Pattern Recognition 90 (2019) 99–108

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning a bi-level adversarial network with global and local perception for makeup-invariant face verification

Yi Li[a,b,d], Lingxiao Song[a,b], Xiang Wu[a,b], Ran He[a,b,c,d,*], Tieniu Tan[a,b,c]

[a] National Laboratory of Pattern Recognition, CASIA
[b] Center for Research on Intelligent Perception and Computing, CASIA
[c] Center for Excellence in Brain Science and Intelligence Technology, CAS
[d] University of Chinese Academy of Sciences

ARTICLE INFO

Article history:
Received 1 March 2018
Revised 22 October 2018
Accepted 7 January 2019
Available online 17 January 2019

Keywords:
Face verification
Makeup-invariant
Generative adversarial network

ABSTRACT

Makeup is widely used to improve facial attractiveness and is well accepted by the public. However, different makeup styles will result in significant facial appearance changes. It remains a challenging problem to match makeup and non-makeup face images. This paper proposes a learning from generation approach for makeup-invariant face verification by introducing a bi-level adversarial network (BLAN). To alleviate the negative effects from makeup, we first generate non-makeup images from makeup ones, and then use the synthesized non-makeup images for further verification. Specifically, there are two adversarial sub-networks on different levels in BLAN, with the one on pixel level for reconstructing appealing facial images and the other on feature level for preserving identity information. For the non-makeup image generation module, a two-path network that involves both global and local structures is applied to improve the synthesis quality. Moreover, we make the generator well constrained by incorporating multiple perceptual losses. All the modules are embedded in an end-to-end network and jointly reduce the sensing gap between makeup and non-makeup images. Experimental results on three benchmark makeup face datasets demonstrate that our method achieves state-of-the-art verification accuracy across makeup status and can produce photo-realistic non-makeup face images.

1. Introduction

Face verification focuses on the problem of making machines automatically determine whether a pair of face images refer to the same identity. As a fundamental research task, its development benefits various real-world applications, ranging from security surveillance to credit investigation. Over the past decades, massive face verification methods have achieved significant progress [1–7], especially the ones profiting by the recently raised deep networks. Nevertheless, there are still challenges remaining as bottlenecks in the real-world applications, such as pose [7], NIR-VIS [6] and makeup changes, which are often summarized as heterogeneous tasks. Due to the wide applications of facial cosmetics, the verification task of face images before and after makeup has drawn much attention in the computer vision society.

The history of cosmetics can be traced back to at least ancient Egypt [8]. Nowadays wearing makeup is well accepted in the daily life, and is even regarded as a basic courtesy on many important occasions. With appropriate cosmetic products, one can easily smooth skin, alter lip colour, change the shape of eyebrows, and accentuate eye regions. All these operations are often used to hide facial flaws and improve perceived attractiveness. But in the meanwhile, they also bring about remarkable facial appearance changes as exhibited in Fig. 1, resulting in both global and local appearance discrepancies between the images with and without makeup. Most of the existing face verification methods rely much on the various cues and information captured by the effective appearance features. These methods inherently lack robustness over the application of makeup that is non-permanent as well as miscellaneous. Recent study in [9] has claimed that the application of facial cosmetics decreases the performance of both commercial and academic face verification approaches significantly.

In contrast to the mentioned schemes, we consider from a new perspective and propose to settle the makeup-invariant face verification problem via a learning from generation framework. This framework simultaneously considers makeup removal and face verification, and is implemented by an end-to-end bi-level adversarial network (BLAN). It has the capacity of removing the cosmetics on a face image with makeup, namely synthesizing an appealing non-

* Corresponding author at: National Laboratory of Pattern Recognition, CASIA, Beijing 100190, China.
E-mail addresses: yi.li@cripac.ia.ac.cn (Y. Li), rhe@nlpr.ia.ac.cn (R. He).

https://doi.org/10.1016/j.patcog.2019.01.013
0031-3203/

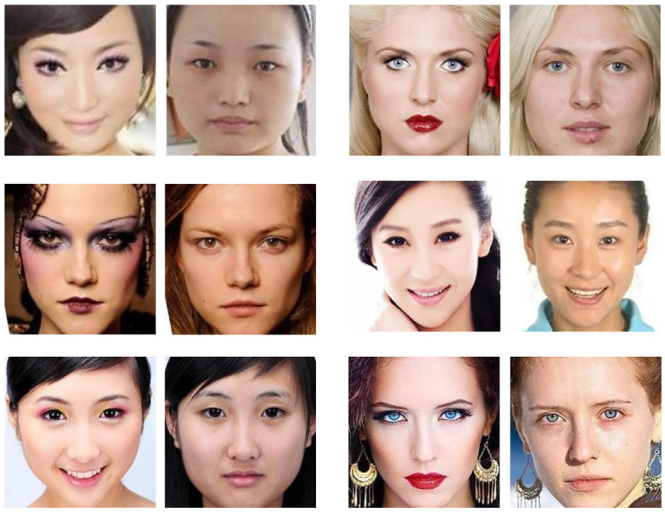


Fig. 1. Samples of facial images with (the first and the third columns) and without (the second and the fourth columns) the application of cosmetics. The significant discrepancy of the same identity can be observed.

makeup image with identity information preserved, effectively reducing the adverse impact of facial makeup. It promotes the verification performance of faces before and after makeup by imposing adversarial schemes on both pixel level and feature level.

Considering the variety and temporality characters of makeup, we first push the images to a uniform cosmetic status, the non-makeup status, by a Generative Adversarial Network (GAN) [10]. And then, deep features are extracted from the synthesized non-makeup faces for further verification task. As is illustrated in Fig. 2, the two steps above are not detached but integrated, for the adversarial loss on pixel level profits to generate perceptually better faces and the adversarial loss on feature level is employed to enhance the identity preservation. However, since the training data including makeup and non-makeup images are all collected on the Internet, it is impractical to criticize them to share the same pose, expression, occlusion, etc. To deal with the inherent unpair nature of training data and improve the synthesis quality, we introduce a two-path sub-network to the generator with one path considering global structure and the other embellishing local details, namely BLAN-2. Moreover, we also make the reconstruction well constrained via incorporating multiple priors such as symmetry and edges. Experiments are conducted on three makeup datasets and favorable results demonstrate the efficiency of our framework.

The major contributions of our work are as follows.

- We propose a learning from generation framework for makeup-invariant face verification. To the best of our knowledge, our framework is the first to account for the possibility of accomplishing the makeup-invariant verification task with synthesized faces.
- The bi-level adversarial network architecture is newly set up for our proposed framework. There are two adversarial schemes on different levels, with the one on pixel level contributing to reconstruct appealing face images and the other on feature level serving for identity maintenance. A two-path generator with multiple reconstruction losses are introduced to deal with the unpaired data and faithfully retain the characteristic facial structure of a certain individual.
- Our generative approach is able to extend the deep feature extraction modals (e.g., light CNN) to make-up problems without retraining or fine-tuning the underlying deep models. Experimental results on three datasets demonstrate that the verification performance is improved by generating non-makeup faces from makeup ones, verifying the efficacy of our framework.

This paper is an extension of our previous conference version [11]. The major differences between this paper and its previous version lie in two-folds: (1) to lessen the non-correspondence in some specific difficult areas between the synthesized non-makeup image and the real non-makeup image, we further introduce a two-path sub-network to the generator considering both global structure and local embellishing details. The new network are used to synthesize natural and realistic images not only in a whole but also in detail. (2) Apart from providing more indepth analysis, extensive additional experiments are carried out to validate the effectiveness of our proposed BLAN and BLAN-2, both quantitatively and qualitatively. Compared to existing anti-makeup methods and the original version of BLAN, our two-path BLAN network can effectively improve the visual performance in the difficult areas such as nose and mouth regions. Moreover, the face verification performance is further improved by BLAN-2.

The rest of the paper is organized as follows. In Section 2, we review some advanced related works in terms of face verification, makeup related studies and generative adversarial network. Section 3 describes the details of our proposed bi-level adversarial network for makeup-invariant face verification including generator architecture, generator losses and discriminator architecture. In Section 4, we present our experimental evaluations to verify the performance of our proposed method. We design extensive contrast experiments and report visual and face verification results on three benchmark datasets. Finally, we give our concluding remarks and future work in Section 5.

2. Related work

2.1. Face verification

As always, the face verification problem has attracted extensive attention and witnessed great progress. In traditional methods, He et al. [12] learned predictable binary codes to obtain fast query speeds as well as reduce storage requirements in face indexing and [13] presented a structured ordinal measure method in a data driven way to help face recognition. Recent impressive works are mostly based on deep networks [14–16]. Sun et al [1]. proposed a hybrid convolutional network - Restricted Boltzmann Machine (ConvNetRBM) model, which directly learns relational visual features from raw pixels of face pairs, for verification task in wild conditions. The Deepface architecture was expounded in [2] to effectively leverage a very large labeled dataset of faces for obtaining a representation with generalization. It also involved an alignment system based on explicit 3D modeling. The Deep IDentification-verification features (DeepID2) were learned in [3] which uses both identification and verification information as supervision. As various sensing conditions exist, heterogeneous face recognition enjoy broad research significance. Work [17] explored solutions by using data synthesis to map data from one modality into another modality and work [18] proposed a hierarchical hyperlingual-words (Hwords) and a distance metric through the hierarchical structure of Hwords to ease the similarity relationship measurement. With the further development of the face verification task, there are approaches customized for some certain conditions. For instance, Zhang et al. [5] aimed at facilitating the verification performance between the clean face images and the corrupted ID photos. Huang et al. [7] attempted to accomplish the recognition task of face images under a large pose. In this paper, we focus on the negative effects of the application of cosmetics over the verification systems, which is one of the most practical issue to be resolved in the real-world applications.

2.2. Makeup studies

Makeup related studies, such as makeup recommendation [19], have become more popular than ever. However, relatively less ar-

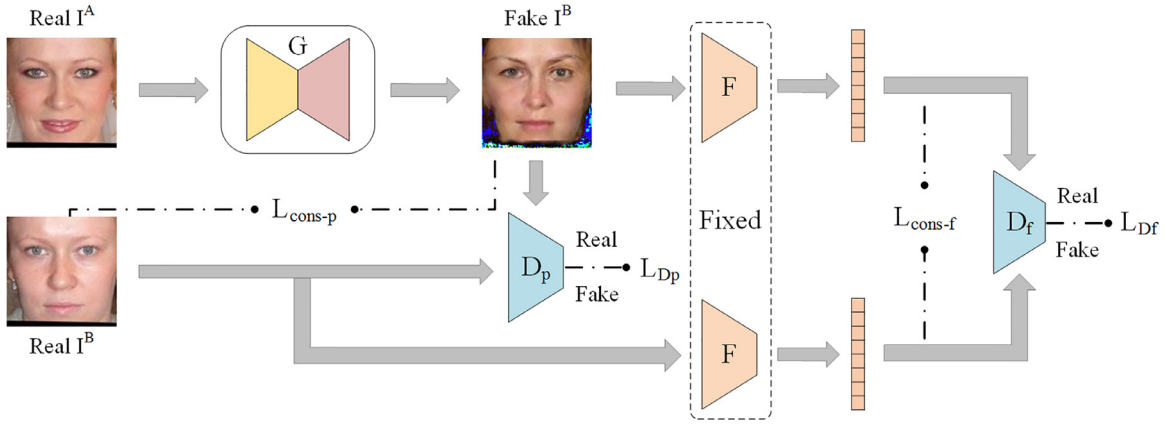


Fig. 2. Diagram of the proposed Bi-level Adversarial Network. I^A is an input image with makeup while I^B stands for the corresponding non-makeup image. The generator G learns to fool the two discriminators, where D_p is on the pixel level and D_f on the feature level. F represents an existing feature extractor that stays fixed during training and testing.

ticles pay attention on the challenge of makeup impact on face verification. Among these existing works, most of them contrive to design a feature scheme artificially to impel the pair images of the same identity to have the maximum correlation. To increase the similarity between face images of the same person, a meta subspace learning method was proposed in [20]. Guo et al. [21] explored the correlation mapping between makeup and non-makeup faces on features extracted from local patches. Chen et al. [22] introduced a patch-based ensemble learning method that uses subspaces generated by sampling patches from before and after makeup face images. A hierarchical feature learning framework was demonstrated in [23] that seeks for transformations of multi-level features. In addition, Convolutional Neural Network (CNN) based schemes have been recently developed. For example, [24] proposed to pre-train network on the free videos and fine-tune it on small makeup and non-makeup datasets.

2.3. Generative adversarial network

Contemporarily, GAN [10] is deemed as one of the most successful deep generative models and is applied in various vision related tasks (e.g., saliency detection [25], image style transition [26–28], image detection [29,30], image generation [7,31]). It corresponds to a min-max two-player game which ensures its ability of commendably estimating the target distribution and generating images that does not exist in the training set. Thereafter, multifariously modified GANs are explored, especially the ones in conditional settings [32]. The work in [27] investigated conditional adversarial networks as a solution to image-to-image translation problems. Zhu et al. [28] designed a new GAN framework with cycle constraint in it (CycleGAN) for unpaired image-to-image translation. Method in [30] proposed to learn an adversarial network that generates examples with occlusions and deformations to boost detection performance. A Two-Pathway Generative Adversarial Network (TP-GAN) [7] was established for photorealistic frontal view synthesis. Pathak et al. [33] proposed Context Encoders to cope with the image inpainting and Ledig et al. [34] applied GAN to super-resolution.

3. Bi-level adversarial network

To refrain from the influence induced by facial makeup, we propose to synthesize a non-makeup image I^B from a face image with makeup I^A first, via a generative network. And then, a deep feature is extracted from the synthesized I^B to further accomplish the verification task. We depict the overall structure of the proposed network in Fig. 2, with the details described below.

3.1. Notation and overview

The original GAN in [10] takes random noise as input and maps it to output images in domains such as MNIST. Different from it, we take images as input and set up our network as a conditional GAN. The generator denoted as G aims at learning a mapping from elements in domain A (with makeup) to elements in domain B (without makeup): $\mathbb{R}_A^{h\times w\times c} \rightarrow \mathbb{R}_B^{h\times w\times c}$, where the superscripts stand for the image size. If not constrained, the learned mapping can be arbitrary. Whereas, our network is tailored for further face verification application. And the two key intuitions are that the non-makeup facial image should be well synthesized and that the input and output of G should be identity invariant. We thus impose the constraint on G through introducing two adversarial discriminators on pixel level and feature level respectively.

During the training phase, image pairs $\{I^A, I^B\}$ with identity information y are required. Some existing conditional GANs based methods [7,27,33] have found that the generator is enhanced by adding a more traditional loss (e.g., L1 and L2 distances) to the GAN objective. The reason lies in that the generator is required to produce images close to the ground truth, not just to fool the discriminators in a conditional GAN. We thus enrich our training losses with some reconstruction items. Suppose that the training set consists of N training pairs, the generator G receives four kinds of losses for parameter updating: two reconstruction loss denoted by L_{cons-p} and L_{cons-f}, and two adversarial losses denoted by L_{D_p} and L_{D_f} in the Fig. 2. And the generator parameters are obtained by the solving the following optimization:

$$G^* = \frac{1}{N}\arg\min_G \sum_{n=1}^{N} L_{cons-p} + \lambda_1 L_{D_p} + \lambda_2 L_{cons-f} + \lambda_3 L_{D_f} \tag{1}$$

where the contributions of the losses are weighted by λ_1, λ_2 and λ_3. And the details of each loss will be discussed in the following section. As for both the discriminators, we apply the standard GAN discriminator loss formulated in Eqs. (2) and (3), since their duty of telling the fake from the real remains unchanged.

$$D_p^* = \arg\max_D \{\mathbb{E}_{I^B\sim p(I^B)} \log D(I^B) + \mathbb{E}_{I^A\sim p(I^A)} \log(1 - D(G(I^A)))\} \tag{2}$$

$$D_f^* = \arg\max_D \{\mathbb{E}_{I^B\sim p(I^B)} \log D(F(I^B)) + \mathbb{E}_{I^A\sim p(I^A)} \log(1 - D(F(G(I^A))))\} \tag{3}$$

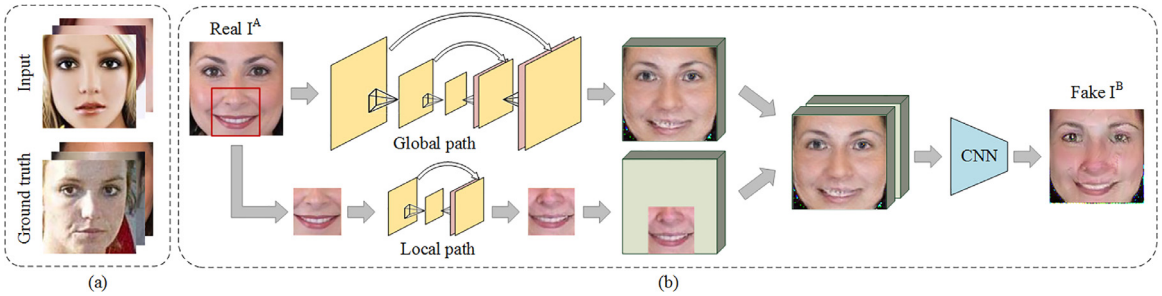


Fig. 3. Generator architecture. (a) illustrates the inherent unpair nature of training data. (b) describes the two-path network used to generate the non-makeup image from makeup images.

Here, the operation of $F(\cdot)$ represents the feature extraction. When training the network, we follow the behavior in [10] and alternately optimize the min-max problem described above. By this means, the generator is constantly driven to produce high-quality images that agree with the target distribution or the ground truth. Specifically, the synthesized non-makeup facial images from makeup ones will become more and more reliable and finally benefit the verification task.

3.2. Generator architecture

The generator in our proposed BLAN aims to learn a desirable mapping between facial images with and without makeup of the same person. An encoder-decoder network [35] can carry the duty out well and has been widely utilized in existing conditional GANs [7,33,36,37]. However, we notice an inherent property here in our task that the input and output of the generator are roughly aligned and share much of the information, both locally and globally. In this situation, a simple encoder-decoder network appears to be insufficient. The reason is that all the information in the input image has to go through the intermediate bottleneck whose size is usually much smaller than the input. This fact determines much of the low level priors captured by the first few layers would be abandoned before the bottleneck, thus makes the encoder-decoder network lack the ability to effectively take advantage of the low level information.

To address a similar problem in biomedical image segmentation, Ronneberger et al. [38] proposed an architecture named "U-net" to directly deliver context information to the corresponding layers with higher resolution, yielding the network shape of "U". Thereafter, Isola et al. [27] applied a semblable network to its generator for solving the image-to-image translation problem. Inspired by these works, we also adopt a network with skip connections to let the information acquired by the encoder benefit the output of decoder as much as possible. In specific, we follow the settings in [27] and concatenate the duplicate of layer i straight to layer $n-i$, with n denoting the total layer amount of the generator.

Although above-mentioned architecture can generate rather realistic synthesized non-makeup images, we find it cannot ensure that all the details are natural and close to the corresponding ones, especially in the regions of nose and mouth. A reasonable explanation is that the difficulty of transiting the facial areas are not in the same level, only taking the global structure into consideration will fail to well manage the difficult areas. Therefore, we employ a two-path architecture which recovers the global and local structures simultaneously, as shown in Fig. 3. Upon observation, we notice that the outline of nose is relatively blurry and the mouth region shows particularly more deformation between the synthesized non-makeup images and the real non-makeup images than other facial regions. Hence, it is necessary to add a local path for better transition details in these two local regions.

3.3. Generator losses

In the sections above, we have elaborated the overall structure and the generator architecture we employ. This part will focus on the four kinds of losses that the generator receive, which has been briefly described in Eq. (1). Besides the double adversarial losses, we also integrate various perceptual losses in L_{cons-p} to guarantee the quality of generated images. Particularly, the reconstruction loss L_{cons-p} is composed of three subordinates — a pixel-wise loss, a first-order loss and a symmetry loss, as is defined in Eq. (4). Since the local path shares the network structure with the global path (i.e., the "U-net" structure), we also train them with the same loss schemes. In the following, we will take the global path in generator as an example and discuss the losses in details. And for concision and convenience, the subscripts $*_{global}$ of all the generated results are omitted in the section. For instance, $G(I^A)$ stands for $G(I^A)_{global}$ in the method, which is the generated result of the global path in Fig. 3.

$$L_{cons-p} \triangleq L_{pxl} + L_{edg} + L_{sym}. \tag{4}$$

It has been mentioned that incorporating traditional losses helps to improve the outcome quality. There are generally two options for pixel wise loss — L1 distance or L2 distance. Since L1 distance is generally deemed to arouse less blur than L2 distance, we formulate the pixel-wise loss function as

$$L_{pxl} = \mathbb{E}_{(I^A,I^B)\sim p(I^A,I^B)} \|G(I^A) - I^B\|_1. \tag{5}$$

Given the paired data $\{I^A, I^B\}$, the pixel-wise loss continuously push the synthesized non-makeup facial image $G(I^A)$ to be as close to the ground truth I^B as possible. In our experiments, we also find that the pixel-wise loss helps to accelerate parameters convergence in some degree.

Although the pixel-wise loss in form of L1 distance would bring about blurry results, the adversarial scheme in GANs can alleviate it to some extent. However, this is based on the premise that there is adequate training data to learn a qualified discriminator, while the scale of existing makeup datasets are rather limited. To further cope with the blurring problem, we propose to train our network with the help of a first-order loss, which takes the form of

$$L_{edg} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} \Big\{ \Big\| |G(I^A)_{i,j} - G(I^A)_{i,j+1}| - |I^B_{i,j} - I^B_{i,j+1}| \Big\|_1 + \Big\| |G(I^A)_{i,j} - G(I^A)_{i+1,j}| - |I^B_{i,j} - I^B_{i+1,j}| \Big\|_1 \Big\}, \tag{6}$$

where $G(I^A)_{i,j}$ stands for the (i,j) pixel of the synthesized image $G(I^A)$. The first-order loss can also be referred as the edge loss, for it aims at fully explore the gradient priors provided in I^B. It actually needs to calculate the edges in images and then drives the edge image of the synthesized face to be close to the edge image of the ground truth.

As one of the most prominent characteristics of human faces, the symmetric structure is well exploited in many previous face related studies. Here in our network, we take it into consideration as well and imposes a symmetric constraint to guarantee the essential legitimacy of the synthesized face structure. The corresponding symmetry loss is calculated by

$$L_{sym} = \frac{1}{h \times w/2} \sum_{i=1}^{h} \sum_{j=1}^{w} \| G(I^A)_{i,j} - G(I^A)_{i,w-j+1} \|_1 \quad (7)$$

The responsibility of the discriminator on the pixel level is to distinguish real non-make facial images from the fake one and it serves as a supervision to produce relatively more pleasing synthesized results. Its corresponding adversarial loss on the generator is

$$L_{D_p} = \mathbb{E}_{(I^A) \sim p(I^A)}[-\log D_p(G(I^A))] \quad (8)$$

In addition to removing makeups, we also expect the synthesized images to facilitate the verification performance across makeup status. Since the verification task is accomplished on image features (e.g. Light CNN [39] feature in our experiments), the key issue is converted to produce images with high quality features, which is crucial for identity preserving. To this end, we propose to further cascade an adversarial network centering on the feature level at the end of the original conditional GAN model. The discriminator D_f is in charge of differentiating between features from real non-makeup images and fake ones, driving to synthesizing images with features close to the target. We formulate the adversarial loss on the feature level as

$$L_{D_f} = \mathbb{E}_{(I^A) \sim p(I^A)}[-\log D_f(F(G(I^A)))]. \quad (9)$$

Similar to the scheme on the pixel level, we incorporate a reconstruction loss with the adversarial loss which takes the following form:

$$L_{cons-f} = \mathbb{E}_{(I^A, I^B) \sim p(I^A, I^B)} \| F(G(I^A)) - F(I^B) \|_1. \quad (10)$$

3.4. Discriminator architecture

Inspired by the concepts in Conditional Random Field [40], we address an assumption on deciding whether the input of the discriminator D_p is real or fake: in a certain image, pixels that are apart from each other are relatively independent. Based on the assumption, we first divide an image into $k \times k$ patches without overlapping. And then the discriminator runs on each patch to obtain a score indicating whether this part of the image is real or not. Thus for each input image, the outcome of D_p is a probability map containing $k \times k$ elements. In our experiments, we empirically set $k = 2$. By this means, D_p is able to pay more attention to local regions instead of the whole image. Additionally, the operation simplifies the required structure of D_p and significantly reduces the parameter amount in the network, which is friendly to small datasets. As for the discriminator on the feature level (i.e. D_f), we concisely set it up with two linear layers, considering the conflict between the complexity of the BLAN structure and the fact of limited available training data.

4. Experiments and analysis

We evaluate our proposed method on three makeup datasets. In this following, we use "BLAN" to indicate the original version without two-path architecture in the generator and use "BLAN-2" represents the improved network. Both visualized results of synthesized non-makeup images and quantitative verification performance are present in this section. Furthermore, we explore the effects of all losses and report them in the ablation studies. The overall results demonstrate that our framework is able to achieve state-of-the-art verification accuracy across makeup status, with appealing identity-preserved non-makeup images synthesized from the ones with makeup.

Table 1
Data comparison of the three makeup datasets.

Dataset	Individual Num.	Female Num.	Male Num.
Dataset 1	501	501	0
Dataset 2	203	203	0
FAM	519	297	222

4.1. Datasets

Dataset 1: This dataset is collected in [21] and contains 1002 face images of 501 female individuals. For each individual, there are two facial images — one with makeup and the other without. The females span mainly over Asian and Caucasian descents. *Dataset 2:* Assembled in [24], there are 203 pairs of images with and without makeup, each pair corresponding to a female individual. *Dataset 3 (FAM)* [20]: Different from the other two datasets, FAM involves 222 males and 297 females, with 1038 images belonging to 519 subjects in total. We compare the data distribution of the three datasets in Table 1. It is worthy noticing that all these images are not acquired under a controlled condition for they are collected from the Internet. Thus there also exist pose changes, expression variations, occlusion and other noises in these datasets except for makeup alteration. Some sample images from the three datasets are showed in Fig. 4.

Following the settings in [20,21,24], we adopt five-fold cross validation in our experiments. In each round, we use about 4/5 paired data for training and the rest 1/5 for testing, no overlap between training set and testing set. All the positive pairs are involved in the testing phase and equal pairs of negative samples are randomly selected. Hence, taking Dataset 1 as an example, there are about 100 pairs of faces for testing each time. We report the average rank-1 accuracy and true positive rate over the five folds as quantitative evaluation.

4.2. Implementation details

In our experiments, all the input images are resized to $128 \times 128 \times 3$ and the generator output synthetic images of the same size. BLAN is composed of a generator G, two discriminator D_p and D_f, and a feature extractor Light CNN. The Light CNN used for feature extracting is pre-trained on MS-Celeb-1M [41] without fine-tuning on makeup datasets. The global path of G in BLAN-2 is an encoder-decoder network with U-Net structure and consists of 8×2 Convolution-BatchNorm-ReLU layers. It contains about 41, 833 k parameters and about 5.6 G FLOPS. The local path has the same network structure as the global path, but with smaller scale. There are about 10,549k parameters in the local path sub-network. As for the fusion sub-network, we apply the ResNet [42] architecture with one block. D_p is a network with 4 convolution layers followed by a Sigmoid function. It contains about 667k parameters and 1.1G FLOPS. D_f is made of 2 fc layers and contains about 26k parameters. We accomplish our network on PyTorch [43]. It takes about 3 h to train BLAN and about 4 hours to train BLAN-2 on Dataset 1, with a learning rate of 10^{-4}. Data augmentation of mirroring images is also adopted in the training phase. Considering the limited number of images in Dataset 2, we first train BLAN and BLAN-2 on Dataset 1 and then fine-tune them on Dataset 2 in our experiments. As for the loss weights, we empirically set $\lambda_1 = 3 \times 10^{-3}$, $\lambda_2 = 0.02$ and $\lambda_3 = 3 \times 10^{-3}$. In particular, we also set a weight of 0.1 to the edge loss and 0.3 to the symmetry loss inside L_{cons-p}.

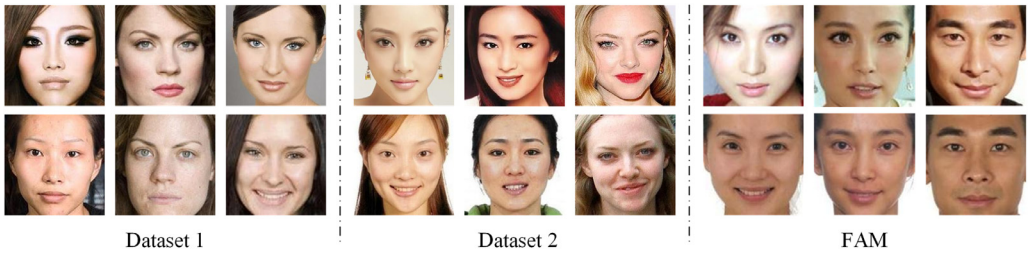


Fig. 4. Sample image pairs of three datasets.

Table 2
Rank-1 accuracy (%) on three makeup datasets.

Dataset	Method	Accuracy
Dataset 1	[21]	80.5
	[24]	82.4
	DenseNet-121	87.6
	ResNet-50	88.0
	ResNet-101	89.4
	VGG	89.4
	Light CNN	92.4
	BLAN	94.8
	BLAN-2	95.5
Dataset 2	[24]	68.0
	DenseNet-121	81.0
	ResNet-50	79.0
	ResNet-101	77.0
	VGG	86.0
	Light CNN	91.5
	BLAN	92.3
	BLAN-2	93.4
FAM	[44]	59.6
	[20]	62.4
	DenseNet-121	74.8
	ResNet-50	75.3
	ResNet-101	76.5
	VGG	81.6
	Light CNN	86.3
	BLAN	88.1
	BLAN-2	90.0

Table 3
True Positive Rate (%) on three makeup datasets.

Method	Dataset	TPR@FPR=0.1%	TPR@FPR=1%
BLAN	Dataset 1	65.9	99.8
	Dataset 2	38.9	82.7
	FAM	52.6	97.0
BLAN-2	Dataset 1	68.2	99.5
	Dataset 2	44.6	84.2
	FAM	53.8	94.9

4.3. Comparisons with existing methods

The ultimate goal of our proposed BLAN is to facilitate face verification performance across makeup status by generating non-makeup facial images. We demonstrate the effectiveness of BLAN by conducting verification task on the mentioned three makeup datasets. The results on VGG [45], ResNet [42], DenseNet [46] and Light CNN [39] serve as baselines. Particularly, we adopt VGG-16, ResNet-50, ResNet-101, DenseNet-121 and Light CNN without any fine-tuning on the makeup datasets. For all these networks, We use the public released model for test. In these experiments, we extract deep features from images with and without makeup via the corresponding networks and directly use them for matching evaluation. While in the BLAN and BLAN-2 experiments, a non-makeup image is first produced by the generator for each makeup image. Then the generated non-makeup image is sent to Light CNN for deep feature extraction. It should be noted that our method is actually accomplishing verification task on synthetic images, which is of significant progress.

We compare the rank-1 verification accuracy with some existing methods in Table 2 and report the true positive rate in Table 3. The similarity metric used in all experiments is cosine distance. Except for the mentioned baselines, the methods listed are all tailored for makeup-invariant face verification. Among them, the works in [20,21,44] explore the correlation between images of a certain identity with and without makeup in traditional ways, while the approach in [24] is based on deep networks.

From Table 2, we can observe that our proposed BLAN brings prominent improvement to rank-1 accuracy comparing with existing makeup-invariant schemes, both traditional and deep ones. In specific, a boost of at least 10% is achieved on each dataset. It demonstrates that our architecture is able to achieve state-of-the-art performance on the datasets. Although ResNet-50, ResNet-101 and DenseNet-121 are well known for their strong representation abilities, their verification performances on three makeup datasets are inferior to Light CNN, BLAN and BLAN-2. The reason may lie in that: technically, ResNet and DenseNet are designed for image classification rather than embedding learning (e.g., face related task). Additionally, it is worth noticing that VGG, ResNet, DenseNet and Light CNN are all trained on much larger datasets than the makeup datasets. Their produced deep features are thus rather powerful, resulting in much higher accuracies than the traditional schemes. Comparing the feature extraction processes in BLAN and in Light CNN, one can find that the only difference is the input. Even though, our network still outperforms the baselines. These phenomena consistently validate that our learning from generation framework has the ability of promote verification performance by alleviating impact from makeup.

Compared with the results of BLAN, we observe that BLAN-2 achieves slightly better performance on most occasions. Notice the major difference of these two network is the generator structure. The aim of the local path is to improve the visualized quality of synthesized non-makeup images. Nevertheless, the results in Tables 2 and 3 demonstrates that learning the generator with both global and local perception is benefit to face verification task. It also proves that the improved BLAN-2 is able to generate non-makeup face images with better quality, at least on the feature level.

4.4. Synthetic non-makeup images

For the existing makeup-invariant face verification methods we discussed, none of them has the capacity of generating non-makeup images from that with makeup. In contrast to them, we propose to extract deep features directly from synthetic non-makeup images for face verification. To evaluate our BLAN perceptually, we exhibit some synthetic samples in Fig. 5 along with results from Pix2pix [27] and Makeup-Go [47]. Noticing Pix2pix is

Fig. 5. Comparisons of synthetic non-makeup images from various methods.

not tailored for makeup-related tasks, we train the networks from scratch on three makeup datasets using public released codes provided by the authors. It is observed that our methods, including BLAN and BLAN-2, achieve better synthesized non-makeup images. The reason is that pix2pix merely adopted a pixel-wise reconstruction loss and an adversarial loss as supervision, but did not take the face priors and identity information into account like our work. For Makeup-Go, we use both the network and the model provided by the authors without fine-tuning in our experiments. Since Makeup-Go deals with portrait images, we input the original images without cropping for this method. The major difference between our work and Makeup-Go is that they are trying to solve disparate problems on makeup. Our work intends to settle makeup-invariant face verification by removing the makeup on faces, while Makeup-Go aims to restore portrait images from operations of skin smoothing and skin color change brought by camera or image processing Apps without considering verification task. Also, Makeup-Go proposes Component Regression Network (CRN) to accomplish the task of blind reverse of unknown beautification process. Yet we propose a bi-level adversarial network (BLAN) to alleviate the negative effects from makeup on face verification. Instead of removing the makeup, Makeup-Go changes the skin color and the texture on faces. Observing the BLAN and BLAN-2 rows, we can find that both holistic face structure and most local attributes of the original faces are kept when removing makeup. The developed BLAN-2 achieves better synthesized non-makeup quality than the others. The reason is that in addition to the discriminator on pixel level, we propose to impose another discriminator on feature level to maintain the identity prior as well as facial structure.

Different makeup datasets have different characteristics. Dataset 1 and Dataset 2 only contain female subjects and the paired images have higher resolution compared with FAM. Thus, BLAN achieves perceptually better synthetic images and results in higher verification accuracy on the first two datasets. In contrast, more than 40% of the subjects are male in FAM. To evaluate our BLAN perceptually, we show both male and female results of BLAN (both with and without two-path architecture) in Fig. 5. The images in the fourth rows are generated by the original BLAN, i.e., the version without two-path architecture. We can find that these images keep holistic face structure as well as some local attributes of the original faces in a certain extend. These results benefit from not only the discriminator on pixel level, but also the another imposed discriminator on feature level for maintaining the identity prior as well as facial structure. The results produced by the BLAN-2 are listed in the fifth row. As we can observe, the synthesized non-makeup faces appear to be more locally natural and realistic compared to the images in the BLAN row and are more close to the ground truth in addition to maintain the global structure. This fact proves the effectiveness of local path in achieving local optimization.

On the other hand, the makeup removing results of males are not so satisfied as that of females. For male individuals, the gap between makeup images and non-makeup ones is relatively narrower than that of females and the training data of males is much less than females, which are determined by the fact that males trend to wear less makeup in reality. In addition, we also notice that there exist blurs in our synthetic non-makeup images compared with ground truth. And the clarity of facial component outlines, e.g. eyes contour, is not so compelling as expected. The reasons lie in multiple folds. (1) In reconstruction loss on pixel level, we adopt L1 distance. It has been reported in [27] and [7] that L1 distance loss in generator will bring about image blurs for it leads to overly smooth results. Even though there are adversarial networks, the overly smooth problem can not be swept away. (2) We merely utilize the data from the three makeup datasets to train BLAN, without any help from other data. Compared with other face related datasets, the data sizes of these makeup datasets are rather limited. It consequently decreases the training quality of the network. (3) As has been introduced in Section Datasets, all the paired images are collected from the Internet. In other words, the images are not acquired under a controlled condition. Even the facial key points are not strictly aligned as standard facial datasets. We present some images pairs with pose, expression and occlusion changes and their synthetic non-makeup results in Fig. 6. Our

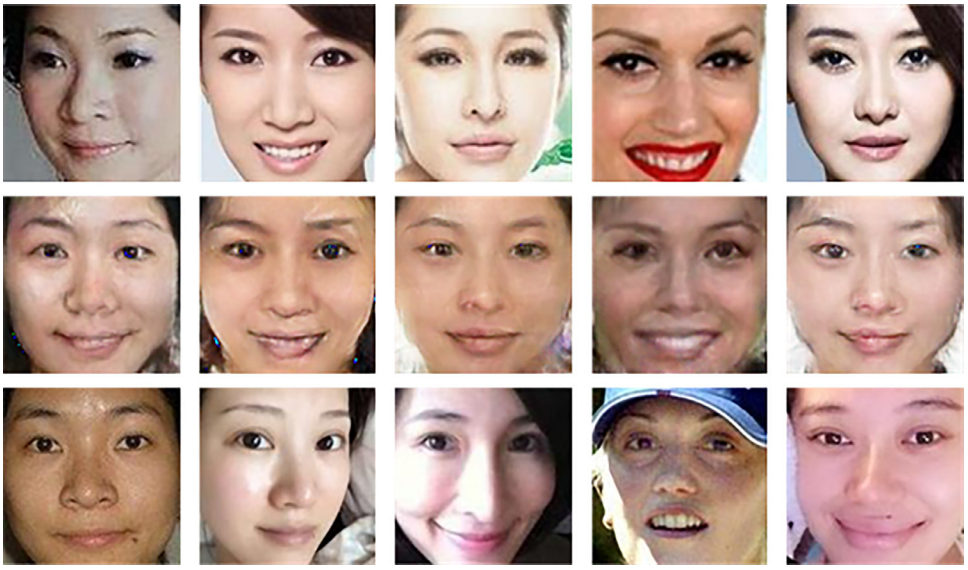


Fig. 6. Sample results with pose, expression and occlusion changes. From top to down, there are input makeup images, synthetic non-makeup images of BLAN and ground truth, respectively.

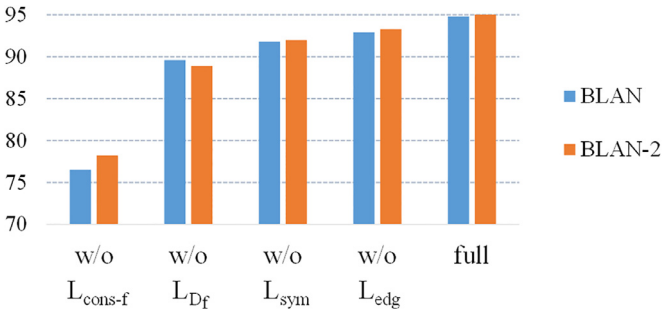


Fig. 7. Rank-1 accuracy (%) on Dataset 1 with ablation.

methods are data-driven, i.e., the generated output is determined by the training data and the input. These changes will severely hinder the network training and thus impact the generated image quality. However, different training losses help to deal with such ambiguity. In this work, we focus on the negative effect induced by makeups and dedicate to addressing the makeup-invariant face verification problem. Accordingly, we adopt reconstruction losses and adversarial losses on different levels to ensure the makeup removal quality and preserve the identity information. As is well-known, a disadvantage of GAN is that it is difficult to control the output properly. Therefore, how to freely handle the ambiguity of our generated image remains a future work direction for us.

4.5. Ablations

To fully explore the contribution of each loss, we conduct experiments on different architecture variants of BLAN and BLAN-2. The quantitative verification results are presented in Fig. 7 for comprehensive comparison. We remove one of the losses in generator training each time and examine the corresponding accuracy change. As expected, both BLAN and BLAN-2 with all the losses achieves the best accuracy. It is evident that L_{cons-f} and L_{D_f} bring the greatest declines, indicating the effectiveness and importance of adversarial network on feature level especially in the aspect of identity preserving. As for L_{edg} and L_{sym}, they also help to promote the performance, though not as much remarkable as the fore discussed two losses. Therefore, we can conclude that all the four losses contribute to improve the final performance. On the other hand, the Rank-1 accuracies of BLAN-2 and its variants outperform those of BLAN in most conditions, which indicates the validness of the added local path.

The corresponding visual results of BLAN and BLAN-2 listed in Fig. 8 intuitively present the effectiveness of each loss. The generated images without the edge loss and the symmetry loss tend to suffer from more unnatural artifacts. And the absence of adversarial loss on feature level causes serve blur to the synthesized results. Finally, L_{cons-f} contributes most to the identity preservation, as can be distinctly observed by comparing the last three rows in Fig. 8. We can see that similar tendencies of the absence of each loss are shared between BLAN and BLAN-2. On one hand, for both BLAN and BLAN-2, the best synthesized results are achieved by the full model. On the other hand, we observe that the makeup removal quality from BLAN-2 is better than BLAN in general, especially around the nose and mouth area.

5. Conclusion

In this paper, we have proposed a new learning from generation framework to address the makeup problem in face verification. A synthesized non-makeup image is generated with its identity prior well preserved from a makeup image. And then, the produced non-makeup images are used for face verification, which effectively bypasses the negative impact incurred by cosmetics. Specifically, we have proposed a novel architecture, named bi-level adversarial network (BLAN). To further improve the quality of our synthesized images, a two-path generator considering both global structure and local details has been introduced. Extensive experiments on three makeup datasets show that our network not only generates pleasing non-makeup images but also achieves state-of-the-art verification accuracy under makeup conditions. However, there are aspects to be improved. Our methods are data-driven and the generated image quality depends much on the training data. The current network structure is difficult to control the output ambiguity. Hence, for the future research, we raise two directions. (1) Both the data quality and quantity are of great importance to the development of makeup-related researches. Therefore, we contrive to establish a new makeup dataset with high quality images. (2) We are considering to apply the learning from generation framework to settling other problems in face verification, e.g. large pose, extreme illu-

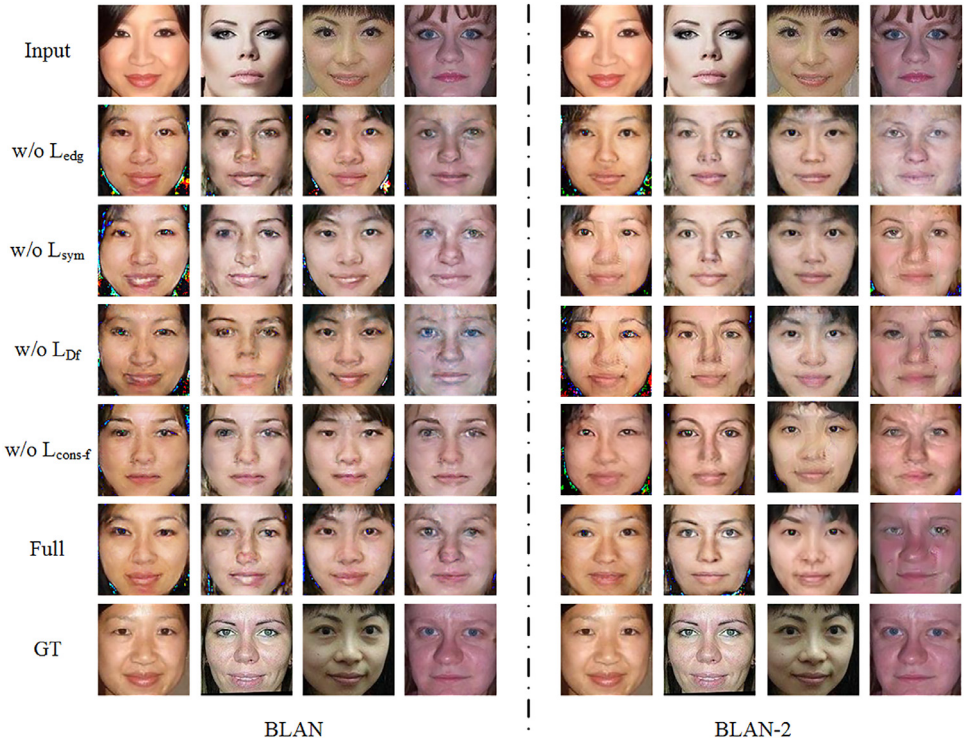


Fig. 8. Synthetic results of BLAN with its variants on the left column, and results of BLAN-2 with its variants on the right.

mination and facial disguise, and try to handle the corresponding diversity of the generated image.

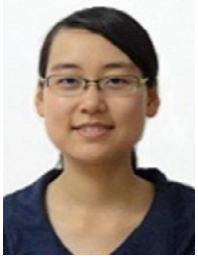
Acknowledgments

This work was funded by State Key Development Program (Grant no. 2016YFB1001001), and National Natural Science Foundation of China (Grant nos. 61622310, 61473289).

References

[1] Y. Sun, X. Wang, X. Tang, Hybrid deep learning for face verification, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1489–1496.
[2] 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.
[3] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 1988–1996.
[4] X.-Y. Jing, F. Wu, X. Zhu, X. Dong, F. Ma, Z. Li, Multi-spectral low-rank structured dictionary learning for face recognition, Pattern Recognit. 59 (2016) 14–25.
[5] S. Zhang, R. He, Z. Sun, T. Tan, Multi-task convnet for blind face inpainting with application to face verification, in: Proceedings of the International Conference on Biometrics, 2016, pp. 1–8.
[6] R. He, X. Wu, Z. Sun, T. Tan, Learning invariant deep representation for nir-vis face recognition, in: Proceeding of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press, 2017, pp. 2000–2006.
[7] R. Huang, S. Zhang, T. Li, R. He, Beyond face rotation: Global and local perception Gan for photorealistic and identity preserving frontal view synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2439–2448.
[8] B. Burlando, L. Verotta, L. Cornara, E. Bottini-Massa, Herbal Principles in Cosmetics: Properties and Mechanisms of Action, CRC Press, 2010.
[9] A. Dantcheva, C. Chen, A. Ross, Can facial cosmetics affect the matching accuracy of face recognition systems? in: Proceedings of the Fifth International Conference on Biometrics: Theory, Applications and Systems, IEEE, 2012, pp. 391–398.
[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Advances in neural information processing systems, 2014, pp. 2672–2680.
[11] Y. Li, L. Song, X. Wu, R. He, T. Tan, Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[12] R. He, Y. Cai, T. Tan, L. Davis, Learning predictable binary codes for face indexing, Pattern Recognit. 48 (10) (2015) 3160–3168.
[13] R. He, T. Tan, L. Davis, Z. Sun, Learning structured ordinal measures for video based face recognition, Pattern Recognit. (2017).
[14] M. Kan, S. Shan, H. Chang, X. Chen, Stacked progressive auto-encoders (spae) for face recognition across poses, in: Proceedings of the Computer Vision and Pattern Recognition, 2014, pp. 1883–1890.
[15] T. Luan, X. Yin, X. Liu, Disentangled representation learning Gan for pose-invariant face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1283–1292.
[16] X. Zhu, Z. Lei, J. Yan, D. Yi, S.Z. Li, High-fidelity pose and expression normalization for face recognition in the wild, IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 787–796.
[17] J. Lezama, Q. Qiu, G. Sapiro, Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 6807–6816.
[18] M. Shao, Y. Fu, Cross-modality feature learning through generic hierarchical hyperlingual-words, IEEE Trans. Neural Netw. Learn. Syst. PP (99) (2017) 1–13.
[19] T. Alashkar, S. Jiang, S. Wang, Y. Fu, Examples-rules guided deep neural network for makeup recommendation, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press, 2017, pp. 941–947.
[20] J. Hu, Y. Ge, J. Lu, X. Feng, Makeup-robust face verification, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 2342–2346.
[21] G. Guo, L. Wen, S. Yan, Face authentication with makeup changes, IEEE Trans. Circuits Syst. Video Technol. 24 (5) (2014) 814–825.
[22] C. Chen, A. Dantcheva, A. Ross, An ensemble of patch-based subspaces for makeup-robust face recognition, Inf. Fus. 32 (2016) 80–92.
[23] Z. Zheng, C. Kambhamettu, Multi-level feature learning for face recognition under makeup changes, in: Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition, 2017, pp. 918–923.
[24] Y. Sun, L. Ren, Z. Wei, B. Liu, Y. Zhai, S. Liu, A weakly supervised method for makeup-invariant face verification, Pattern Recognit. 66 (2017) 153–159.
[25] X. Hu, X. Zhao, K. Huang, T. Tan, Adversarial learning based saliency detection, in: Proceedings of the 4th Asian Conference on Pattern Recognition, 2017.
[26] J.Y. Zhu, W.S. Zheng, J.H. Lai, S.Z. Li, Matching NIR face to vis face using transduction, IEEE Trans. Inf. Forensi. Secur. 9 (3) (2017) 501–514.
[27] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 5967–5976.
[28] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, IEEE International Conference on Computer Vision, 2017.
[29] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, Perceptual generative adversarial networks for small object detection, IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1951–1959.

[30] X. Wang, A. Shrivastava, A. Gupta, A-Fast-RCNN: Hard positive generation via adversary for object detection, IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[31] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[32] M. Mirza, S. Osindero, Conditional generative adversarial nets, Comput. Sci. (2014) 2672–2680.
[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
[34] 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, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 105–114.
[35] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[36] X. Wang, A. Gupta, Generative image modeling using style and structure adversarial networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 318–335.
[37] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, International Conference on Machine Learning, 2017, pp. 1857–1865.
[38] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[39] X. Wu, R. He, Z. Sun, T. Tan, A light cnn for deep face representation with noisy labels, IEEE Trans. Inf. Forens. Secur. 13 (11) (2018) 2884–2896.
[40] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the 18th International Conference on Machine Learning, vol. 1, 2001, pp. 282–289.
[41] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-celeb-1m: a dataset and benchmark for large-scale face recognition, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 87–102.
[42] 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.
[43] A. Paszke, S. Gross, S. Chintala, Pytorch, 2017. https://github.com/pytorch/pytorch.
[44] H.V. Nguyen, L. Bai, Cosine similarity metric learning for face verification, in: Proceedings of the Asian Conference on Computer Vision, Springer, 2010, pp. 709–720.
[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the 3rd International Conference on Learning Representations, 2015.
[46] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2017, p. 3.
[47] Y.-C. Chen, X. Shen, J. Jia, Makeup-go: Blind reversion of portrait edit, in: Proceedings of the IEEE International Conference on Computer Vision, vol. 2, 2017.



Yi Li received the B.E. degree in Electronic and Information Engineering from Dalian University of Technology in 2014, and the M.E. degree in Information and Communication Engineering from Dalian University of Technology in 2017. She is currently a Ph.D. student in the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), CASIA, Beijing, China. Her current research interests include computer vision and pattern recognition.

Lingxiao Song received the B.E. degree in Automation from Zhejiang University (ZJU), Hangzhou, China, and the M.S. degree in Computer Application Technology from Chinese Academy of Sciences (CASIA), Beijing, China, in 2013 and 2016 respectively. She is currently an assistant engineer with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), CASIA, Beijing, China. Her research interests include biometrics, pattern recognition, computer vision, and machine learning.

Xiang Wu received the B.E. degree in Electronic Engineering from University of Science and Technology Beijing in 2013, the M.S. degree in Electronic Engineering from University of Science and Technology Beijing in 2016. He is a research assistant in Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests focus on deep learning, computer vision and biometric recognition.

Ran He received the B.E. degree in Computer Science from Dalian University of Technology, the M.S. degree in Computer Science from Dalian University of Technology, and Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences in 2001, 2004 and 2009, respectively. Since September 2010, Dr. He has joined NLPR where he is currently Professor. He currently serves as an associate editor of Neurocomputing (Elsevier) and serves on the program committee of several conferences. His research interests focus on information theoretic learning, pattern recognition, and computer vision.

Tieniu Tan received his B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and his M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. He is currently a Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, China. His current research interests include biometrics, image and video understanding, and information forensics and security.

He is a Fellow of the IEEE and the IAPR.