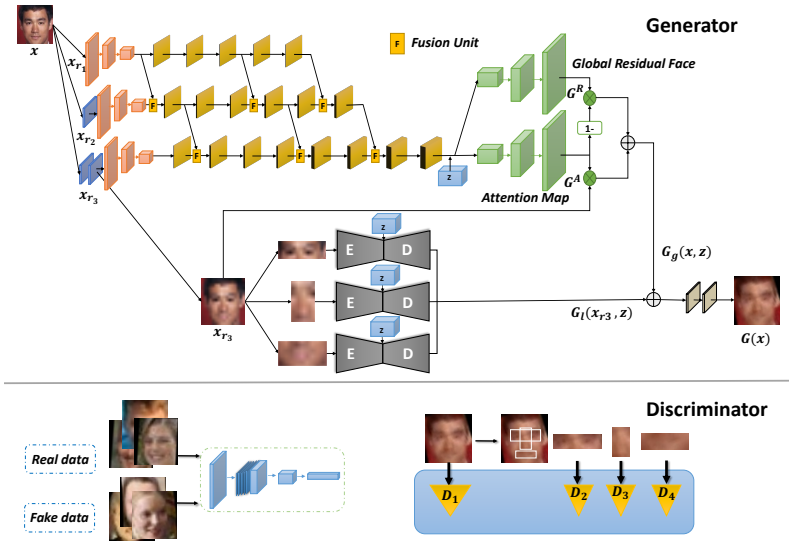


Fig. 2. The architecture of MR-GAN.

map in the global stream. The sub-networks adopt three strided convolutional layers to encode faces into feature maps. Then residual block is used to further deepen the network and make feature maps to maintain the same width and height. To aggregate the information from different streams, fusion units are adopted. We illustrate the details of fusion unit in Figure 3, where all the operated feature maps are learned from residual blocks. The feature maps in the fusion unit can be denoted as $\{\mathbf{F}^1_{r_1}, \mathbf{F}^1_{r_2}, \mathbf{F}^1_{r_3}, ..., \mathbf{F}^k_{r_1}, \mathbf{F}^k_{r_2}, \mathbf{F}^k_{r_3}\}$, where superscript k indicates the feature map from k-th residual block and subscript r shows the feature map in the stream of resolution r. To fuse $\mathbf{F}_r$ from different resolutions, we concatenate two feature maps to deepen the channels. For instance, $F^a_{r_1}$ of $C_1 \times W \times H$ and $F^b_{r_2}$ of $C_2 \times W \times H$ could be integrated to get feature map of $(C_1 + C_2) \times W \times H$. To enhance resolution and identity-relevant informa-

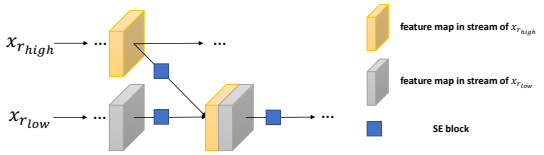


Fig. 3. Illustrate how the fusion unit connects the feature maps from different streams. The representations are selected by SE block before and after connection and flow into the stream of lower-resolution with the deeper channel.

tion in the fusion unit, squeeze-and-excitation (SE) blocks [19] are aggregated before and after feature connection. With repeated feature fusion, the feature maps of higher-resolution are gradually injected into the stream of lower resolution. Meanwhile, since multi-resolution can be preserved at the most extent by connecting feature maps of different streams, we can inject the vector of random noise z to effectively select and simulate different degrees of resolution degradation randomly. To decode the low-resolution information and focus more on the resolution-relevant changes, we introduce spatial attention to ignore the background. So the output of global stream can be summarized as:

$$G_g(x,z) = G^A(x,z) \cdot x_{r_3} + (1 - G^A(x,z)) \cdot G^R(x,z), \tag{1}$$

where $G^R(x,z)$ is the output residual face and $G^A(x,z)$ represents the attention map to describe the contribution to output $G_g(x,z)$. So the important regions can be learned by the generator, and irrelevant pixels could be directly retained from x_{r_3}. The local generator G_l contains three identical sub-networks that learn separately to refine three center-cropped local patches: eyes, nose and mouth. These regions are obtained by the detected landmark and fixed. By passing encoder-decoder stream and injecting random vector z, three cropped local patches can be refined, which are further combined with global face $G_g(x,z)$ and then fed into two 1×1 strided convolutional layers to generate the faces $G(x)$.

Global-local Focused Discriminator. We employ a series of discriminators to distinguish both global and local area, enhancing discrimination ability. Considering characteristics of LR faces, we adopt the same receptive regions as the local branch of generator, consisting of eyes, nose, and mouth to construct local discriminators, while a global discriminator receive the entire face. As shown in Figure 2. These four discriminators ($D_k, k = 1, 2, 3, 4$) pay attention to discriminating different regions respectively. Compared with simple down-sampling and bicubic interpolation, MR-GAN attaches importance to guaranteeing the texture of local region keep fixed and naturally blurred with great visual quality.

3.3 Loss Function

The key objective of our MR-GAN is to generate LR face, while preserving the identity information to avoid artifacts. Several loss terms are proposed to learn realistic representations.

Perceptual Loss. To ensure the generated LR face preserve the same identity as input face, perceptual loss is introduced to reduce the differences in high-dimensional feature space. And the high-level feature representation $\boldsymbol{F}$ are extracted by the pre-trained expert network. So the loss can be formulated as:

$$L_{perceptual} = \sum \|\boldsymbol{F}(x) - \boldsymbol{F}(G(x))\|_1. \tag{2}$$

Adversarial Loss. Adversarial loss is employed for cross domain adaption from source to target distribution. The loss functions are presented as follows:

$$
\begin{aligned}
L_{adv}^{D} &= \sum_{k=1}^{4} \boldsymbol{E}[(D_k(y)-1)^2] + \sum_{k=1}^{4} \boldsymbol{E}[D_k(G(x))^2], \\
L_{adv}^{G} &= \sum_{k=1}^{4} \boldsymbol{E}[(D_k(G(x))-1)^2],
\end{aligned}
\tag{3}
$$

where x is the input HR face and y represents the realistic LR face. Subscript k points the discriminator of corresponding regions. Least square loss is adopted [28] to ensure the discriminator cannot distinguish the synthesized faces.

Pixel Loss. Besides the specially designed adversarial criticism and identity penalty, L_1 loss in the image space is also adopted for further refining the simple down-sampling and bridging the input-output gap, which is defined as follows:

$$
L_{pixel} = \frac{1}{W \times H \times C} \| G(x) - x_{r_3} \|_1. \tag{4}
$$

As shown before, x_{r_3} is the input of the lowest resolution, which can be employed to accelerate the convergence speed and stabilize optimization.

Attention Activation Loss. As shown in Equation 5, when all elements in $G^A(x)$ saturate to 0, all the output is treated as global output. To prevent learning identity-irrelevant information, attention activation loss is adopted to constrain the activation on the important mask and ignore the information around the background. So the loss function can be written as:

$$
L_{att} = \| G^A(x,z)_{center} - 0 \|_1 + \| G^A(x,z)_{edge} - 1 \|_1, \tag{5}
$$

where $G^A(x,z)_{center}$ represents the 85×82 central patch of attention map and $G^A(x,z)_{edge}$ is the edge of attention map.

In summary, we have four loss functions for generating LR face and use hyper-parameters $\lambda_1, \lambda_2, \lambda_3$ and λ_4 to balance them. The overall objective is:

$$
\begin{cases}
L_D = \lambda_1 L_{adv}^{D}, \\
L_G = \lambda_1 L_{adv}^{G} + \lambda_2 L_{perceptual} + \lambda_3 L_{pixel} + \lambda_4 L_{att}.
\end{cases}
\tag{6}
$$

3.4 Feature Adaption Network

Due to the lack of enough LR faces in large-scale datasets, we propose to add generated target samples to balance the multi-resolution representations. However, due to the domain shift between HR and LR domains, it is hard to directly apply the method of simply minimizing distance with the same identities into surveillance face recognition. To overcome this issue, we propose the feature adaption network to preserve the discrimination ability in HR domain and apply it to improve competitiveness in LR domain dynamically.

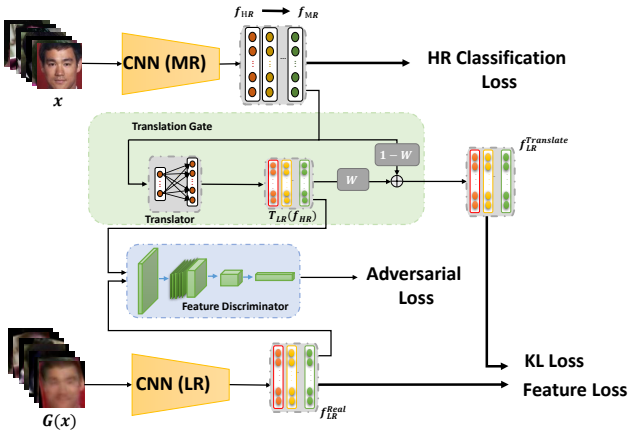


Fig. 4. The pipeline of feature adaption network.

The whole framework is shown in Figure 4, which contains two streams. The stream at the bottom is trained by the generated LR face to offer the realistic LR representation and fixed in the following adaption learning. Stream at the top is used to learn the final multi-resolution embedding f_{MR}. To preserve discriminatory in HR faces, We employ ArcFace [10] as classification loss L_c^{HR}, making the model of top stream directly supervised by HR faces. Meanwhile, to improve the performance on LR face and avoid deteriorating the HR feature space by directly minimizing domain gap, we propose the translation gate. The translate gate employs translator to balance the LR component of f_{HR} and determine the source of $f_{LR}^{Translate}$. The translator consists of two batch normalization, ReLU and fully connected layers in sequence, which plays an intermediate role in amplifying the LR representations to obtain LR features $T_{LR}(f_{HR})$, making HR features f_{HR} focus on preserving LR information. By translating realistic LR features gradually, HR model at the top of stream can preserve more LR representations to obtain the multi-resolution embedding f_{MR}. To achieve this goal, we apply low-resolution adversarial network to ensure that translated LR embedding $T_{LR}(f_{HR})$ is realistic enough to confuse the discriminator. LSGAN [28] is adopted to pull them together. And the loss function can be seen as follows:

$$\begin{aligned} L_{feature}^{D} =& \boldsymbol{E}[(D(f_{LR}^{Real})-1)^2] + \boldsymbol{E}[D(T_{LR}(f_{HR}))^2], \\ L_{feature}^{G} =& \boldsymbol{E}[(D(T_{LR}(f_{HR}))-1)^2]. \end{aligned} \tag{7}$$

By adopting LSGAN, $|D(T_{LR}(f_{HR}))-0.5|$ is used to represent the confidence level of the translated LR features. The closer output of discriminator is to 0.5, the more realistic LR features are translated, representing that f^{HR} can preserve more LR information and obtain f^{MR} with balanced multi-resolution knowledge. With the increase of confidence, f_{HR} can also preserve and provide enough LR representations directly without translation. So, our translation gate adopts a

weighted architecture to determine the final LR features:

$$\begin{cases} W = 1 - |D(T_{LR}(f_{HR})) - 0.5|, \\ f_{LR}^{Translate} = W \cdot T_{LR}(f_{HR}) + (1 - W) \cdot f_{HR}, \end{cases} \tag{8}$$

where W is the weight to balance $T_{LR}(f_{HR})$ and f_{HR}. After obtaining $f_{LR}^{Translate}$, we add L_1 loss and KL loss to learn the low-resolution face distribution in feature and probabilistic representation, further pulling translated embedding close to realistic embedding. The losses can be seen as follows:

$$\begin{aligned} L_f &= \| \frac{f_{LR}^{Translate}(x_{HR})}{\|f_{LR}^{Translate}(x_{HR})\|_2} - \frac{f_{LR}^{Real}(x_{LR})}{\|f_{LR}^{Real}(x_{LR})\|_2} \|, \\ L_p &= \sum p^{Real}(x_{LR}) \cdot \log \frac{p^{Real}(x_{LR})}{p^{Translate}(x_{HR})}. \end{aligned} \tag{9}$$

Considering that f_{HR} contains the limited LR representations in the early stage of training, $T_{LR}(f_{HR})$ plays the dominant role in the feature and probabilistic supervision. Then as HR features can preserve and provide more realistic LR representations gradually, W will maintain within a stable range to balance two sources of low-resolution knowledge. With this weighted translation, f_{HR} can retain enough LR representation to construct resolution-robust embedding. So, total loss can be seen as:

$$L_c = L_c^{HR} + \alpha L_{feature}^{G} + \beta L_p + \gamma L_f. \tag{10}$$

4 Experiments

4.1 Experiment Settings

In this section, we present results for proposed resolution adaption network. CASIA-WebFace [42] is used as HR faces to train both MR-GAN and feature

Table 1. Evaluation results on IJB-C TinyFace 1:1 covariate protocol. Results from row 2 to row 10 are implemented in the same ResNet-34 backbone network.

Method	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
MS1Mv2 (ResNet100 + ArcFace) [10]	0.0300	0.0436	0.1002	0.2191	0.3842	0.5246	0.6948
CASIA-WebFace [42] (ResNet34 + ArcFace)	0.0261	0.0291	0.0420	0.0917	0.1961	0.3219	0.5409
Down-Sampling	0.0434	0.0629	0.1000	0.1486	0.2201	0.3510	0.5853
Cycle-GAN	0.0279	0.0468	0.0897	0.1399	0.2016	0.3065	0.5261
High-to-Low	0.0332	0.0454	0.0638	0.0916	0.1335	0.2113	0.3873
MR-GAN	0.0508	0.0715	0.1159	0.1736	0.2535	0.3861	0.6147
Down-Sampling + Adaption	0.0488	0.0764	0.1168	0.1890	0.2870	0.4452	0.6751
Cycle-GAN + Adaption	0.0524	0.1032	0.1508	0.2058	0.2819	0.4048	0.6254
High-to-Low + Adaption	0.0665	0.0940	0.1428	0.2132	0.2977	0.4281	0.6477
MR-GAN + Adaption (RAN)	**0.0699**	**0.1031**	**0.1616**	**0.2287**	**0.3273**	**0.4817**	**0.7095**

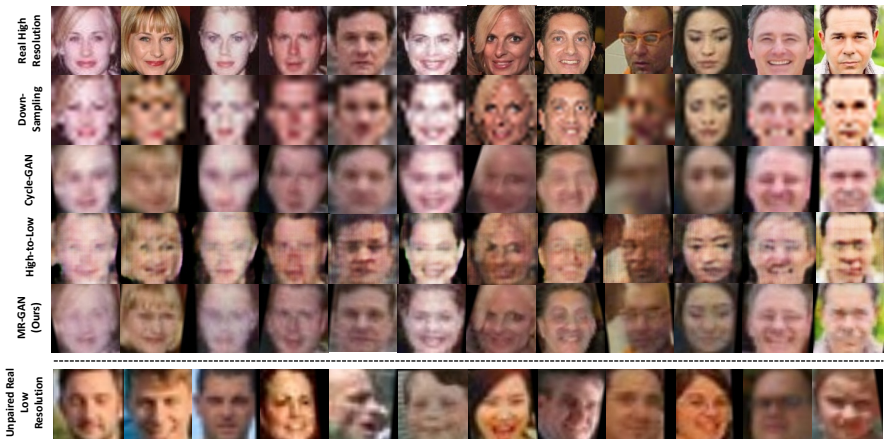
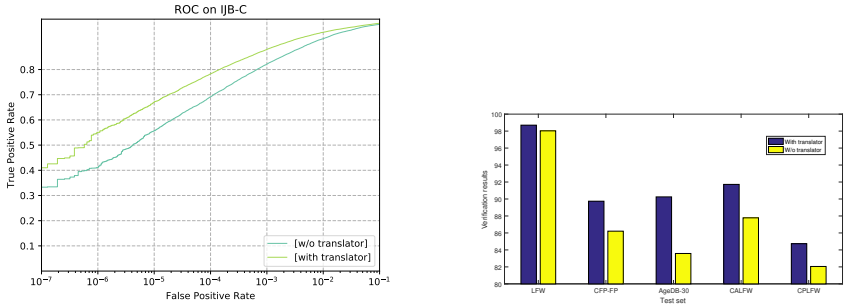


Fig. 5. Face images synthesized by different methods.

adaption network, which contains 494,414 images and 10,575 subjects. The realistic LR faces are selected from MillionCelebs [50]. We use MTCNN [48] for face detection and alignment. The detected landmarks are utilized to measure distance between the center point of eyes and mouth center. Faces whose distances less than 30 and more than 10 are selected as realistic LR faces. To evaluate the performance of feature adaption network, we utilize 34-layer deep residual architecture [17] as backbone and adopt SCface [15], QMUL-SurvFace [9] and low resolution subset of IJB-C [29] (IJB-C TinyFace) as test set. IJB-C [29] is a video-based face database which contains natural resolution variation. We follow the same rule to select realistic LR faces. All the detected LR faces are adopted and faces with same identity are selected for each anchor to construct the positive pairs, including 158,338 genuine comparisons. Following IJB-C 1:1 covariate verification protocol, the same 39,584,639 negative pairs are used in IJB-C TinyFace. SCface [15] consists of face images of 130 subjects. Following [27], 80 subjects are for testing and the other 50 subjects are used for fine-tuning. Face identification is conducted where HR faces are used as the gallery set and LR images captured at 4.2m (d_1), 2.6m (d_2) and 1.0m (d_3) as the probe respectively. QMUL-SurvFace [9] consists of very-low resolution face images captured under surveillance cameras.

4.2 Implementation Details

All the training and testing faces are cropped and aligned into 112×112. In MR-GAN, We train the discriminator and generator by iteratively minimizing the discriminator and generator loss function with Adam optimization. Pixel-critic is employed at every 5 generator iterations. MobileFaceNets [6] has been adopted as the expert network and all the parameters are fixed. The hyper-parameters



(a) (b)

Fig. 6. (*a*). ROC curves depict the effectiveness of translator on IJB-C [29]. (*b*): The comparisons of "with / without translator" on HR domain is depicted to show discriminative recognition ability with translator. The results of LFW [21], CALFW [53], CPLFW [52], AgeDB-30 [30] and CFP-FP [34] are reported.

are empirically set as follows: $\lambda_1 = 2$, $\lambda_2 = 20$, $\lambda_3 = 20$, $\lambda_4 = 0.4$ and batch size= 16. And we set hyper-parameters of the optimizer as follows: $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate= 0.0002. ArcFace is adopted in feature adaption network as the classification loss. Following [10], the feature scale and angular margin m are set as 64 and 0.5 respectively. We set the batch size to 256 to train the pre-trained HR and LR model. There are three steps to obtain the MR embedding. First, we pre-train ResNet-34 by using CASIA-WebFace to obtain the HR model. The learning rate starts from 0.1 and is divided by 10 at 60,000, 100,000 and 140,000 iterations. Second, we fine-tune HR model by adopting the generated LR CASIA-WebFace as training set to get the LR model. And the learning rate starts from 0.01 and is divided by 10 at 50,000 and 100,000 iterations. To simulate more changes of resolution, random Gaussian blur is added when training LR model. Finally, HR model continues to be finetuned by using HR faces with indirect supervision of fixed LR model to train MR model. The batch size is set to 200 in this step and learning rate starts from 0.01, which is divided by 10 at 50,000 iterations. The hyper-parameters can be set as follows: $\alpha = 0.05$, $\beta = 0.04$, $\gamma = 10$. We adopt SGD optimization for recognition and Adam optimization for adversarial learning. $L^G_{feature}$ is updated and utilized at every 4 discriminator iterations. Please refer to the supplementary material for full details on network architectures.

4.3 Ablation Study

Effects of LR Face Synthesis. Since existing large-scale datasets such as CASIA-Webface [42] and MS-Celeb-1M [16] contain a lot of HR faces, our method aims to generate LR faces with different resolutions to augment training set. However, most existing works adopt down-sampling to obtain LR face,

Table 2. Evaluation results on IJB-C TinyFace 1:1 covariate protocol. HR, LR and MR models trained on cleaned MS-Celeb-1M [16] are reported and compared.

Method	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
HR model	0.0307	0.0418	0.0811	0.1801	0.3641	0.5591	0.7491
LR model	0.0643	0.0854	0.1356	0.2240	0.3479	0.5025	0.7033
MR model (RAN)	0.0664	0.1049	0.1678	0.2635	0.4135	0.5819	0.7597

which doesn't match the real environment. As shown in Figure 5, faces generated by down-sampling are full of irregular twist and noise. The GAN-based synthesis method can keep the realism of faces when resolution is reduced. However, the faces generated by Cycle-GAN [55] are over-smoothed. Bulat et al. [3] aimed to adopt High-to-Low and Low-to-High for face super-resolution. They ignored to preserve the information around the facial details and employed the limited supervision in LR faces. So, the LR faces generated by High-to-Low generator can not be used for recognition directly. In contrast, our MR-GAN can integrate multi-resolution information to utilize the specific representation and focus more on the simulation of local region to obtain coarse, but discriminative details. More visualizations can be found in supplementary material.

To quantitatively compare with results on face recognition, we evaluate different methods on IJB-C TinyFace and report the results in Table 1. We translate all the faces of CASIA-WebFace to LR faces including: Down-sampling, Cycle-GAN [55], High-to-Low [3] and MR-GAN, and adopt the generated training set to fine-tune HR model. The results are depicted from row 3 to row 6. And with adaption, the performances are further improved. Since faces generated by High-to-Low [3] are very small which can not be recognized directly, the results are relatively low. However, High-to-Low still provides the coarse enough details during adaption learning, which shows the effectiveness. To better demonstrate the effect of RAN, we report the results of model [10] using larger datasets and more parameters, which is shown at the top. Our method utilizes the smaller model and training set to achieve the same performance and even far beyond them in some cases.

Effects of MR Feature Adaption. To prevent directly minimizing the distances of HR and LR domains due to the domain gap, translation gate is proposed to use translator to balance the source of translated LR features. Without translator, f_{HR} is directly adopted to minimize the distances between different domains. In Figure 6(a), discrimination ability declines fast with the decrease of FAR by directly minimizing distance in feature and probabilistic distribution. In Figure 6(b), the accuracy of LFW decreases to 97.7. However, with intermediate role of translator, translation gate can adopt weighted architecture to generate $T_{LR}(f_{HR})$ progressively. So, the accuracy of LFW can be kept into 98.7. The preserved results on IJB-C and high-resolution testing sets reveal that our MR embedding with translation gate can be adapted into two domains and shows significant effectiveness to handle difficult situations.

Table 3. Rank-1 performance of face identification on SCface testing set. 'w/o FT' means testing with the trained model directly without fine-tuning.

Methods	d_1	d_2	d_3
RICNN [45]	23.00	66.00	74.00
LDMDS [41]	62.70	70.70	65.50
Coupled-ResNet [27]	73.30	93.50	98.00
TCN-ResNet [46]	74.60	94.90	98.60
Selective knowledge distillation [13]	43.50	48.00	53.50
Triplet Loss [26]	70.69	95.42	97.02
Quadruplet Loss [7]	74.00	96.57	98.41
DATL [14]	76.24	96.87	98.09
DAQL [14]	77.25	96.58	98.14
ArcFace [10](w/o FT)	35.00	85.80	98.00
MR-GAN (w/o FT)	65.00	91.50	86.50
RAN (w/o FT)	70.50	96.00	98.00
ArcFace [10]	56.80	91.00	97.50
MR-GAN	71.80	94.30	91.00
RAN	**81.30**	**97.80**	**98.80**

Performance on Large-scale Dataset. To show the effectiveness of our RAN on large-scale datasets, cleaned MS-Celeb-1M [16] which contains 5,084,598 faces and 97,099 subjects is used as training set. ResNet-50 and ArcFace are adopted as the basic training architecture and loss function. Same training steps are employed in this experiment. The results of HR, LR and MR models are depicted in Table 2. Since the large-scale datasets already contain a lot of low-resolution images, only adopting ArcFace loss for supervision can get high performance in HR model. By using MR-GAN to transform all the data set to the LR data set, LR model outperforms HR model where FAR is less than 10^{-4}. Furthermore, our RAN achieves the highest performance in all cases by integrating multi-resolution knowledge.

4.4 Compare with SOTA Methods

Comparisons on SCface. SCface defines face identification protocol. For each subject, there are 15 faces taken at three distances (five faces at each distance) by surveillance cameras, and one frontal mugshot image taken by a digital camera. For fair comparison, we implemented SOTA face recognition method ArcFace [10] as HR model and follow [27] to fine-tune on SCface. The compared methods focus more on minimizing distance of intra-class in different resolutions. However, these methods directly minimize the distance of class, ignoring the resolution gap. And they simply adopt down-sampling to increase the diversity of resolutions and provide paired multi-resolution faces, which don't match the real scenarios. Selective knowledge distillation [13] adopted HR model as teacher and LR model as student to try to restore LR model's ability to discriminate on facial details. Since high resolution information is already lost, sufficient

Table 4. Performance of face identification on QMUL-SurvFace. Most compared results are directly cited from [9] except ArcFace and RAN. In these face super-resolution methods including SRCNN [11], FSRCNN [12], VDSR [22], DRRN [38] and LapSRN [24], SphereFace [25] is used as recognition model.

Methods	TPIR20(%)@FPIR				AUC
	30%	20%	10%	1%	
DeepID2 [37]	12.8	8.1	3.4	0.8	20.8
VggFace [31]	5.1	2.6	0.8	0.1	14.0
FaceNet [33]	12.7	8.1	4.3	1.0	19.8
SphereFace [25]	21.3	15.7	8.3	1.0	28.1
SRCNN [11]	20.0	14.9	6.2	0.6	27.0
FSRCNN [12]	20.0	14.4	6.1	0.7	27.3
VDSR [22]	20.1	14.5	6.1	0.8	27.3
DRRN [38]	20.3	14.9	6.3	0.6	27.5
LapSRN [24]	20.2	14.7	6.3	0.7	27.4
ArcFace [10]	18.7	15.1	10.1	2.0	25.3
RAN	**26.5**	**21.6**	**14.9**	**3.8**	**32.3**

representation cannot be recovered. Instead, our RAN focuses on retaining LR information from HR features through the resolution adaption, which can learn enough multi-resolution knowledge and achieve the best performance.

Comparisons on QMUL-SurvFace. QMUL-SurvFace contains very low LR faces which are drawn from real surveillance videos. We compare our RAN with face super-resolution (SR) methods and common recognition methods. As shown in Table 4, we conduct face identification. Large margin loss (ArcFace and SphereFace) have achieved the SOTA results in large-scale datasets. So, they improve the performance in HR domain, and also can be applied to LR domain. However, these face SR methods struggle to recover the identity information and focus more on the visual quality, inevitably degrading performance. By dynamically extracting MR knowledge in feature space from HR face, our method can perform better than face SR and common recognition methods.

5 Conclusion

This paper proposes Resolution Adaption Network (RAN) for realistic LR face synthesis and surveillance face recognition. We aim to generate LR faces for data augmentation and bridge the cross-resolution gap. In RAN, MR-GAN employs multi-resolution and global-local architecture, blurring face in random resolutions, to generate the identity-preserved and realistic LR faces. To use LR faces to better match with both LR faces and HR faces, feature adaption network is proposed to enhance LR knowledge and balance multi-resolution representations progressively. SOTA results are achieved for surveillance face recognition.

References

1. Ataer-Cansizoglu, E., Jones, M., Zhang, Z., Sullivan, A.: Verification of very low-resolution faces using an identity-preserving deep face super-resolution network. arXiv preprint arXiv:1903.10974 (2019)
2. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 109–117 (2018)
3. Bulat, A., Yang, J., Tzimiropoulos, G.: To learn image super-resolution, use a gan to learn how to do image degradation first. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 185–200 (2018)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
5. Chaitanya Mynepalli, S., Hu, P., Ramanan, D.: Recognizing tiny faces. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
6. Chen, S., Liu, Y., Gao, X., Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition. pp. 428–438. Springer (2018)
7. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 403–412 (2017)
8. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2492–2501 (2018)
9. Cheng, Z., Zhu, X., Gong, S.: Surveillance face recognition challenge. arXiv preprint arXiv:1804.09691 (2018)
10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
11. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. pp. 184–199. Springer (2014)
12. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European conference on computer vision. pp. 391–407. Springer (2016)
13. Ge, S., Zhao, S., Li, C., Li, J.: Low-resolution face recognition in the wild via selective knowledge distillation. IEEE Transactions on Image Processing **28**(4), 2051–2062 (2018)
14. Ghosh, S., Singh, R., Vatsa, M.: On learning density aware embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4884–4892 (2019)
15. Grgic, M., Delac, K., Grgic, S.: Scface–surveillance cameras face database. Multimedia tools and applications **51**(3), 863–879 (2011)
16. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision. pp. 87–102. Springer (2016)

17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Hennings-Yeomans, P.H., Baker, S., Kumar, B.V.: Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
20. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
21. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)
22. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
23. Kumar, A., Chellappa, R.: Landmark detection in low resolution faces with semi-supervised learning. arXiv preprint arXiv:1907.13255 (2019)
24. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017)
25. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
26. Liu, X., Song, L., Wu, X., Tan, T.: Transferring deep representation for nir-vis heterogeneous face recognition. In: 2016 International Conference on Biometrics (ICB). pp. 1–8. IEEE (2016)
27. Lu, Z., Jiang, X., Kot, A.: Deep coupled resnet for low-resolution face recognition. IEEE Signal Processing Letters **25**(4), 526–530 (2018)
28. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802 (2017)
29. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB). pp. 158–165. IEEE (2018)
30. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–59 (2017)
31. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: bmvc. vol. 1, p. 6 (2015)
32. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8503–8512 (2018)
33. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)

34. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. arXiv preprint arXiv:1902.09212 (2019)
37. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems. pp. 1988–1996 (2014)
38. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3147–3155 (2017)
39. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
40. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (2018)
41. Yang, F., Yang, W., Gao, R., Liao, Q.: Discriminative multidimensional scaling for low-resolution face recognition. IEEE Signal Processing Letters **25**(3), 388–392 (2017)
42. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
43. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 908–917 (2018)
44. Zangeneh, E., Rahmati, M., Mohsenzadeh, Y.: Low resolution face recognition using a two-branch deep convolutional neural network architecture. Expert Systems with Applications **139**, 112854 (2020)
45. Zeng, D., Chen, H., Zhao, Q.: Towards resolution invariant face recognition in uncontrolled scenarios. In: 2016 International Conference on Biometrics (ICB). pp. 1–8. IEEE (2016)
46. Zha, J., Chao, H.: Tcn: Transferable coupled network for cross-resolution face recognition. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3302–3306. IEEE (2019)
47. Zhang, K., Zhang, Z., Cheng, C.W., Hsu, W.H., Qiao, Y., Liu, W., Zhang, T.: Super-identity convolutional neural network for face hallucination. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 183–198 (2018)
48. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
49. Zhang, X., Zhao, R., Qiao, Y., Wang, X., Li, H.: Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10823–10832 (2019)
50. Zhang, Y., Deng, W., Wang, M., Hu, J., Li, X., Zhao, D., Wen, D.: Global-local gcn: Large-scale label noise cleansing for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7731–7740 (2020)
51. Zhao, K., Xu, J., Cheng, M.M.: Regularface: Deep face recognition via exclusive regularization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1136–1144 (2019)

52. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying crosspose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Tech. Rep pp. 18–01 (2018)
53. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)
54. Zhou, C., Zhang, Z., Yi, D., Lei, Z., Li, S.Z.: Low-resolution face recognition via simultaneous discriminant analysis. In: 2011 International Joint Conference on Biometrics (IJCB). pp. 1–6. IEEE (2011)
55. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)