Pattern Recognition 120 (2021) 108107

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

DDAT: Dual domain adaptive translation for low-resolution face verification in the wild

Qianfen Jiao[a], Rui Li[a], Wenming Cao[a], Jian Zhong[a], Si Wu[b], Hau-San Wong[a,*]

[a] Department of Computer Science City University of Hong Kong Hong Kong SAR, China
[b] School of Computer Science and Engineering South China University of Technology Guangzhou, Guangdong 510006, China

ARTICLE INFO

Article history:
Received 23 March 2020
Revised 19 May 2021
Accepted 5 June 2021
Available online 12 June 2021

Keywords:
Low-resolution face verification
Domain adaptation
Image translation
GAN

ABSTRACT

Low-resolution (LR) face verification has received much attention because of its wide applicability in real scenarios, especially in long-distance surveillance. However, the poor quality and scarcity of training data make the accuracy far from satisfactory. In this paper, we propose an end-to-end LR face translation and verification framework to improve the generation quality of face images and face verification accuracy simultaneously. We design a dual domain adaptive structure to generate high-quality images. On one hand, the structure can reduce the domain gap between training data and test data. On the other hand, the structure preserves identity consistency and low-level attributes. Meanwhile, in order to make the whole model more robust, we treat the generated images of the target domain as an extension of the training data. We conduct extensive comparative experiments on multiple benchmark data sets. Experimental results verify that our method achieves improved results in high-quality face generation and LR face verification. In particular, our model DDAT reduces FID to 18.63 and 39.55 on the source and the target domain from 254.7 and 206.19 of the up-sampling results, respectively. Our method outperforms competing approaches by more than 10 percentage points in terms of face verification accuracy on multiple surveillance benchmarks.

1. Introduction

Face is a critical characteristic to distinguish one person from another, therefore face recognition has been extensively applied to security system, person re-identification, video surveillance, *et al.* In the above scenarios, native and unconstrained low-resolution (LR) face images [1] are common. Here, native and unconstrained LR face images are due to constraints of the camera setup and image capture environment, not through artificial down-sampling of high-resolution (HR) images. As shown in Fig. 1, for high-resolution face images of LFW [2], it is straightforward to verify whether the two faces are from the same person. However, when the images are native LR, like QMUL-SurvFace [3], it is difficult to judge whether the two faces correspond to the same identity. The right sub-figure in Fig. 1 displays the ROC curves of two CenterLoss models tested on HR and LR data, respectively. One of them is trained on HR data and another is trained on LR data, where HR data is LFW and LR data is 4× down-sampled LFW. We find that when tested on HR data, the model trained on HR data obtains the best performance. Compared to the model trained and tested on HR data, the performance of the model trained and then tested on LR data decreases slightly, due to the loss of some face details. But if the model trained on HR data is applied to LR data, the performance will drop significantly because of the domain gap between HR and LR data. In other words, if high performance face verification models are trained on HR images, like LFW, it is difficult to generalize the model to LR data sets directly. Furthermore, face recognition is an open-set problem, so it is difficult to acquire enough training data for native LR face verification. Based on the aforementioned constraints, we find native LR face verification represents a very challenging problem.

In order to enhance the verification accuracy of native LR face, a number of approaches attempt to utilize super-resolution methods [1,4,5] to reconstruct face details. As a mainstream approach, super-resolution models obtain remarkable performance in generating high-quality images. However, when the super-resolution model is trained on a source domain which is significantly different from the target domain, that may negatively affect face verification performance on the target domain. As reported in [1,3], there are two reasons why a super-resolution model trained on the source domain does not perform well on the target domain. First,

* Corresponding author.
E-mail addresses: qjiao4-c@my.cityu.edu.hk (Q. Jiao), ruili52-c@my.cityu.edu.hk (R. Li), wenmincao2-c@my.cityu.edu.hk (W. Cao), jianzhong7-c@my.cityu.edu.hk (J. Zhong), cswusi@scut.edu.cn (S. Wu), cshswong@cityu.edu.hk (H.-S. Wong).

https://doi.org/10.1016/j.patcog.2021.108107
0031-3203/

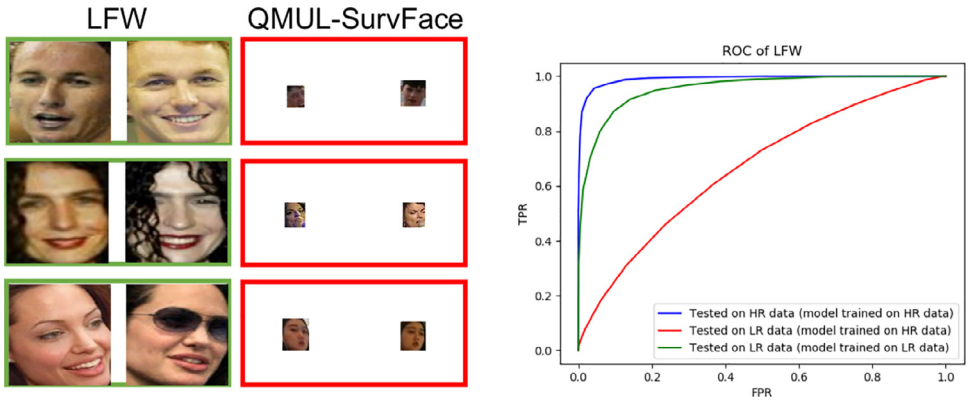


Fig. 1. Left sub-figure displays samples from LFW (Green rectanges) and QMUL-SurvFace (Red rectangles). QMUL-SurvFace is a surveillance benchmark and the average image size is 20 × 16, and it is difficult to verify whether the two faces are from the same person. Right sub-figure displays ROC curves of LFW result obtained by the model trained on LFW (Blue curve), down-sampled LFW result obtained by the model trained on LFW (Red curve) and down-sampled LFW result obtained by the model trained on down-sampled LFW (Green curve). When LFW images are down-sampled to a similar size (25 × 25) as QMUL-SurvFace, the performance of the model trained on LFW drops significantly.

the widely-used mean squared error (MSE) in super-resolution methods leads to a loss of high-frequency details on target domain images. Consequently, the generated images become too smooth. Second, the domain gap between synthesized and native LR images can be large. The super-resolution model is usually trained on the synthesized LR images down-sampled from HR data sets like LFW [2], CASIA-webface [6], *et al*, but it will be applied to native LR data sets. Based on these analyses, we conjecture that super-resolution-based methods may not be suitable for enhancing the accuracy of native LR face verification. As discussed above, the main challenges of enhancing LR face verification accuracy can be summarized as follow: 1) Important features for face verification are lost in LR face images; 2) There are not enough labelled LR face data to train a LR face verification model. A feasible solution is to utilize the existing labelled data sets to supplement useful details for LR images. However, the wide domain gap between synthesized and native LR images makes face verification models trained on the former hard to generalize to the latter. In order to minimize the domain gap, the mainstream solution is to align the data distributions of source domain and target domain in a latent space [7–9]. Nevertheless, aligning data distributions cannot enforce identity consistency, and may lead to loss of low-level attributes [10] which will negatively affect LR face verification accuracy. In order to address these troublesome issues, we propose **D**ual **D**omain **A**daptive **T**ranslation (DDAT) method, which is illustrated in Fig. 2. DDAT generates high-resolution images for LR face images, and the generated HR images are input into the verification module, which is beneficial for improving the LR face verification accuracy.

The architecture of our framework includes two modules, an adaptive adversarial module in the dashed dark yellow polygon, and an anti-perturbation verification module in the dashed purple rectangle. More specifically, we first utilize the adaptive adversarial module to generate high-quality face images to supplement necessary information for LR face images. In the process, in order to minimize the domain gap, we use feature-level domain adaptation in the latent space to align the LR image distributions of the source domain and target domain. Meanwhile, we perform image-level domain adaptation between the generated target domain images and HR images of the source domain to preserve identity consistency and low-level attributes. We then treat the generated images of the target domain as an extension of the source domain, and incorporate them into the training process of the classifier to increase the model generalization capability and robustness.

The main contributions of our paper can be summarized as follows:

1. Our model proposes a dual domain adaptive translation structure to address the challenging low-resolution face verification task. To the best of our knowledge, it is the first time to introduce a dual domain adaptation to face verification model.
2. As an end-to-end model, we focus on native LR face verification in the wild, which can be applied to many real scenarios effectively. We treat the generated images of the native LR images as an extension of the training data, which strengthen the system robustness and generalization capability.
3. Benefitting from the introduction of the dual domain adaptive translation and anti-perturbation verification module, our method achieves state-of-the-art face verification performances on multiple benchmarks. In particular, on the down-sampled CelebA [11], DDAT achieves 70.6% verification accuracy, which is 13.1 percentage points higher than the contrastive method.

The rest of this paper is organized as follows. We review related literatures on face verification in Section 2. In Section 3, we introduce our proposed networks in detail. We present experimental results in Section 4, and draw conclusions in Section 5.

2. Related works

Face recognition has wide applications in real scenarios, and it has been studied extensively for several decades. In this section, we review related works from two aspects: general face recognition and low-resolution face recognition.

2.1. General face recognition

Traditional face recognition has obtained good performance by utilizing handcrafted features or learned descriptors [12–15]. Common descriptors include LBP [12], SIFT [16], HOG [17], Gabor [18], *et al*. Traditional methods adopt a shallow network, and the highest attained accuracy is just 95.17% on LFW by the LBP descriptor [12].

With the development of deep learning, face recognition accuracy has attained impressive improvement, and even surpassed human performance [19–21]. DeepFace [22] is the first work of applying deep learning to face recognition, which achieves 97.35% accuracy on LFW. Google proposes FaceNet [23] using the triplet loss, which increases the inter-class distance and makes the intra-class features more compact. FaceNet uses Inception [24] to achieve

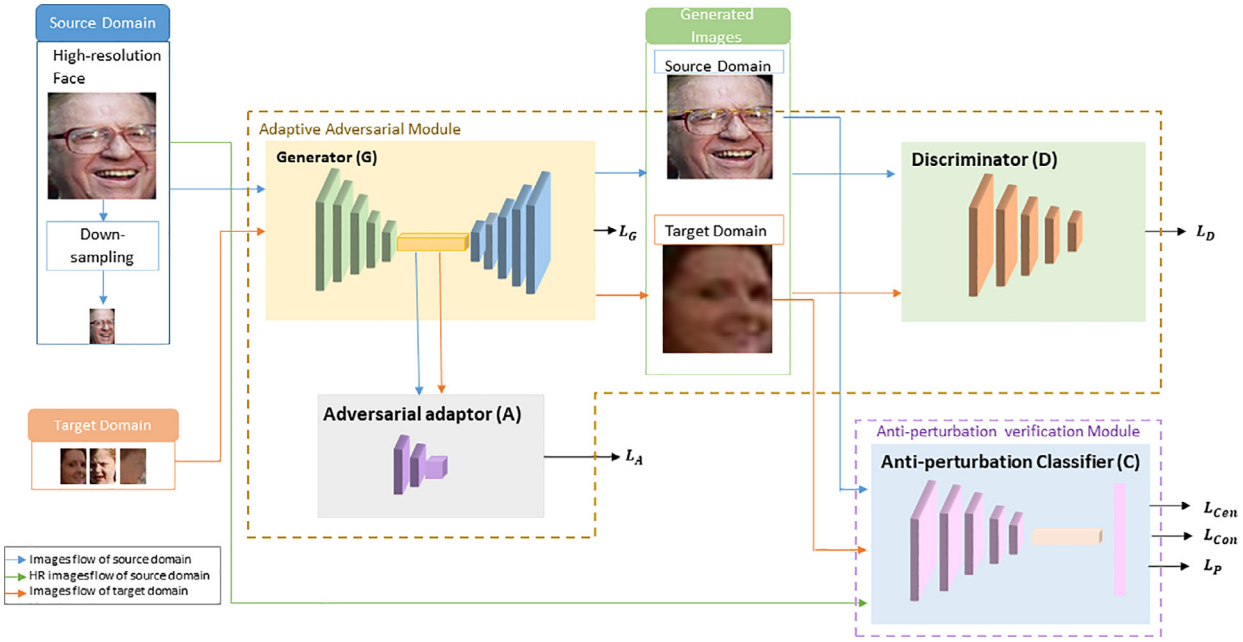


Fig. 2. The framework of our approach, which includes two modules: an adaptive adversarial module in dashed dark yellow polygon and an anti-perturbation verification module in dashed purple rectangle. *A* aims to reduce the negative effect of domain gap between the down-sampled images and native LR images. *D* judges whether the result is fake or real, and constrains *G* to generate target domain images similar to those in source domain. The input of our model includes down-sampled LR and labeled HR images on the source domain and native LR images on the target domain. The down-sampled and native LR images are input into G and A, which decreases their domain gap. Referring to HR images on the source domain, D judges the generated images through G to determine whether they are fake or real. The adversarial architecture can generate high-resolution images on both the source and target domains, which are input into C together with labeled HR images on the source domain. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

an accuracy of 99.63% on LFW. However, face recognition is actually an open-set problem, and learning the discriminative features is vital to the model's effectiveness in the real scenarios. To this end, there are many works to optimize the loss function [25–33]. For example, [25] replaces the traditional softmax classification with the angular margin, which is helpful to learn a discriminative feature. Wang et al. [26] proposes the feature normalization to enhance the accuracy further based on the angular margin. Liu et al. [31] propose a minimum hyperspherical energy (MHE) objective as a regularization for neural networks. Specifically, if MHE is added to SphereFace, then it is referred to as SphereFace+, which can increase inter-class separability. Liu et al. [32] propose a generalized large-margin softmax (L-Softmax), which multiplies a constant with the angle between the sample and the classifier. L-Softmax reduces intra-class variations and improves inter-class separability. Liu et al. [33] propose hyperspherical convolution (SphereConv) through an angular representation on hyperspheres. In particular, SphereFace, ArcFace [34], CosFace and other angular loss functions can be realized by SphereConv. Specifically, the paper proposes three variants of SphereConv operators: linear, cosine and sigmoid.

2.2. LR Face recognition

LR face recognition is a branch of general face recognition and includes two types of solutions: 1) the first type maps features of the probe images and gallery images into a common feature space to compare the distance between them [35–40]; 2) the second type reconstructs useful details for low-resolution images to obtain higher-quality images [1,5,41–49]. Specifically, [36] adopts a teacher-student model to extract critical features[37]. maps LR images and HR images to a common space and minimizes their diversity to enhance accuracy[38]. maps images of different resolutions to a common space and minimize the distance between LR and HR images[40]. presents a multi-dimensional scaling method based on a transformation matrix for both LR and HR face images[1]. presents a Complement Super-Resolution and Identity (CSRI) model that utilizes synthetic LR images, which are down-sampled from HR images, and the corresponding HR images in the source domain to train the super-resolution and classification networks, followed by using native LR images in the target domain to fine-tune the whole network[44]. presents a coarse-to-fine face super-resolution method, which updates LR patches and preserves corresponding geometry of HR patches by dictionary training[5]. supplements residual features from LR images and semantics to the super-resolution network[45]. proposed incorporating face features extracted by a face recognition algorithm to the super-resolution module as prior. Zhang et al. [49] propose a face hallucination network to recover identity information by generating face details. They also formulate a dynamic domain divergence problem, which can be solved by a domain-integrated training method.

Our work focuses on using LR images as the probe to find a LR face image. Different from other works, we enhance the LR face verification accuracy by generating high-quality images of the target domain. We regard the generated target domain images as an extension of source domain training data. Because the target domain data takes part in the training process, the domain gap in both latent space and data space can be minimized at the same time. The resulting small domain gap can further enhance face verification accuracy.

3. The proposed approach

In our setting, the average size of LR images is 20 × 16. Meanwhile, there are not enough labelled training data in low-resolution face image data sets, and at the same time there is a considerable divergence between synthesised low-resolution and native low-resolution face image distributions. Therefore, a face verification model trained on source domain data cannot perform well

on target domain data. In order to address the LR face verification problem, we propose an end-to-end model which incorporates a dual domain adaptive translation structure to enhance LR face verification accuracy by improving the generated image quality. In this section, we introduce our proposed scheme, which is referred to as dual domain adaptive translation for low-resolution face verification in the wild (DDAT). DDAT includes an adaptive adversarial module (the dark yellow dash-line polygon in Fig. 2) and dummyTXdummy- an anti-perturbation classifier module (the purple dash-line rectangle in Fig. 2). The adaptive adversarial module has an image generator, an adversarial adaptor and a discriminator. Incorporating the adversarial adaptor into the generator makes the image translation module adaptive to unseen LR images. The discriminator plays a key role in discriminating the generated images as fake or real, and judging if the generated images preserve identity information from the original LR images. The anti-perturbation classifier module enhances the accuracy of face verification by adding unlabeled target domain data during training. Our model improves the quality of generated images and further enhances the LR face verification accuracy by increasing the model's generalization capability and robustness.

3.1. Adaptive adversarial module

The adaptive adversarial architecture aims to generate HR face images in the target domain. As shown in Fig. 2, the generator (G) and the adaptive discriminator (D) constitute a GAN [50]. The domain adaptive generator contains a feature-level domain adaptor (A) including a reversed layer which plays a similar role as GAN. In the discriminator, we add an image-level domain adaptation loss to the overall loss. Therefore, there are two adversarial structures and dual domain adaptation structures in the whole architecture.

3.1.1. Domain adaptive generator

U-Net [51] combines low-level and high-level features, and uses low-level features to improve the image quality. We thus adopt U-Net [51] as the backbone of our generator, as shown in Fig. 2. Specifically, we input the source domain and target domain data into our model. The source domain data includes HR images (X_H^S) and LR images (X_L^S), while the target domain data only has LR images (X_L^T). The generated face images should be realistic to deceive the discriminator and preserve the same identity with input images. Toward this end, the objective function of the adversarial generator includes two terms: 1) GAN [50] loss to deceive the discriminator, and 2) L_I loss to preserve identity with input images as follows:

$$L_G = L_{G_A} + \lambda L_I, \tag{1}$$

where

$$L_I = E_{x_i \in X_L^S, x_j \in X_H^S}\left(\left\|G(x_i) - x_j\right\|_1\right) + E_{x_k \in X_L^T}(\|F(G(x_k)) - x_k\|_1), \tag{2}$$

$$L_{G_A} = E_{x_i \in X_L^S} \log(1 - D(G(x_i))) + E_{x_k \in X_L^T} \log(1 - D(G(x_k))), \tag{3}$$

where $G(\cdot)$ is the output of the generator, $D(\cdot)$ is the output of the discriminator, and $F(\cdot)$ is the 4× down-sampling operator. To make the generator insensitive to another different domain, we implement an adversarial adaptor at the U-Net [51] bottleneck, as shown in Fig. 2. We align the distributions of the source and target domains on feature-level representations, which makes the generator difficult to discriminate which domain the input image belongs to. The adversarial adaptor will be beneficial to face verification on the target domain, as demonstrated in the experiment section. We set the source domain label as 1 and the target domain label as 0. The objective function L_A for feature-level domain adaptation is defined as follows:

$$\min_A(-\textstyle\sum_i^N(\alpha \log A(f_i^S) + \beta \log(1 - A(f_i^T)))), \tag{4}$$

where α and β are weights of source domain labels and target domain labels, respectively. $A(\cdot)$ is the output of the adversarial adaptor.

3.1.2. Adaptive discriminator

Another component of the adversarial architecture is the adaptive discriminator. Since LR face images of the target domain do not have corresponding HR images, we use HR images of source domain to judge the quality when discriminating generated images. Specifically, during the optimization of the generator, we preserve the same identity by using L_I loss. During the optimization of the discriminator, we make the distribution of synthetic images similar to HR images of the source domain. The objective function L_D of this discriminator is defined as follows:

$$\begin{aligned}&\max_D(E_{x_i \in X_H^S} \log D(x_i) + E_{x_j \in X_L^S} \log(1 - D(G(x_j))) \\ &+E_{x_k \in X_L^T} \log(1 - D(G(x_k)))),\end{aligned} \tag{5}$$

The discriminator makes the generated images realistic enough in the source domain and target domain, and meanwhile the generated images of target domain have close similarities with those of source domain except the identities.

Based on these analyses, we summarize the objective function of the adaptive adversarial module as follows:

$$L_{IG} = \omega_G \cdot L_G + \omega_A \cdot L_A + \omega_D \cdot L_D. \tag{6}$$

In summary, the whole model has dual adversarial structures and dual domain adaptation structures for collaborative training, as explored in the experiment.

3.2. Anti-perturbation verification module

To enhance face verification accuracy on the target domain, it is not enough to merely enhance the quality of the generated images. It is necessary to improve the face verification module. If we just input generated images into current face verification modules such as center loss [14] and triplet loss [23], the face verification accuracy is satisfactory on the source domain but not good on the target domain. In this setting, we introduce an anti-perturbation loss to the verification module to enhance the face verification accuracy on the target domain, as shown in Fig. 3. We add a consistency loss into C to preserve the identity between super-resolution images and HR images in the source domain. Moreover, we add an anti-perturbation loss to C to make the whole module more robust. We choose the center loss [14], denoted by L_{Cen}, as the classification cost in the source domain, which is composed of the classification loss L_{CS} and the sum of distances between each feature and the center of features L_{CD}. To generalize the model to the target domain, we introduce an anti-perturbation loss L_P. Meanwhile, we add a consistency loss, L_{Con}, to keep the identity consistent with the input from the generator to the classifier. The overall objective function for the anti-perturbation verification module can be expressed as follows:

$$L_C = \omega_{Cen} \cdot L_{Cen} + \omega_P \cdot L_P + \omega_{Con} \cdot L_{Con}, \tag{7}$$

where $L_{Cen} = L_{CS} + L_{CD}$, including the classification loss L_{CS} and the feature distance L_{CD} for the classifier on the source domain. They are defined as follows:

$$\begin{aligned}L_{CS} &= -\sum_{i=1}^{N} log\frac{\left(exp\left(f_{S_i}^S[y_i]\right)\right)}{\sum_{j=1}^{n} exp\left(f_{S_i}^S[j]\right)}, \\ L_{CD} &= \frac{1}{2}\sum_{i=0}^{N}\left\|X_{S_i}^S - C_{S_i}\right\|_2^2,\end{aligned} \tag{8}$$

where $f_{S_i}^S$ is the i^{th} feature of a generated image on the source domain, and y_i is the corresponding label. N is the batch number,

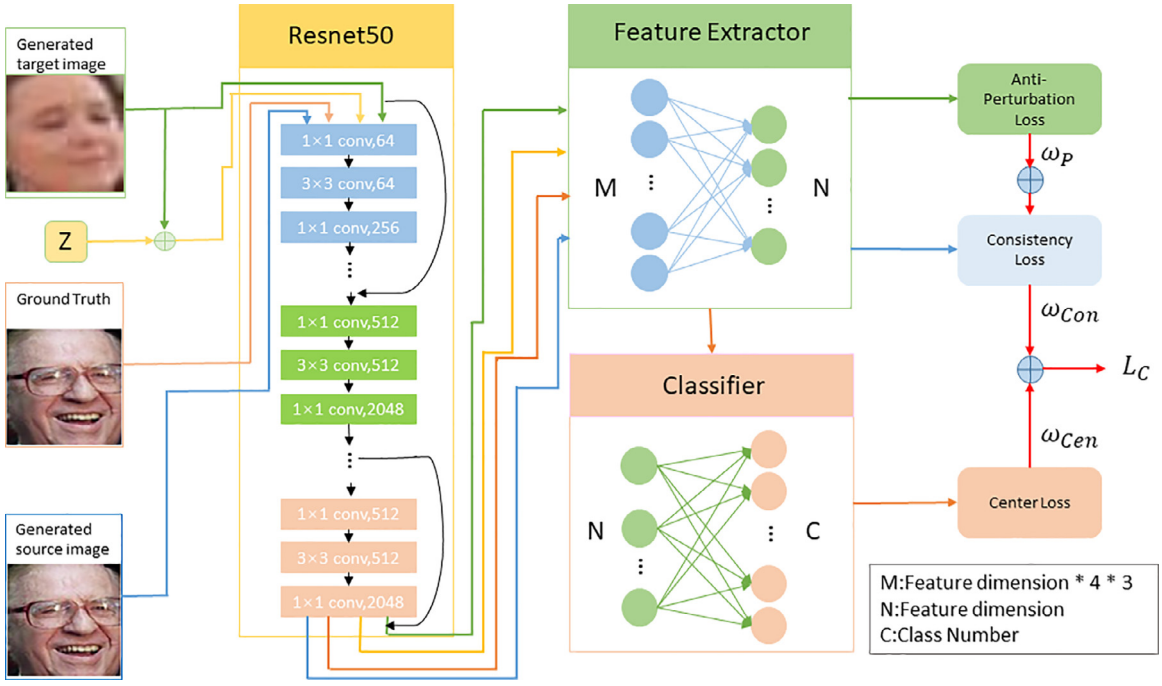


Fig. 3. The architecture of our anti-perturbation classifier. Based on Resnet50, we build a face feature extractor for computing anti-perturbation loss and consistency loss. The features are used to classify face images. The green flow denotes the generated target images. Injecting noise into the yellow flow brings perturbation to source domain data, which includes 2 branches. The orange flow denotes the ground-truth of HR images and the blue flow denotes the generated super-resolution images. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and n is the class number of source domain. C_{S_i} is a class center of the generated images on the source domain. The center loss increases the inter-class distance while reducing the intra-class distance, which contributes to enhancing the verification accuracy on the source domain. Besides, we adopt an anti-perturbation loss L_P to generalize the verification module to the target domain by following VAT [52]. The objective function is defined by:

$$\min_{\Theta_n} E_{x \in X_S^T}[D_{KL}(h_{\Theta_n}(x+\epsilon) \parallel h_{\Theta_n}(x))], \tag{9}$$

where X_S^T denotes the generated images on the target domain, $h_{\Theta_n}(\cdot)$ denotes a feature-extractor function and Θ_n is the nth layer parameter of the feature-extractor. L_P computes the KL-divergence between X_S^T and $X_S^T + \epsilon$, where ϵ denotes noise. A smaller loss indicates that the classifier is more robust.

Since we aim to enhance face verification accuracy on the target domain, we should preserve image identities, and the classification result will impact the quality of image generation. Thus, we add a consistency loss function L_{Con} to obtain a better generation and classification result as follows:

$$L_{Con} = E_{x_i \in X_H^S, x_j \in X_S^S}[D_{KL}(C(x_i) \parallel C(x_j))], \tag{10}$$

where X_H^S denotes the HR images on the source domain and X_S^S denotes the generated images on the source domain. $C(\cdot)$ is the output of the anti-perturbation classifier. In summary, our method includes an adaptive adversarial module and an anti-perturbation verification module, and the final objective function is as follows:

$$L = L_{IG} + \xi L_C, \tag{11}$$

where L_{IG} denotes the total loss of the adaptive adversarial module, ξ is the weight of L_C and L_C denotes the total loss of the anti-perturbation verification module

3.3. Implementation details

3.3.1. Generate high-quality face images on the target domain

In the training process, we use LFW [2] as the source domain and degrade high-quality images in LFW[2] to low-quality as LR images. We degrade LFW[2] images using Gaussian blur and 4× down-sampling by nearest neighbour interpolation to generate LR images of source domain. We input LR images X_L^S and HR images X_H^S of source domain and LR images X_L^T of target domain to the adaptive adversarial module. The proportion of the three types of images is 1:1:1. We adopt U-Net as the backbone of the generator to extract features. In the bottleneck, we add an adversarial adaptor to reduce the domain gap between X_L^S and X_L^T. In the training of adversarial adaptor, we use a reverse layer which introduces an adversarial loss into the generator. The adversarial structure generates high-quality images which are as similar to X_H^S as possible. Besides, we utilize L_I loss to preserve face identities with input images.

3.3.2. Discriminate the quality of the generated images

In the adaptive adversarial module, the discriminator tells apart the generated X_S^S, X_S^T from the real HR images X_H^S. For the target domain, since LR images X_L^T do not have the corresponding HR images, we use X_H^S as the reference and use L_I loss to preserve face identities. The process is similar to style transfer which keeps the same face identity as X_L^T and learns styles of X_H^S like resolution, illumination and so on. Image-level domain adaptation will complement the adversarial adaptor in the generator. The discriminator and the generator compose the entire adaptive adversarial network. The dual adversarial and dual domain adaptation structures help to generate high-quality images and enhance generalization on the target domain.

3.3.3. Enhance face verification accuracy on the target domain

In our classifier, we generalize the classifier trained on the source domain to the target domain. Since source domain images

are labelled and of high-quality, we can obtain a well-behaved model on the source domain. Based on the model, we improve the module by adding perturbation to source domain images. The perturbation comes from target domain images, the amount of which is limited to 10% of source domain images. Since the target domain images do not have labels, we compare the divergence at the feature-level. To further enhance LR face verification accuracy, we use the classification result to improve image generation. In the process, we compare the distributions between the generated X_S^S and the corresponding HR images X_H^S.

The generator is based on U-Net. We resize X_L^S, X_H^S and X_L^T to 128 × 128. Only 10% of the generated target images are used for the classifier. The training process is presented in Algorithm 1.

Algorithm 1 Pseudo-code of our model training process.

Input: LR images X_L^S, HR images X_H^S of source domain, LR images X_L^T of target domain;

Output: θ_G for adversarial generator G, θ_A for adversarial adaptor A, θ_D for adaptive discriminator D and θ_C for anti-perturbation classifier C;

Initialize learning rate η for G, A, D, C, and training epoch N;

for *epoch* = 1 to N **do**
  Randomly sample images x_L^S, and x_H^S from source domain and x_L^T from target domain;
  for each mini-batch **do**
    Extract feature$\{f_L^S, f_H^S, f_L^T\}$ from $\{x_L^S, x_H^S, x_L^T\}$;
    Update A by $\{f_L^S, f_L^T\}$ with Adam:
      $\theta_A \leftarrow Adam(\bigtriangledown(L_A, \theta_A), \eta)$;
    Update $\{f_L^S, f_L^T\}$ to $\{f_L^{S'}, f_L^{T'}\}$ by minimizing the domain divergence with A;
    Update G by $\{f_L^{S'}, f_L^{T'}\}$ with Adam:
      $\theta_G \leftarrow Adam(\bigtriangledown(L_G, \theta_G), \eta)$;
    Update D by $\{x_S^S, x_S^T, x_H^S\}$ with Adam:
      $\theta_D \leftarrow Adam(\bigtriangledown(L_D, \theta_D), \eta)$;
    if n≥ T **then**
      Optimize L_{CS} and L_{CD} in Eq. (8) by $\{x_S^S\}$;
      Optimize L_P in Eq. (9) by $\{x_S^T\}$;
      Optimize L_{Con} in Eq. (10) by $\{x_S^S, x_H^S\}$;
      Update C with SGD: $\theta_C \leftarrow SGD(\bigtriangledown(L_C, \theta_C), \eta)$.
    end if
  end for
end for

4. Experiment

This section first introduces the evaluation metrics, benchmark datasets and experiment settings. We then compare state-of-the-art image generation and LR face verification methods with our proposed approach, followed by experimental result analysis. We adopt the accuracy, FAR@TAR and ROC curve as the criteria to evaluate the performance. For LR face verification, supplementing the details is a key step, so we compare the generation quality in terms of Fret Inception Distance (FID) [53].

4.1. Datasets

1) LFW [2] is one of the most popular labelled face data sets for face recognition research, which includes 5349 people and more than 13,000 images. Face recognition achieves high accuracy on LFW, which is better than human performance. We train our model on LFW which is considered as the source domain, and generalize it to the target domain. *2) CelebA [11]* has 10,177 identities and 202,599 images. This is a labelled data set and has abundant attributes which increase the difficulty of face recognition and lead to a big domain gap with respect to LFW. We adopt down-sampled images of this dataset as a target domain. Since CelebA is widely accepted as a standard benchmark for face recognition, it is used to compare the generation ability among different generation models[11] in our experiment. *3) QMUL-TinyFace [1]* is a large-scale LR face recognition benchmark which consists of 5139 identities and 169,403 native low-resolution images. The images show different poses, illuminations, occlusions and backgrounds, and the average size is 20 × 16. In our experiment, we use it as the target domain and do not involve the labels during training. *4) QMUL-SurvFace [3]* is much more challenging than QMUL-TinyFace, since the data are captured using real surveillance cameras and can be considered as an open-set setting. The data set has 463,507 face images from 15,573 identities. Some of the images are very small, with size below 10 × 10, and are not used in our experiment. However, in the benchmark [3], these small images are used to train the supervised models. *5) VGGFace2 [54]* is a million-level face dataset, which includes 9131 identities. The images are downloaded from Google search and have large variations in pose, illumination, ethnicity and so on. We utilize down-sampled images of the dataset to form the target domain, which is used only in the ablation study to check the sensitivity to different structures.

4.2. Experiment details

4.2.1. Data pre-processing

Images of QMUL-TinyFace and QMUL-SurvFace only include the face region so in order to keep the same setting, we use MTCNN [55] to detect the face region and crop face regions of LFW, CelebA and VGGFace2 images as the training data and test data, respectively. To synthesize LR images for LFW and CelebA, we first perform 4× down-sampling and reshape cropped images to 128 × 128 by nearest neighbor interpolation. For VGGFace2, we downsample the cropped images to 32 × 32 first and reshape the down-sampled images to 128 × 128 by nearest neighbor interpolation. On the other side, the images of QMUL-TinyFace and QMUL-SurvFace are reshaped to the same size as the images of the other data sets.

4.2.2. Hyper-parameter settings

Our model consists of a generator, a discriminator, an adversarial adaptor and an anti-perturbation classifier. The entire model is randomly initialized. In the training process, we utilize Adam optimization with $\beta_1 = 0.5$, $\beta_2 = 0.999$ for G, D and A. The initial learning rate is 0.0002 and training is performed for 200 epochs. We use LSGAN [56] as the adversarial training framework, which is observed to be most suitable for our model. After G starts to generate stable images, C will participate in the training. We initialize SGD [57] with learning rate 0.0002 and momentum 0.9, which performs better than Adam [58] in the training of classifier. The training process will last for 100 epochs. In the experiment, we implement our model using a computer equipped with an Intel Core-i7 CPU and Asus DUAL-RTX2080TI-O11G graphic card.

4.2.3. Architecture details

Our generator follows the architecture of U-Net [51]. The shape of input is 128 × 128×3 and the output shape is 256 × 256×3. The kernel size is 4 × 4, stride is 2 and padding size is 1. In the bottleneck of U-Net, we add the adversarial adaptor, which we will introduce next. The kernel size starts from 64, and is doubled each time until the size reaches 512. Other details of the model are shown in Table 1.

In the adversarial adaptor, we adopt a fully-connected layer to reduce features from 2048 dimensions to 100 dimensions and then to 1 dimension. We use Batch Normalization [59] to stabilize training. The activation function is leaky ReLU [60].

Table 1
The architectures of the generator, the adversarial adaptor and the discriminator. Parameters of FC include input feature dimension and output dimension. As for the Conv, Conv1 and deConv, kernel size is 4 × 4, and padding is 1. The step size of Conv and deConv is 2 and Conv1 is 1. The numbers after Conv and deConv are kernel numbers. The kernel number of Conv1 is 1.'FC-100' denotes FC(2048,100). 'FC-1' denotes FC(100,1). $P_{S/T}$ represents the probability that the image belongs to the source or target domain.

Generator (G) **Input**: $x_L^S \in X_L^S, x_L^T \in X_L^T, x_H^S \in X_H^S$	Adversarial adaptor (A) **Input**: f_i^S and f_i^T	Discriminator (D) **Input**: $x_S^S \in X_S^S, x_S^T \in X_S^T, x_H^S \in X_H^S$
[*Conv*.64, *LReLU*], [*Conv*.128, *BN*, *LReLU*] [*Conv*.256, *BN*, *LReLU*] [*Conv*.512, *BN*, *LReLU*] × 3, [*Conv*.512, *ReLU*] [*deConv*.512, *BN*, *ReLU*] [*deConv*.512, *BN*, *Dropout*, *ReLU*] × 2 [*deConv*.256, *BN*, *ReLU*], [*deConv*.128, *BN*, *ReLU*] [*deConv*.64, *BN*, *ReLU*], [*deConv*.32, *Tanh*]	*FC-100, BN, ReLU* *FC-1* *Sigmoid*	[*Conv*.64, *LReLU*] [*Conv*.128, *BN*, *LReLU*] [*Conv*.256, *BN*, *LReLU*] [*Conv*.512, *BN*, *LReLU*] *Conv1*
Output: X_S^S, X_S^T	**Output**: $P_{S/T}$	**Output**: Fake / True

Table 2
Comparison of FID values obtained by nearest neighbor interpolation, SRGAN, pix2pix and DDAT, respectively. (S) denotes source domain and (T) denotes target domain.

Model	Test Data set					
	LFW (S)	CelebA (T)	CelebA (S)	LFW (T)	QMUL-TinyFace	QMUL-SurvFace
Nearest neighbor interpolation	254.70	206.19	206.19	254.70	301.12	315.02
SRGAN [62]	21.12	101.16	22.08	90.88	199.73	187.66
pix2pix [63]	**17.84**	99.77	**19.61**	83.34	185.61	163.93
DDAT	18.63	**39.55**	21.06	**34.22**	**129.19**	**159.58**

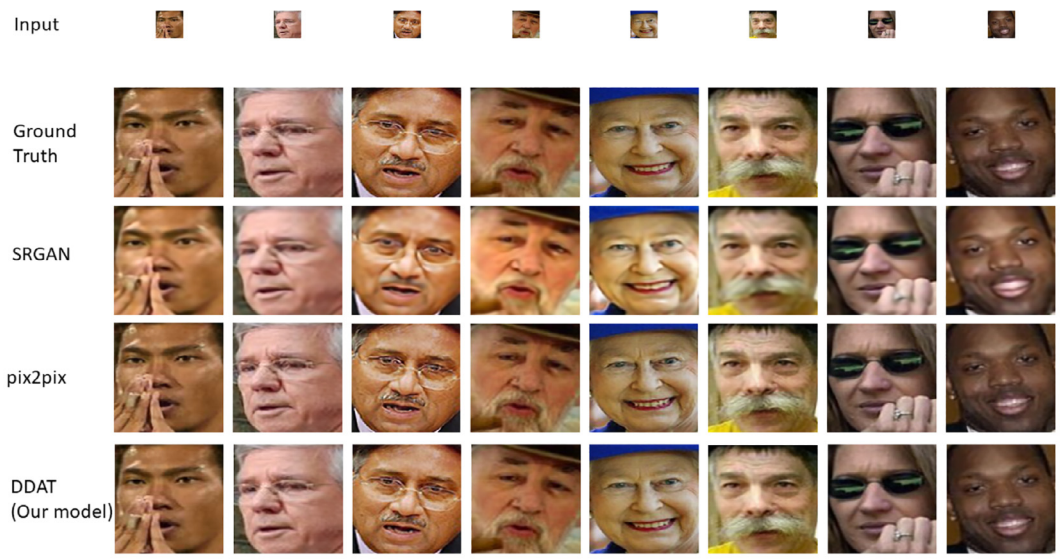
In the anti-perturbation verification module, we use ResNet50 [61] to extract features and reshape features according to the number of face identities in different data sets. After we obtain the features, we minimize their difference to preserve the identity between generated images and the ground-truth. Meanwhile, based on these features, an anti-perturbation loss is added to the classification loss to enhance generalization. Since labeled data are available in the source domain, we utilize CenterLoss [14] as the basic component of the classification loss, which has demonstrated good performance on face recognition.

4.3. Comparison of GAN models on LFW and CelebA

We compare the quality of images generated by our model DDAT against those by SRGAN [62] and pix2pix [63] on LFW and CelebA, which are considered as the source and the target domain, respectively. Among the target domain data sets, CelebA [11] has the corresponding HR images, and we adopt it as the target domain data set when comparing the quality of the generated image. Fig. 4 displays the generated images in the source and the target domains obtained by pix2pix, SRGAN and DDAT, respectively. From Fig. 4(a), we can observe that all these three methods obtain good quality in the source domain. The images generated by SRGAN tend to be a little smoother, compared with the other two methods. The pix2pix obtains results very close to those by DDAT. On the other hand, as shown in Fig. 4(b), our DDAT achieves a significantly higher quality in the target domain. On the contrary, SRGAN may generate fake images, and the images generated by pix2pix are not satisfactory enough. In order to show the details, we display the image generation results of pix2pix and DDAT on CelebA in Fig. 5. A number of image regions are highlighted in red and green rectangles for comparison, from which we observe that DDAT performs better in terms of the generation of high-resolution details.

In addition, we compare these three methods quantitatively. In order to evaluate the quality of the generated images, we compute FID of the nearest neighbor interpolated input, SRGAN, the pix2pix and our DDAT, respectively. Table 2 displays the FID results obtained on both the source and the target domains, and a lower FID result corresponds to a better generation capability. We observe that SRGAN can achieve good generation performance in the source domain, with a FID of 21.12 which is significantly lower than that of nearest neighbor interpolation. Interestingly, the pix2pix obtains the best performance, which is slightly outperforms our model by about 1 point. A possible reason is that the perturbation added to the source domain in our model leads to slight degradation of the generation quality. On the other hand, our model achieves the best performance on the target domain, which is 39.55 in terms of FID. In the target domain testing process, we use 90% source domain data from LFW and 10% target domain data from CelebA [11] to train SRGAN, pix2pix and our DDAT. Meanwhile, our generator adopts domain adaptation in the image generation process through A and D, which leads to good generalization performance of the generation model. The improved generalization capability is important for translating high-quality images in the target domain. However, SRGAN and pix2pix cannot attain such generalization ability, and their translation results are not as good as our model. We observe that the FID value corresponding to our approach is much lower than those of SRGAN and pix2pix in the target domain, which indicates that our generated images can provide a better match to the target distribution. In summary, in the target domain (CelebA), our model significantly outperforms SRGAN and pix2pix. We have also exchanged the source domain data set and target domain data set. In this setting, CelebA becomes the source domain and LFW becomes the target domain. In this case, our model also results in the lowest FID value in the target domain.

To highlight the quality of the generated images on native LR data, we also have calculated the FID score between native LR data and LFW, where the former includes QMUL-TinyFace and QMUL-SurvFace. We observe that all the three generation models can effectively produce high resolution images, and the proposed DDAT model outperforms the competing models significantly. In particular, on QMUL-TinyFace, the FID score of our DDAT is 129.19, while that of the second-best model is 185.61 (obtained by pix2pix). This significant performance improvement is due to the incorporation of the domain adaptation module in our generation net, which decreases the domain gap between training and test LR data. In this case, the generated images contain more detail. Overall, the FID re-



(a) Generated images on the source domain.

(b) Generated images on the target domain.

Fig. 4. Generated images by SRGAN, pix2pix and our DDAT on the source and the target domains, respectively.

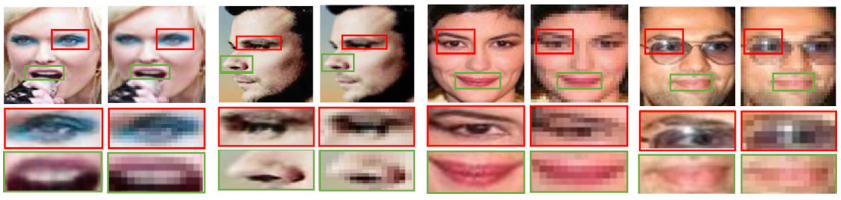


Fig. 5. Details of the generated images by our DDAT and pix2pix, respectively. In every example, the left image is generated by DDAT and the right one is generated by pix2pix. The generated images by DDAT have much better quality in terms of high-resolution details, as highlighted in the red and green rectangles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sult demonstrates that DDAT can successfully generate high-quality images.

In addition, we check how the GAN models impact the face verification results. We compare the results obtained by different generation methods based on the same verification model, CenterLoss. These results are shown in Table 3. Here, Original means using LR images directly without translating them to HR images before face verification. For LFW, we downsample the images to 32×32 and for QMUL-Tinyface, we reshape the images to 32×32 by nearest neighbor interpolation. DDAT refers to our approach. In the source domain data set (LFW), SRGAN, pix2pix and DDAT attain verification accuracies of 80.3%, 90.6%, and 94.4%, respectively,

Table 3
Comparison of mean accuracy obtained by the original, nearest neighbor interpolation, SRGAN, pix2pix and DDAT on down-sampled LFW and QMUL-TinyFace, respectively.

Model	Test Data set	
	LFW	QMUL-TinyFace
Original	74.0%	73.0%
Nearest neighbor interpolation	76.0%	71.7%
SRGAN [62]	80.3%	65.0%
pix2pix [63]	90.6%	74.9%
DDAT	**94.4%**	**76.2%**

which are significantly higher than that of the original. However, for the target domain data set (QMUL-TinyFace), a performance drop to 65.0% is observed for SRGAN, since the super-resolution model is trained on LFW and cannot generalize well to QMUL-TinyFace. Pix2pix brings about a slight enhancement of less than 2 percentage points compared to the original. On the other hand, our generation method can achieve 76.2% in terms of accuracy, which outperforms the result of pix2pix by 1.3 percentage points. The face verification results demonstrate that DDAT can generate high-quality images which in turn enhances the verification accuracy.

4.4. Low-Resolution face verification

In this section, we evaluate LR face verification performance in terms of accuracy, and TAR@FAR on the data sets CelebA, QMUL-TinyFace and QMUL-SurvFace, which can demonstrate the effectiveness of DDAT model comprehensively. In each data set, there are four types of experiments: 1) comparisons among traditional face recognition models, all of which are trained on HR images of LFW and are used on LR images of target domain data sets directly; 2) comparisons among different operators of SphereConv [33], which can realize SphereFace, ArcFace [62], CosFace and other angular loss functions; 3) comparisons among traditional face recognition models, all of which are trained on LR images of LFW and are used on LR images of target domain dataset; and 4) comparisons among a baseline method, CSRI [1], our DDAT and DDAT trained with labels in the target domain, all of which involve the process of generating HR images on the target domain. The baseline uses pix2pix to generate high-quality images, and the accuracy and TAR@FAR are evaluated based on CenterLoss [14]. The pix2pix model adopts UNet-128 as the backbone and implements pixel-level consistency between the generated images and corresponding HR images. The key parameter is the ratio of GAN loss to pixel-level consistent loss, which is 150. CenterLoss implements the distance loss between features and corresponding feature centers and softmax loss simultaneously, which is trained by labeled HR images on the source domain. The key parameter of CenterLoss is the ratio of softmax loss to the distance between the feature and the feature center, which is 0.03. Baseline is trained on the source domain, so the hyperparameters are the same for the three datasets. We conduct these experiments on all three data sets, as shown in Table 4, Table 5 and Table 6. The methods of the first three parts are traditional ones and the four methods of the last part apply image generation before the verification.

Although traditional face recognition models have achieved state-of-the-art performances on HR face images, they perform poorly on LR face images. CenterLoss [14], FaceNet [23], CosFace [27], SphereFace [25], SphereFace+ [31] and L-Softmax [32] are good face recognition models whose face verification accuracies surpass 90% on LFW. However, from Table 4, we observe that the accuracy drops significantly on LR images by about 40% on down-sampled data of CelebA images, and the highest accuracy is 56.3%. This is more noticeable with respect to TAR@FAR. When FAR is 0.1%, TAR@FAR is lower than 1%. These results demonstrate that on the one hand, traditional face verification models perform poorly on LR face verification, and on the other hand, LR face verification is still an open research problem. The six classical models of the first part apply different loss functions for face recognition, and from our experiments results, we observe that CenterLoss perform well with respect to the accuracy metric. Therefore, in our model, we adopt CenterLoss as a component of our classification loss.

From the results of the second type of experiments, We find that SphereConv can achieve a comparable result with CosFace and SphereFace. For example, in Table 4, the accuracy of CelebA by CosFace and SphereFace are 55.3% and 55.7% and the best result of SphereConv is 54.8%. As shown in the third part of Table 4, Table 5 and Table 6, the verification performance can be enhanced when training a model on LR data, especially for CelebA. From Table 4, we observe that the accuracy of the model trained on LR data is improved by more than 10 percentage points, compared to the result from the model trained on HR data. We consider that the domain gap between down-sampled CelebA and LFW is relatively small, and a model trained on down-sampled LFW can thus maintain reasonable performance on down-sampled CelebA. From Table 5 and 6, we can also observe that the inclusion of the down-sampled training data lead to the improvement of about 3 and 4 percentage points in terms of mean accuracy.

DDAT, CSRI and the baseline can generalize to LR face images by translating LR images to HR images before face verification. Here, the baseline adopts pix2pix as an image translation model and uses CenterLoss for face verification. CSRI utilizes SRGAN to generate super-resolution images, which is observed to lead to a lower face verification accuracy, as shown in Table 3. To refine the model, CSRI adds a refinement stage to the training process, which enhances face verification performance. Different from the traditional methods, the two methods attempt to improve the verification accuracy by first generating high-quality images.

For CelebA, as shown in Table 4, the traditional model cannot obtain good performance on down-sampled images. It is observed that the performance drops significantly when compared with face verification accuracy on HR images. Moreover, because CelebA data has much more variations than LFW in terms of illumination conditions, occlusions, poses, expressions and so on, the accuracy of down-sampled images of CelebA is the lowest among the three target domain data sets. Traditional models achieve a similar accuracy of around 55%. The baseline method can bring about a slight improvement to 57.5%, and CSRI can further enhance the accuracy to 61.2%. On the contrary, DDAT can significantly improve the accuracy by 13 percentage points with respect to the baseline, and by 9.4 percentage points compared to CSRI.

For QMUL-TinyFace, as shown in Table 5, traditional models obtain an above 70% face verification accuracy, which is still not satisfactory enough. Both the baseline and CSRI improve the accuracy slightly by around 2 percentage points. Our model obtains a significant improvement in terms of accuracy by 11 percentage points and 9.7 percentage points, compared to the baseline and CSRI, respectively. At the same time, TAR@FAR=1% of our model achieves 46.1%, which is much higher than 37.1% obtained by FaceNet which is the best result obtained by the competing methods.

For QMUL-SurvFace, the accuracy of the supervised CenterLoss model is 88.0% and TAR@FAR=0.1% is 26.8%[3]. As shown in Table 6, our model obtains 82.5%, which is trained in an unsupervised setting on the target domain while only using the source domain labels. The result is higher than that of the baseline (73.9%) and CSRI (73.4%) by 8.6 percentage points and 9.1 percentage points, respectively. Although the accuracy and TAR@FAR cannot attain the performance of the supervised model, the accuracy and TAR@FAR are better than other competing methods including CSRI and baseline except when TAR@FAR=0.1%.

Table 4
Face verification comparison results on CelebA. * indicates the model trained on LR data.

Model	CelebA				
	TAR(%)@FAR				Mean Accuracy(%)
	30%	10%	1%	0.1%	
CenterLoss [14]	42.5%	19.7%	3.12%	0.24%	56.2%
FaceNet [23]	63.5%	13.1%	2.23%	0.75%	54.2%
CosFace [27]	68.6%	15.2%	1.96%	0.28%	55.3%
SphereFace [25]	68.4%	12.2%	2.12%	0.21%	55.7%
SphereFace+ [31]	67.4%	17.9%	2.56%	0.44%	54.3%
L-Softmax [32]	62.1%	16.0%	1.62%	0.32%	56.3%
SphereConv w/ linear operator [33]	-	-	-	-	54.8%
SphereConv w/ cosine operator [33]	-	-	-	-	53.6%
SphereConv w/ sigmoid operator[33]	-	-	-	-	53.3%
CenterLoss* [14]	64.1%	35.3%	7.53%	0.56%	67.3%
FaceNet* [23]	70.3%	30.1%	5.04%	0.63%	67.3%
CosFace* [27]	71.4%	35.2%	**11.1%**	**2.45%**	66.2%
SphereFace* [25]	69.4%	29.2%	4.92%	0.58%	66.5%
Baseline	43.0%	17.5%	2.56%	0.52%	57.5%
CSRI [1]	51.8%	23.3%	4.07%	0.20%	61.2%
DDAT	**70.0%**	**40.7%**	8.48%	1.24%	**70.6%**
DDAT w/ target domain labels	71.6%	42.9%	9.84%	1.82%	71.2%

Table 5
Face verification comparison results on QMUL-TinyFace. * indicates the model trained on LR data.

Model	QMUL-TinyFace				
	TAR(%)@FAR				Mean Accuracy(%)
	30%	10%	1%	0.1%	
CenterLoss [14]	76.0%	55.4%	25.4%	6.72%	73.0%
FaceNet [23]	82.5%	64.8%	37.1%	13.5%	74.2%
CosFace [27]	81.9%	63.8%	36.1%	10.3%	73.5%
SphereFace [25]	82.2%	65.9%	26.7%	5.35%	72.2%
SphereFace+ [31]	75.9%	41.3%	25.8%	6.74%	75.5%
L-Softmax [32]	78.0%	42.3%	26.9%	8.64%	74.1%
SphereConv w/ linear operator [33]	-	-	-	-	71.3%
SphereConv w/ cosine operator [33]	-	-	-	-	69.0%
SphereConv w/ sigmoid operator[33]	-	-	-	-	73.5%
CenterLoss* [14]	81.7%	58.4%	17.9%	5.42%	76.9%
FaceNet* [23]	83.1%	53.5%	18.1%	7.57%	77.0%
CosFace* [27]	82.3%	57.9%	16.0%	6.76%	76.3%
SphereFace* [25]	82.0%	59.1%	17.5%	7.56%	77.3%
Baseline	76.1%	52.7%	20.8%	6.37%	74.9%
CSRI [1]	76.6%	49.3%	16.3%	5.04%	75.3%
DDAT	**93.1%**	**81.1%**	**46.1%**	**13.8%**	**85.9%**
DDAT w/ target domain labels	94.6%	86.4%	62.2%	41.5%	88.3%

Table 6
Face verification comparison results on QMUL-SurvFace. * indicates the model trained on LR data.

Model	QMUL-SurvFace				
	TAR(%)@FAR				Mean Accuracy(%)
	30%	10%	1%	0.1%	
CenterLoss [14]	75.1%	54.2%	26.4%	11.8%	72.8%
FaceNet [23]	81.5%	64.1%	36.6%	13.5%	70.6%
CosFace [27]	80.6%	65.5%	31.7%	12.1%	71.6%
SphereFace [25]	80.2%	62.3%	32.5%	12.5%	71.3%
SphereFace+ [31]	71.7%	32.2%	14.4%	12.6%	73.1%
L-Softmax [32]	78.9%	60.2%	31.3%	10.4%	72.0%
SphereConv w/ linear operator [33]	-	-	-	-	70.6%
SphereConv w/ cosine operator [33]	-	-	-	-	66.1%
SphereConv w/ sigmoid operator[33]	-	-	-	-	65.3%
CenterLoss* [14]	72.9%	45.3%	12.4%	1.41%	71.6%
FaceNet* [23]	83.6%	47.2%	19.2%	9.73%	74.2%
CosFace* [27]	80.2%	42.1%	17.2%	8.69%	72.2%
SphereFace* [25]	81.3%	44.9%	18.8%	8.72%	73.4%
Baseline	75.0%	54.5%	22.1%	**15.8%**	73.9%
CSRI [1]	78.6%	53.1%	18.9%	12.4%	73.4%
DDAT	**88.7%**	**72.6%**	**37.0%**	7.71%	**82.5%**
DDAT w/ target domain labels	90.4%	75.5%	40.4%	16.4%	83.6%

Table 7
Ablation comparisons of our DDAT on CelebA, QMUL-TinyFace, QMUL-SurvFace.

Model	Training Data		CelebA	QMUL-TinyFace	QMUL-SurvFace
	Target Domain(10%)	Source Domain(90%)			
Baseline		√	56.2%	74.9%	73.9%
Baseline w/ Domain adaptation	√	√	58.3%	76.2%	74.7%
Baseline w/ Anti-perturbation	√	√	69.2%	84.8%	81.7%
DDAT	√	√	**70.6%**	**85.9%**	**82.5%**
DDAT w/o domain adaptation loss	√	√	69.5%	82.6%	81.5%
DDAT w/o anti-perturbation loss	√	√	66.8%	80.6%	74.0%
DDAT w/o consistency loss	√	√	69.2%	84.8%	81.7%

As shown in Table 4, Table 5 and Table 6, 'DDAT w/ target domain labels' represents the DDAT model trained with labeled target data. We observe that inclusion of target labels can lead to significant improvements in the verification performance. It is notable that the labeled target data improves TAR significantly when FAR is small. In particular, when FAR is 0.1%, the values of TAR are enhanced significantly, from 13.8% to 41.5% on QMUL-TinyFace and from 7.71% to 16.4% on QMUL-SurvFace.

4.5. Ablation study

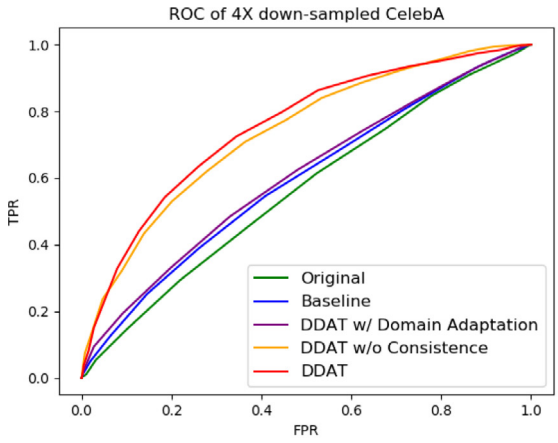
To verify contributions of the different components in our model to performance improvement, we conduct ablation studies on the CelebA, QMUL-TinyFace and QMUL-SurvFace data sets. Our ablation study includes two parts. One of them is based on the baseline model which uses pix2pix to generate high-quality images and then use CenterLoss to test face verification accuracy. We construct three variants and measure the corresponding verification accuracy. First, we add domain adaptation to the baseline model to check how this will enhance the verification accuracy. This is referred to as Baseline w/ Domain Adaptation. Second, based on the first step, we add the anti-perturbation loss, which is referred to as Baseline w/ Anti-perturbation. Finally, we combine the consistency loss with the above two variants to form the complete DDAT model. In another part of our ablation study, we build three variants by removing the domain adaptation term L_A, anti-perturbation term L_P, and consistency term L_{Con} from the overall loss function of DDAT, respectively. We perform a comprehensive analysis of the receiver operating characteristic (ROC) curves and mean accuracy. In Fig. 6, we present ROC curves of face verification results on the baseline and our model w/o certain components. We also include the ROC curve of the original LR data sets based on the CenterLoss model, denoted as Original in Fig. 6. We can observe that the complete DDAT model (red line) corresponds to the best performance. Meanwhile, every component of our model can enhance the performance above that of the original and the baseline.

The face verification results of the variants are summarized in Table 7. We can observe that DDAT achieves a higher face verification accuracy than the baseline by about 14 percentage points, 11 percentage points and 9 percentage points, respectively on these three data sets. Baseline w/ domain adaptation improves performances by about 2.1 percentage points, 1.3 percentage points and 0.8 percentage points, respectively. When the domain gap is larger, the effects of the domain adaptive module is more obvious. The most important term for generalization is the anti-perturbation loss, which brings about significant enhancement to face verification accuracy on the target domain. As shown in Table 7, baseline w/ anti-perturbation enhances face verification accuracy by 10.9 percentage points, 8.6 percentage points and 7.0 percentage points, respectively on these three data sets. In order to preserve the identity, DDAT includes the consistency loss, which plays a key role and can bring about a 1 percentage point performance improvement on the three data sets. As shown in the second part of Table 7, we build three variants by removing the domain adaptation, anti-perturbation, and consistency from the overall loss function of DDAT, respectively. In all the cases, we can observe performance drops compared with the complete DDAT, which indicates the effectiveness of our improvement techniques in the three aspects. In particular, the inclusion of the anti-perturbation term improves the performance by about 3–8 percentage points. On the other hand, the domain adaptation term is important when domain discrepancy is significant. The consistency term is also an effective choice and leads to an improvement across the datasets. Based on the above analysis, we verify that every variant of our model can bring enhancement to the verification accuracy.

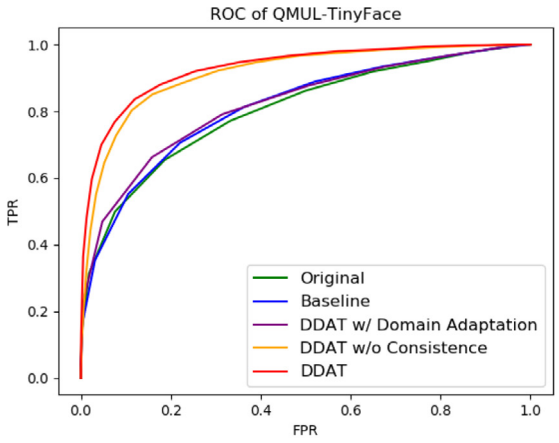
We explore the sensitivity to different structures on down-sampled VGGFace2, and show the result in Table 8. When there is no adversarial adaptor, the accuracy decreases by 1.8 percentage points compared to DDAT. DDAT w/o anti-perturbation denotes replacing our anti-perturbation classifier with CenterLoss, whose accuracy decreases by 5.3 percentage points. When there are no adversarial adaptor and anti-perturbation classifier, DDAT is reduced to the baseline which incorporates pix2pix and CenterLoss, and the accuracy decreases by 7.3 percentage points compared to DDAT. The different modules play complementary roles in improving the accuracy.

4.6. Further analysis

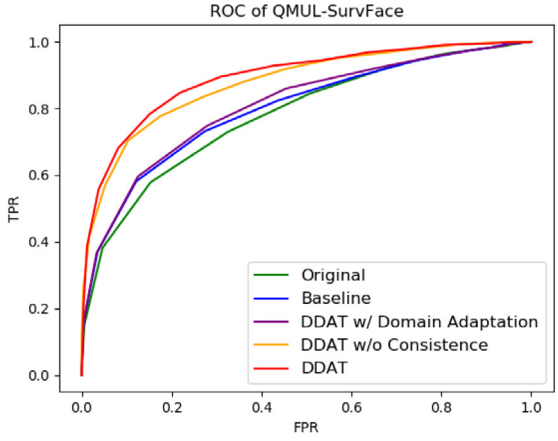
To explore the impact of the large-margin softmax loss in our model, we train our model using a modified loss function, in which we replace the softmax in the CenterLoss model with a large-margin softmax. Here, we use the SphereFace loss and the resulting model is referred to as æAngular CenterLossg. The experiments are conducted on QMUL-TinyFace, and the results are shown in Table 9. To ensure fair comparison, testing data is first translated by our adaptive adversarial module, which is the same process as our proposed DDAT. We observe that the Angular CenterLoss indeed leads to performance gains in terms of mean accuracy and different values of FAR. Compared to the baseline (CenterLoss), the gain resulting from Angular CenterLoss can reach about 5 percentage points in terms of mean accuracy. The large-margin loss benefits the verification task, since it is designed to reduce intra-class variants and improve inter-class separability. When combined with other improvement techniques of the proposed DDAT, the verification performance can be further enhanced, but the degree of improvement becomes smaller. A possible reason is that the large-margin loss and the anti-perturbation loss play similar roles in the training process. Both of them aim to make the data points distant from the decision boundaries. Compared to the large-margin loss, the anti-perturbation loss does not depend on the identity information, and it is thus useful for regularizing the model in an unsupervised fashion.



(a) ROC on down-sampled CelebA.



(b) ROC on QMUL-TinyFace.



(c) ROC on QMUL-SurvFace.

Fig. 6. ROC curves on down-sampled CelebA, QMUL-TinyFace and QMUL-SurvFace datasets.

Table 8
Sensitivity analysis on down-sampled VGGFace2.

Model	DDAT	DDAT w/o adversarial adaptor	DDAT w/o anti-perturbation	Baseline
ACC	75.2%	73.4%	69.9%	67.9%

Table 9
Face verification comparison results with a large-margin loss on QMUL-TinyFace.

Model	QMUL-TinyFace				
	TAR(%)@FAR				Mean Accuracy(%)
	30%	10%	1%	0.1%	
Baseline	77.0%	53.8%	21.0%	5.03%	74.8%
Baseline w/ Angular CenterLoss	85.6%	67.3%	37.8%	19.6%	79.4%
DDAT w/ Angular CenterLoss	86.5%	68.3%	38.3%	25.2%	80.2%
DDAT	93.1%	81.1%	46.1%	13.8%	85.9%

5. Conclusion

In this paper, we propose a dual domain adaptation image translation model for LR face verification. This model addresses the issue of LR face verification and delivers good generalization performance in different domains. This is important for enhancing the face verification accuracy on low-resolution data sets which do not have enough labels or are unlabeled. Toward this end, we design an adversarial adaptor in the generator and utilize the adaptor to form a domain adaptive structure, which makes the generated image more realistic. Furthermore, in the discriminator we use the high-resolution images in the source domain to constrain the generated images in the target domain, which is another domain adaptation mechanism in our proposed framework. Apart from enhancing the generated image quality, we improve the classifier by combining a consistency loss and an anti-perturbation loss to ensure that the generated images maintain the same identities as the input and to make the model more robust, respectively. Experi-

mental results demonstrate that the proposed DDAT achieves better performance in terms of the generated image quality and face verification accuracy than other competing approaches. In addition, it can be easily adapted to any other LR image data sets.

6. Discussion and future work

Although our model can bring significant enhancement to LR image quality and LR face verification, there are some limitations. First, LR images on the source domain are synthesized via downsampling, which leads to extra loss of image details. Second, when the quality of the real LR images is low, for example, due to low levels of illumination, and the synthesized images on the source domain are significantly different from the real images on the target domain, it is difficult for the adversarial adaptor of our model to decrease the domain discrepancy effectively. Finally, if the LR images include too few identity-related features, the proposed model cannot ensure that the identity is preserved in the process of image translation. In view of these limitations, we plan to further improve our model from the following aspects. For the first two limitations, we can apply a GAN model to degrade HR images directly, which avoids the extra loss of image details due to downsampling on the source domain. On the other hand, to preserve more identity-related features, we will apply robust identity regularization or adopt a progressive architecture. In addition, we will further improve LR face verification performance by generating identity-preserved HR images.

Declaration of Competing Interest

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

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (Project No. 62072189), in part by the Natural Science Foundation of Guangdong Province (Project No. 2020A1515010484), and in part by the Research Grants Council of the Hong Kong Special Administrative Region (Project No. CityU 11201220).

References

[1] Z. Cheng, X. Zhu, S. Gong, Low-resolution face recognition, in: Asian Conference on Computer Vision, Springer, 2018, pp. 605–621.
[2] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, 2008.
[3] Z. Cheng, X. Zhu, S. Gong, Surveillance face recognition challenge, arXiv preprint: 1804.09691(2018).
[4] Y. Bai, Y. Zhang, M. Ding, B. Ghanem, Finding tiny faces in the wild with generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 21–30.
[5] Z. Wang, S. Chang, Y. Yang, D. Liu, T.S. Huang, Studying very low resolution recognition using deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4792–4800.
[6] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch, arXiv preprint: 1411.792(2014).
[7] M. Wang, W. Deng, Deep visual domain adaptation: a survey, Neurocomputing 312 (2018) 135–153.
[8] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans Knowl Data Eng 22 (10) (2009) 1345–1359.
[9] Y. Taigman, A. Polyak, L. Wolf, Unsupervised cross-domain image generation, arXiv preprint: 1611.02200(2016).
[10] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, CyCADA: Cycle-consistent adversarial domain adaptation, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, volume 80, PMLR, Stockholmsmssan, Stockholm Sweden, 2018, pp. 1989–1998. http://proceedings.mlr.press/v80/hoffman18a.html
[11] Z. Liu, P. Luo, X. Wang, X. Tang, Deep Learning Face Attributes in the Wild, in: Proceedings of International Conference on Computer Vision (ICCV), IEEE, 2015.
[12] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Transactions on Pattern Analysis & Machine Intelligence (12) (2006) 2037–2041.
[13] M. Turk, A. Pentland, Eigenfaces for recognition, J Cogn Neurosci 3 (1) (1991) 71–86.
[14] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European conference on computer vision, Springer, 2016, pp. 499–515.
[15] K. Etemad, R. Chellappa, Discriminant analysis for recognition of human face images, Josa a 14 (8) (1997) 1724–1733.
[16] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int J Comput Vis 60 (2) (2004) 91–110.
[17] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, 2005.
[18] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Trans. Image Process. 11 (4) (2002) 467–476.
[19] N. Kumar, A. Berg, P.N. Belhumeur, S. Nayar, Describable visual attributes for face verification and image search, IEEE Trans Pattern Anal Mach Intell 33 (10) (2011) 1962–1977.
[20] Y. Sun, D. Liang, X. Wang, X. Tang, Deepid3: Face recognition with very deep neural networks, arXiv preprint: 1502.00873(2015).
[21] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1891–1898.
[22] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.
[23] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[25] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 212–220.
[26] F. Wang, X. Xiang, J. Cheng, A.L. Yuille, Normface: L2 hypersphere embedding for face verification, in: Proceedings of the 25th ACM International Conference on Multimedia, in: MM 17, Association for Computing Machinery, New York, NY, USA, 2017, p. 10411049, doi:10.1145/3123266.3123359.
[27] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.
[28] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[29] X. Zhang, R. Zhao, Y. Qiao, X. Wang, H. Li, Adacos: Adaptively scaling cosine logits for effectively learning deep face representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10823–10832.
[30] X. Zhang, R. Zhao, J. Yan, M. Gao, Y. Qiao, X. Wang, H. Li, P2sgrad: Refined gradients for optimizing deep face models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9906–9914.
[31] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, L. Song, Learning towards minimum hyperspherical energy, in: Advances in neural information processing systems, 2018, pp. 6222–6233.
[32] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, in: ICML, volume 2, 2016, p. 7.
[33] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao, L. Song, Deep hyperspherical learning, in: Advances in neural information processing systems, 2017, pp. 3950–3960.
[34] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[35] M. Pang, Y.-m. Cheung, B. Wang, R. Liu, Robust heterogeneous discriminative analysis for face recognition with single sample per person, Pattern Recognit 89 (2019) 91–107.
[36] S. Ge, S. Zhao, C. Li, J. Li, Low-resolution face recognition in the wild via selective knowledge distillation, IEEE Trans. Image Process. 28 (4) (2018) 2051–2062.
[37] B. Li, H. Chang, S. Shan, X. Chen, Low-resolution face recognition via coupled locality preserving mappings, IEEE Signal Process Lett 17 (1) (2009) 20–23.
[38] S. Biswas, K.W. Bowyer, P.J. Flynn, Multidimensional scaling for matching low-resolution face images, IEEE Trans Pattern Anal Mach Intell 34 (10) (2011) 2019–2030.
[39] C.-X. Ren, D.-Q. Dai, H. Yan, Coupled kernel embedding for low-resolution face image recognition, IEEE Trans. Image Process. 21 (8) (2012) 3770–3783.
[40] H. Zhang, J. Yang, Y. Zhang, N.M. Nasrabadi, T.S. Huang, Close the loop: Joint blind image restoration and recognition with sparse representation prior, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 770–777.
[41] R. Abiantun, F. Juefei-Xu, U. Prabhu, M. Savvides, Ssr2: sparse signal recovery for single-image super-resolution on faces with extreme low resolutions, Pattern Recognit 90 (2019) 308–324.

[42] X. Chen, Z. Zhang, B. Wang, G. Hu, E.R. Hancock, Recovering variations in facial albedo from low resolution images, Pattern Recognit 74 (2018) 373–384.
[43] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, M. Nixon, Super-resolution for biometrics: a comprehensive survey, Pattern Recognit 78 (2018) 23–42.
[44] J. Jiang, R. Hu, Z. Wang, Z. Han, Face super-resolution via multilayer locality–constrained iterative neighbor embedding and intermediate dictionary learning, IEEE Trans. Image Process. 23 (10) (2014) 4220–4231.
[45] P.H. Hennings-Yeomans, S. Baker, B.V. Kumar, Simultaneous super-resolution and feature extraction for recognition of low-resolution faces, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[46] S. Kolouri, G.K. Rohde, Transport-based single frame super resolution of very low resolution face images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4876–4884.
[47] P. Zhang, X. Ben, W. Jiang, R. Yan, Y. Zhang, Coupled marginal discriminant mappings for low-resolution face recognition, Optik (Stuttg) 126 (23) (2015) 4352–4357.
[48] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807.
[49] K. Zhang, Z. Zhang, C.-W. Cheng, W.H. Hsu, Y. Qiao, W. Liu, T. Zhang, Super-identity convolutional neural network for face hallucination, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 183–198.
[50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[51] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[52] T. Miyato, S.-i. Maeda, M. Koyama, S. Ishii, Virtual adversarial training: a regularization method for supervised and semi-supervised learning, IEEE Trans Pattern Anal Mach Intell 41 (8) (2018) 1979–1993.
[53] D.C. Dowson, B.V. Landau, The fréchet distance between multivariate normal distributions, J Multivar Anal 12 (3) (1982) 450–455.
[54] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 67–74.
[55] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process Lett 23 (10) (2016) 1499–1503.
[56] X. Mao, Q. Li, H. Xie, R.Y.K. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2794–2802.
[57] H. Robbins, S. Monro, A stochastic approximation method, The annals of mathematical statistics (1951) 400–407.
[58] Y. Bengio, Y. LeCun (Eds.), 3Rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7 deldDel- deliIns–9, 2015, Conference Track Proceedings, 2015. https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:accepted-main.html.
[59] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, volume 37, PMLR, Lille, France, 2015, pp. 448–456. http://proceedings.mlr.press/v37/ioffe15.html
[60] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, volume 30, 2013, p. 3.
[61] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[62] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.
[63] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

Qianfen JIAO received the M.S. degree from the Nanjing University, Nanjing, China, in 2010. She is currently working towards the Ph.D. Degree at the Department of Computer Science, City University of Hong Kong. Her research interests include domain adaptation and deep learning.

Rui LI is currently working towards the Ph.D. Degree at the Department of Computer Science, City University of Hong Kong. His research interests include domain adaptation and deep learning.

Wenming CAO received the M.S. degree in control engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2015, and the Ph.D. Degree in computer science from the City University of Hong Kong. His research interests include unsupervised learning and transfer learning.

Si WU received the PhD degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in 2013. He is an associate professor in the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include computer vision and pattern recognition.

Jian Zhong is currently working towards the Ph.D. Degree at the Department of Computer Science, City University of Hong Kong. His research interests include domain adaptation and deep learning.

Hau-San WONG received the BSc and MPhil degrees in electronic engineering from the Chinese University of Hong Kong, and the PhD degree in electrical and information engineering from the University of Sydney. He is currently an associate professor at the Department of Computer Science, City University of Hong Kong. He has also held research positions in the University of Sydney and Hong Kong Baptist University. His research interests include multimedia information processing, multimodal human-computer interaction, and machine learning.