This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3044187, IEEE Access



# On Open-Set, High-Fidelity and Identity-Specific Face Transformation

**Longhao Zhang[1], Xipeng Pan[2,3*], Huihua Yang[1,2] and Lingqiao Li[2]**
[1]Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2]Guilin University of Electronic Technology, Guangxi, 541004, China
[3]Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, 510080, China
*Equal contribution

Corresponding author: Huihua Yang (e-mail: yhh@bupt.edu.cn).

This research was supported in part by the National Key R&D Program (Grant No.2018AAA0102600), National Natural Science Foundation of China (Grant No. 62002082, 61906050), Guangxi Natural Science Foundation (Grant No. 2020GXNSFBA238014)

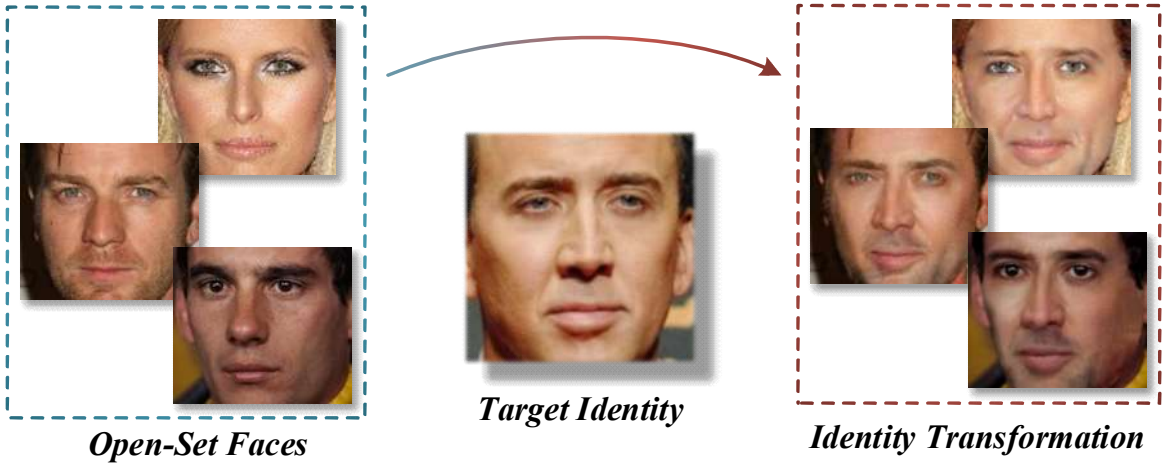


**FIGURE 1.** Target-Specific Face Identity Transformation.

**ABSTRACT** In this paper, a Generative Adversarial Networks-based framework has been proposed for identity-specific face transformation with high fidelity in open domains. Specifically, for any face, the proposed framework can transform its identity to the target identity, while preserving attributes and details (e.g., pose, gender, age, facial expression, skin tone, illumination and background). To this end, an auto-encoder network is adopted to learn the transformation mapping, which encodes the source image into the latent representation, and reconstruct it with the target identity. In addition, the face parsing pyramid is introduced to help the decoder restore the attributes. Moreover, a novel perceptual constraint is applied to the transformed images to guarantee the correct change of the desired identity and to help retrieve the lost details during face identity transformation. Extensive experiments and comparisons to several open-source approaches demonstrate the efficacy of the proposed framework: it can achieve more realistic identity transformation while better preserving attributes and details.

**INDEX TERMS** Auto-encoder, Face transformation, Generative Adversarial Networks, Perceptual constraint

## I. INTRODUCTION

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3044187, IEEE Access



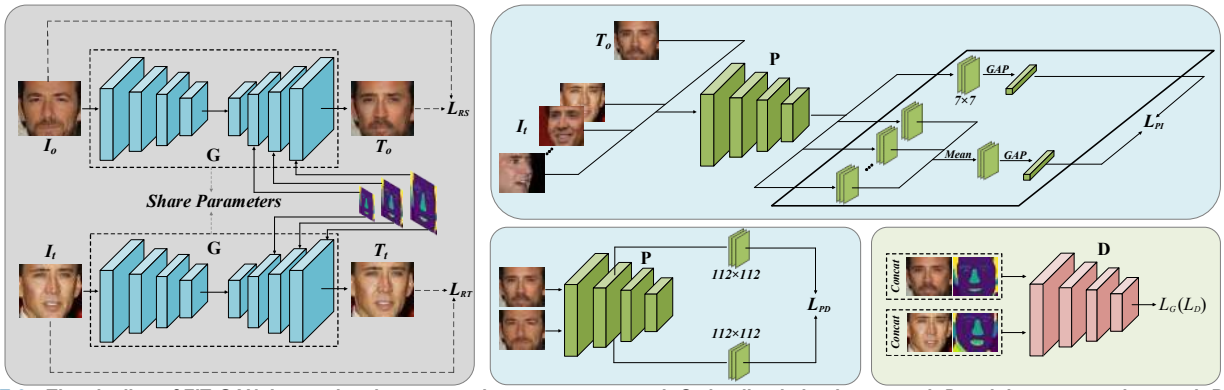


**FIGURE 2.** The pipeline of FIT-GAN. It contains three parts: the generator network *G*, the discriminative network *D* and the perceptual network *P*. *G* takes the source face *I* and face parsing masks as input and output the transformed face *T*. Subscript *o* indicates the face is from an open set, and subscript *t* means the face has the target identity. $T_o$ and a batch of $I_t$ are sent to *P* to calculate the perceptual identity loss $L_{PI}$, while $T_o$ and $I_o$ are sent to *P* to calculate the perceptual detail loss $L_{PD}$. *D* is used to calculate the GAN loss, whose input is the concatenation of transformed face *T* and the face parsing mask

In recent years, with the emergence of deep generative models, such as the Generative Adversarial Networks (GAN) [1] and the Variational Auto-encoder (VAE) [2, 3], researchers have made tremendous progress in building deep networks for image generation [38]. Among them, face transformation is a critical task, owing to its wide real-world applications. It can be divided into two branches: face attribute transformation and face identity transformation.

Many approaches have obtained significant advances in face attribute transformation. For example, TP-GAN [4] and FF-GAN [5] aim to synthesize the frontal view of a face from a single face image. DR-GAN [6] can take one or multiple face images as the input and jointly learn a pose-invariant representation and perform face frontalization. These methods successfully transform the pose of the face while preserving its identity. Other methods are designed to edit the face attributes like age, gender, hair color, expression and so forth. Star-GAN [7] trains a conditional attribute transformation network via attribute classification loss and cycle consistency loss for multiple face attribute editing. Att-GAN [8] shares some similar ideas with Star-GAN but does not include a cyclic process or cycle consistency loss. It focuses on the disadvantages of existing methods [9, 10] on modeling the relationship between the latent representation and the attributes.

Instead of manipulating face attributes, face identity transformation changes the face of a person with the face of another person. As one of the most well-known open-source face identity transformation methods, DeepFakes [11] is based on two auto-encoders with a shared encoder that is trained to reconstruct training images of the source and the target face, respectively. To change the identity of the source image, DeepFakes first detects and crops the face region. Then, the trained encoder and decoder of the target identity are applied to it. Finally, the output is blended with the rest of the source image using Poisson image editing [12] to preserve the details. However, DeepFakes has some limitations: 1) each training model can only achieve the transformation between two fixed identities, but cannot be used for open set; 2) fail to well preserve face attributes such as skin color, age, gender, etc., especially when they differ too much between the source and the target. There are some other face identity transformation methods based on the GAN and the VAE, like CVAE-GAN [13], which proposes a general learning framework that combines a variational auto-encoder with a generative adversarial network under a conditioned generative process. Mutual transformation of multiple identities can be achieved through CVAE-GAN, since it establishes an identity-independent latent representation for further identity editing. However, the encoder of CVAE-GAN only outputs the mean vector and the covariance vector, and the latent vector used for reconstruction is sampled from a random normal distribution. Such an approach leads to the information loss, causes it to fail to preserve attributes and details of the source face.

In this paper, a GAN-based framework is proposed towards open-set and high-fidelity face identity transformation. In particular, for any face from the open set, our framework can transform it into a face with a specific identity. High-fidelity implies that the transformed face retains other attributes and details of the source face as much as possible, except for the identity change. To this end, an encoder-decoder network is trained to learn the transformation mapping between the source face and the reconstructed face with the target identity. For better preservation of attributes, the face parsing pyramid is applied to different levels of the decoder as the prior. Besides, a novel perceptual constraint is proposed to guarantee the correct transformation of the desired identity and to help retrieve lost image details during image reconstruction.

As shown in Fig. 2, our approach consists of three parts: 1) a generator network *G*, which encodes the source image to the latent representation and reconstructs it with the target identity; 2) a discriminator network *D*, which distinguishes real or fake images; 3) a perceptual network *P*, whose

parameters are frozen, to constrain the identity of the reconstructed face and to retain details of the source image.

Extensive experiments and comparisons to several open-source methods demonstrate the efficacy of the proposed framework. It is trained on CelebA [14] and tested on CelebA and Face Scrub [15]. Results show that whether the source face and its identity are included in the training dataset or not, our approach can achieve realistic identity transformation while preserving attributes and details.

## II. RELATED WORK

**Auto-encoder (AE)** is widely used in semi-supervised learning and unsupervised learning, which compressing the input into a latent representation, and then reconstructing this representation into an output. It was first proposed to solve the "encoder problem" in representation learning in 1985 [16], and is now widely used in image generation [37]. Variational Auto-encoder (VAE) [17, 18] is a classic model of AE, it adds the constraints to the encoder, forcing it to generate the latent representation that obeys the Gaussian unit distribution. The disadvantage of using VAE in face transformation task is that, because the encoder only outputs the mean vector and the covariance vector, and the latent vector used for reconstruction is sampled from a random normal distribution, there is a lot of face attributes and image details lost. Therefore, traditional AE architecture is adopted instead of VAE in this paper.

**Generative Adversarial Networks (GAN)** can learn to generate realistic images. It simultaneously trains two networks: a generator network to generate samples, and a discriminator network to differentiate between natural and generated samples. It effectively solves the problem that the generated images of AE are often blurry. However, the GAN is hard to converge in the training stage and the samples generated from GAN are often far from natural. Recently, many works have been proposed to improve the quality of the generated samples and make training stable. For example, the Wasserstein GAN (W-GAN) [19] uses Earth Mover Distance as an objective for training GAN. It improves the stability of learning and gets rid of problems like mode collapse. The Least Squares GAN (LS-GAN) [20] changes the sigmoid cross-entropy loss function commonly used in the regular GAN which might lead to the vanishing gradient problem during the learning process to the least-squares loss function, significantly improving the quality of generated images and the stability of training. This paper adopts the loss function used in [20].

**Face transformation** is an image generation task that allows us to edit the face identity or face attributes in the source image. [39] proposes a novel attributes encoder for extracting multi-level target face attributes, and a new generator with carefully designed Adaptive Attentional Denormalization (AAD) layers to change the identity to the target while preserving attributes. Unlike two-player GANs, [40] generates identity-preserving faces by proposing FaceID-GAN, which treats a classifier of face identity as the third player, competing with the generator by distinguishing the identities of the real and synthesized faces. It generates faces of arbitrary viewpoint while preserving identity.

**Perceptual constraint** appears in several recent papers, depending on high-level features extracted from a convolutional network. Images can be generated to maximize class prediction scores [21] or individual features [22] in order to understand the functions encoded in trained networks. [23] optimizes the perceptual loss function to train a feed-forward network for image style transfer. Experiments prove that it can also be used for face transformation. In this paper, a novel perceptual constraint is proposed to guarantee the correct transformation of the desired identity and to help retrieve lost image details during image reconstruction.

## III. FACE INDENTITY TRANSFORMATION GAN

In this section, we introduce the proposed Face Identity Transformation GAN (FIT-GAN). As shown in Fig. 2, our proposed framework contains three parts: 1) the generator network $G$; 2) the discriminative network $D$; and 3) the perceptual network $P$.

$G$ is an encoder-decoder network to learn the transformation mapping which is identity-specific and attribute-preserving. Whether target identity faces and open-set faces share its encoder or decoder during training. For better preservation of the facial attributes, the face parsing pyramid is introduced since it contains rich identity-invariant facial information (e.g., the position of facial features, facial poses, and even facial expressions). To form the face parsing pyramid, the face parsing mask of $I$ is first obtained through a pre-trained face parsing model, and then scaled to several resolutions. After that, they are applied it to different levels of the decoder through the SPADE Resblock proposed in [24]. Besides, the widely-used self-attention mechanism [25, 35, 36] is adopted to further improve the quality of the generated image. The transformed face $T$ and the face parsing mask $M$ are concatenated and sent to $D$, which learns to distinguish between real and fake samples. To guarantee that the identity of the source face from the open-set can be correctly transformed into the target identity, the transformed face $T$ is constrained using a perceptual network $P$ whose parameters are frozen. Moreover, $P$ can also help retrieve lost details during image reconstruction.

At the heart of our proposed framework lies three loss functions: 1) the reconstruction loss $L_R$, 2) the GAN loss $L_G$ ($L_D$), and 3) the perceptual loss $L_P$. In the following sections, we describe them at length and provide the training details.

### A. RECONSTRUCTION LOSS

The reconstruction loss $L_R$ is used to constrain the similarity between the transformed image $T$ and the source image $I$. In this paper, the pixel-wise $L2$ loss is adopted as the reconstruction loss, the same as [11] and [13].

In the training phase, the generator network $G$ is shared by target identity faces and open-set faces, so there are two situations: whether the identity of the input face is the target identity or not. If the identity of the input face image $I_o$ is from open set (not in the target identity set), $G$ is expected to output a face that transforming the identity to the target identity while preserving other attributes and details. When the input $I_t$ is from the target identity set, $G$ is expected to restore the $I_t$, that is, the input and output are hoped to be exactly the same, just like the traditional AE. Therefore, different loss weights are employed in two situations. Formally, the reconstruction loss is

$$L_R = \begin{cases} L_{RS} = \frac{1}{2}\|I_o - T_o\|_2^2 \\ L_{RT} = \frac{\lambda_1}{2}\|I_t - T_t\|_2^2 \end{cases} \quad (1)$$

where $L_{RS}$ corresponds to the first situation and LRT to the second. λ1 is the reconstruction loss weight.

In practice, eyeballs and mouth details are identity-irrelevant but crucial for facial expression. Therefore, in order to restore the facial expression vividly when changing the identity, the eyeball and the inner region of the mouth are extracted with the help of the face parsing mask, and calculate the reconstruction loss separately from the face, as shown in Fig. 3. The location-aware reconstruction loss can be written as

$$L_R = \begin{cases} L_{RS} = L_{face} + \lambda_2 L_{eyeball} + \lambda_2 L_{mouth}, \quad where \quad L_{region} = \frac{1}{2}\|I_o^r - T_o^r\|_2^2 \\ L_{RT} = \frac{\lambda_1}{2}\|I_t - T_t\|_2^2 \end{cases} \quad (2)$$

where a higher loss weight $\lambda_2$ is assigned to the eyeball and

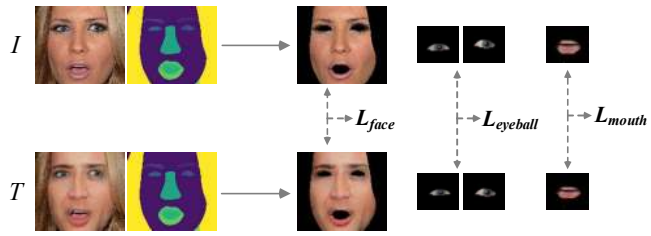


**FIGURE 3. Location-aware reconstruction loss. *I* is the source image and *T* is the transformed image. We use face parsing mask to precisely segment the face, eyeball and mouth, and calculate the reconstruction loss $L_{face}$, $L_{eyeball}$ and $L_{mouth}$ respectively.**

mouth. The loss of each part is calculated in the same way as in (1), except that *L2* distance is calculated respectively in the corresponding region.

### B. GAN LOSS

One of the defects of AE is that the generated image is blurry. However, the use of the GAN loss can solve this problem, synthesizing more realistic images. For the design of the GAN loss, we refer to [20]. Similar to the original GAN, the generator network $G$ competes in a two-player minimax game with the discriminator network $D$. $D$ tries to distinguish real training images from synthesized images, while $G$ tries to fool $D$. The difference is that it uses the least-squares loss instead of the sigmoid cross-entropy loss. It significantly improves the quality of generated images and the stability of training. Concretely, when training $D$, it tries to minimize the loss function

$$L_D = \frac{1}{2}E_{I\sim P_r}\left[D(I)-a\right]^2 + \frac{1}{2}E_{I\sim P_r}\left[D(G(I))-b\right]^2 \quad (3)$$

where $a = 1$ and $b = 0$. When training $G$, it minimizes the loss function

$$L_G = \frac{1}{2}E_{I\sim P_r}\left[D(G(I))-c\right]^2 \quad (4)$$

where $c = 1$.

### C. PERCEPTUAL LOSS

In this paper, a novel perceptual constraint on the transformed face $T$ is proposed to guarantee the correct change of the desired identity and to help retrieve lost details during image reconstruction. To this end, the perceptual network $P$ is introduced. It is pre-trained on VGGFace2 [26], which includes numerous identities and covers a wide range of postures, ages and races. It should be noted that $P$ is frozen during training, and the perceptual loss $L_P$ is used to update the parameters of $G$. $L_P$ consists of two parts: the perceptual identity loss $L_{PI}$ and the perceptual detail loss $L_{PD}$.

The perceptual identity loss $L_{PI}$ utilizes the target identity to constrain the identity of the transformed face. Because the input source image comes from the open set, its identity category may not exist in VGGFace2. Therefore, the classification results of the face recognition network $P$ cannot be used to constrain the identity. However, since $P$ is a face recognition network, as its layers getting deeper, the features become more and more identity-relevant and attribute-irrelevant. To prove that, we select two images $A_1$ and $A_2$ with the same identity and different postures, and image $B$ with the same posture but different identity as $A_2$. Then they are sent into $P$ to extract features at different levels. By comparing their cosine distance, it can be seen that in the deep layer of the network, the feature distance of images with

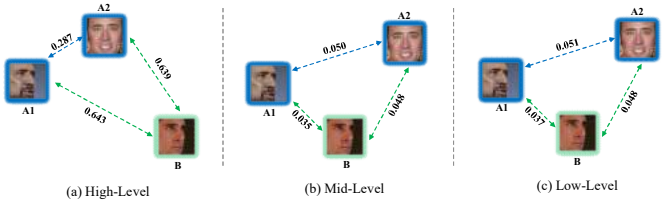


**FIGURE 4. The distance of features in different levels of *P*. High-level indicates features of the last convolutional layer.**

the same identity is far less than that of images with similar postures but different identities. Details are shown in Fig. 4. Therefore, the identity of the transformed face is constrained by feature distance from the faces with the target identity. Specifically, the transformed face $T_o$ and a set of face images with the target identity $\{I_t^1, I_t^2, \ldots I_t^n\}$ are input into $P$ and the high-level feature $F_T$ and $\{F_I^1, F_I^2, \ldots F_I^n\}$ are obtained from

the last convolution layer of the last Resblock. $\{F_I^1, F_I^2, \ldots F_I^n\}$ are then elementwise averaged to get $F_I$. After that, the global average pooling (GAP) is applied on $F_T$ and $F_I$ to get the spatial-independent identity vector $V_T$ and $V_I$. Their cosine distance is used as the perceptual identity loss. Formally, $L_{PI}$ is

$$L_{PI} = 1 - \frac{V_I \cdot V_T}{\|V_I\|_2 \|V_T\|_2} \tag{5}$$

The experiment shows that in the process of identity transformation, especially when applying the perceptual identity constraint, the transformed face loses some details of the source face, such as texture and illumination. To retrieve these lost details, the perceptual detail loss $L_{PD}$ is proposed, referring to the approaches of the image style transfer task [23, 27]. Concretely, the source face $I_o$ and the transformed face $T_o$ are input into the perceptual network $P$, and get the low-level feature $F_I'$ and $F_T'$ from the last convolution layer of the first Resblock. For each of them, its Gram Matrix $GM \in R^{N\times N}$ is calculated where $N$ is the number of channels and $GM_{ij}$ is the inner product between the feature map $i$ and $j$:

$$GM_{ij} = \frac{1}{CHW} \sum_k F_{ik}' F_{jk}' \tag{6}$$

Mid-level's and low-level's features are extracted from the shallow layers.

Then the perceptual detail loss $L_{PD}$ is

$$L_{PD} = \sum_{i,j} \left( GM_{ij}^I - GM_{ij}^T \right)^2 \tag{7}$$

Moreover, we also try to combine features of multiple layers, it only marginally improves the ability of retrieving the details.

### D. OVERALL LOSS FUNCTION AND TRANING DETAIL

The overall loss function is the sum of all the losses defined above:

$$\begin{aligned} L &= L_{RS} + \lambda_1 L_{RT} + \lambda_3 L_G + \lambda_3 L_D + \lambda_4 L_{PI} + \lambda_5 L_{PD} \\ &= L_{face} + \lambda_2 L_{eyeball} + \lambda_2 L_{mouth} + \lambda_1 L_{RT} + \lambda_3 L_G + \lambda_3 L_D + \lambda_4 L_{PI} + \lambda_5 L_{PD} \end{aligned} \tag{8}$$

where we empirically set $\lambda_1$=$\lambda_2$=10, $\lambda_3$=1, $\lambda_4$=0.3 and $\lambda_5$=1 in the experiment.

(8) is optimized using Adam [28] with $\beta_1$=0.5, $\beta_2$=0.999. The learning rate is set to 0.00001. The size of the input face image for training is 128×128, which has been detected, cropped, and aligned using MTCNN [29]. As mentioned above, the face parsing mask pyramid is introduced to help preserve face attributes, due to its identity independence. To achieve this, the SPADE Resblock proposed in [24] is employed. In particular, the face parsing mask of the source input image is first obtained through a pre-trained face parsing model. Then it is scaled to 128×128, 64×64, 32×32, 16×16 and 8×8, and sent to different levels of the decoder. In the training phase, we separate each iteration into two steps: at the first step, we use images with target identity to train the network, which means only images from the target identity set are sampled to form the mini-batch, whose batch size is set to 64 in the experiments; at the second step, we only sample images from the rest training set except the target identity set and use them to train the network. In practice, it is proved that this two-steps training strategy can not only improve the training stability and convergence speed, but also make the generated image more realistic. The details of the training strategy are described in Algorithm 1.

**Algorithm 1** The two-step training strategy.

**Require:** $\theta_G$, initial $G$ network parameters. $\theta_D$, initial $D$ network parameters. $\theta_P$, initial $P$ network parameters. $\lambda_1$=$\lambda_2$=10, $\lambda_3$=1, $\lambda_4$=0.3 and $\lambda_5$=1. *iter*=1

**While** $\theta_G$ not converged **do**

  **if** *iter*%2 = 1 **then**

    Sample $I_t$ form the target identity set

    $L_R = 1/2\| I_t - G(I_t)\|^2$

    $L_D = 1/2(D(G(I_t)) - 1)^2$

    $L_G = 1/2(D(I_t) - 1)^2 + 1/2D(G(I_t))^2$

    $\theta_G \leftarrow -\nabla_{\theta_G} (\lambda_1 L_R + \lambda_3 L_G)$

    $\theta_D \leftarrow -\nabla_{\theta_D} (\lambda_3 L_D)$

  **else**

    Sample $I_o$ form the rest training set except the target identity set

    $L_R = L_{face} + \lambda_2 L_{eyeball} + \lambda_2 L_{mouth}$

    $L_D = 1/2(D(G(I_o)) - 1)^2$

    $L_G = 1/2(D(I_o) - 1)^2 + 1/2D(G(I_o))^2$

    $V_I = P(I_t)$, $V_T = P(G(I_o))$

    $L_{PI} = 1 - (V_I V_T / \|V_I\|_2 \|V_T\|_2)$

    Compute $GM_I$ and $GM_T$

    $L_{PD} = \sum ( GM_I - GM_T)^2$

    $L_P = \lambda_4 L_{PI} + \lambda_5 L_{PD}$

    $\theta_G \leftarrow -\nabla_{\theta_G} ( L_{face} + \lambda_2 L_{eyeball} + \lambda_2 L_{mouth} + \lambda_3 L_G + \lambda_4 L_{PI} + \lambda_5 L_{PD})$

    $\theta_D \leftarrow -\nabla_{\theta_D} (\lambda_3 L_D)$

  **end if**

  *iter* ← *iter* + 1

**end while**

## IV. EXPERIMENTS

In this section, we use experiments to validate the effectiveness of the proposed FIT-GAN. The network is trained on CelebA [14] and evaluated on CelebA and Face Scrub [15]. CelebA is a large-scale face attributes dataset. It has large diversities, large quantities, and rich annotations, including: 1) 10,177 number of identities, 2) 202,599 number of face images and 3) 40 binary attributes annotations per image. However, in our experiment, the above identity annotations and attribute annotations are not used. Instead, the perceptual loss and reconstruction loss are utilized to constrain the network to transform identity and preserve face attributes. CelebA is chosen for training because images in this dataset cover large variations in posture, age, skin color, facial expression, etc., which significantly improves the generalization performance of the network. Face Scrub is a

dataset with over 100,000 face images of 530 people, which is the testing dataset in our experiments.

The generator network $G$ is an auto-encoder, consisting of a down-sampling encoder network and an up-sampling decoder network. In this paper, the self-attention block and SPADE Resblock are introduced to $G$ to improve the quality of the generated image and to better retain face attributes. The discriminative network $D$ is a binary classification network, whose input is the concatenation of the transformed face and the face parsing mask, referring to [24]. For the perceptual network $P$, the ResNet [30] pre-trained on VGGFace2 is adopted. More details of the network architecture can be found in appendix A.

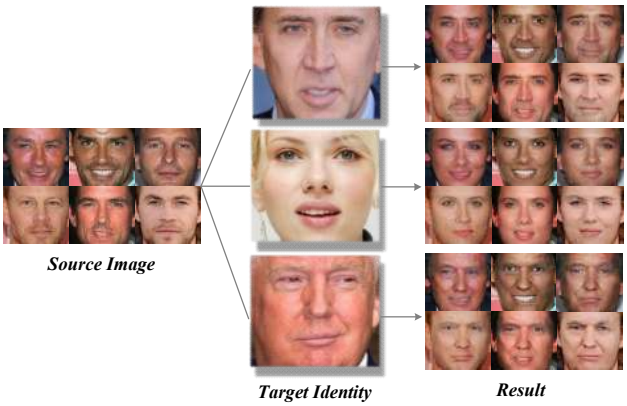


**FIGURE 5.** Face identity transformation results on training set.

### A. FACE INDENTITY TRANSFORMATION of FIT-GAN

This section presents the results of the open-set, high-fidelity and target-specific face identity transformation using our proposed FIT-GAN. The goal of the transformation is to change the identity of the source face from open set to the specific target identity, while preserving other face attributes and details.

Fig. 5 shows the transformation result which the source faces are from CelebA. As can be seen from the figure, our proposed FIT-GAN can successfully transform the identity of the source face of different ages, races and genders into the target identity. Meanwhile, it can also effectively preserve the face attributes such as skin color, posture, facial expression and details such as illumination and background in the source image.

To further prove that the proposed method can achieve open-set face identity transformation, we choose the source image from Face Scrub, which does not exist in the training dataset CelebA, including its identity. Fig. 6 presents the results.

### B. VISUALISATION COMPARISION with OTHER FRANWORKS

In this section, we visually compare the proposed FIT-GAN with other open-source frameworks for face identity transformation. They are analyzed and compared in the following aspects: 1) whether they can achieve realistic identity transformation; 2) whether they can preserve face attributes and details; and 3) the performance on open-set faces. Therefore, we specially select source images with large poses and extreme facial expressions. Besides, open-set faces that do not exist in the training set are tested.

Visualization comparison is shown in Fig. 7. As can be seen from the figure, whether the source faces exist in the training set or not, CVAE-GAN can edit the identity accurately. However, most of the face attributes and details in the source images are lost, especially when dealing with

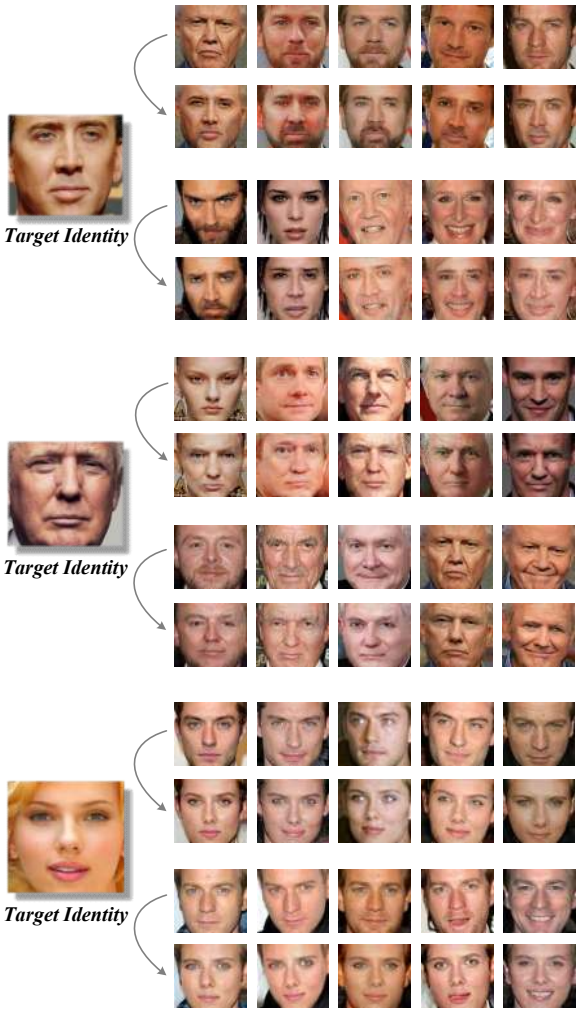


**FIGURE 6.** Face identity transformation results on open set.


(a) From Training Set (b)From Open Set

**FIGURE 7.** Visualization comparison with existing open-source frameworks on both training set and open set.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3044187, IEEE Access



large poses, such as the 1st, 3rd, 4th, and 6th column. In addition, CVAE-GAN cannot retain facial expressions like a surprise, smile and frown in the source image, such as the 5th, 7th, 9th, and 10th column. Besides, other attributes and details in the source image, such as skin color, mouth opening or closing, eye gaze, are also not preserved. Compared with CVAE-GAN, DeepFakes can preserve more attributes and details, but it cannot achieve realistic identity transformation when facing the source images from the open set, which can be seen in the last seven columns of the figure. Furthermore, DeepFakes sometimes fail to generate faces when face poses are large, like the 3rd column. Thanks to the introduced face parsing mask and the perceptual constraint, FIT-GAN can achieve identity transformation under large poses or extreme facial expressions, while preserving attributes and details such as skin color, illumination and background.

### C. QUANTITATIVE COMPARISION with OTHER FRAMEWORKS

In this paper, several criteria are utilized for the following quantitative comparisons: 1) whether the transformed identity is the target identity; 2) whether face attributes and details are preserved during transformation; and 3) the photorealism of output images.

In this work, we use the *Top-1 accuracy* to measure the identity transformation ability. We check whether the identity category of the generated image can be accurately predicted. To this end, an Inception-V3 [31] face recognition network is trained on the test set Face Scrub. Besides, the *ID retrieval* is also adopted as [39] does. We extract identity vector using a different face recognition network [41] and adopt the cosine similarity to measure the identity distance. For each generated face from the test set, we search the nearest face in Face Scrub and check whether it belongs to the correct person. The averaged accuracy of all such retrievals is reported as the *ID retrieval* in Table 1. Attribute and detail preservation is measured by a variant of perceptual distance [32], called the domain-invariant perceptual distance (*DIPD*) [33]. To compute the *DIPD*, the feature of VGG conv5 layer is extracted from the input source image as well as from the output transformed image. Then, the instance normalization is applied to the features, which will remove their mean and variance. It can filter out much identity-specific information in the features [34] and focus on the identity-invariant similarity. The *DIPD* is given by *L2* distance between the instance normalized features. Moreover, we use the open-source pose estimator [42] and 3D face model [43] to estimate *head pose* and *expression* preservation as [39] does. We report the *L2* distances of pose and expression vectors between the generated face and the source face. To quantify the photorealism of output images, the structural similarity

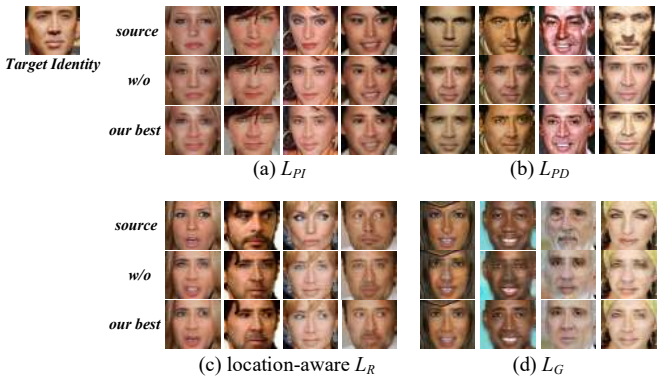


**FIGURE 8. Face identity transformation results with or without the specific loss.**

(*SSIM*) is calculated. *SSIM* measures the similarity between the transformed image and the input image in terms of brightness, contrast and structure. The higher the SSIM value, the lower the distortion of the transformed image.

Tab. 1 shows the quantitative results of the generated image quality of different frameworks. As can be seen from the table, our FIT-GAN is superior to other approaches in terms of identity transformation, attribute and detail preservation, and the photorealism of the generated image.

### D. ABLATION STUDY

As described above, at the heart of our proposed framework lies the composite loss function $L$, which can be written as (8). To understand the effects of each loss component, we repeat the training of FIT-GAN with the same settings but using a different combination of losses and compare the quality of the transformed images. Specifically, we independently remove the perceptual identity loss $L_{PI}$, the perceptual detail loss $L_{PD}$, the GAN loss $L_G$, and change the location-aware reconstruction loss $L_R$ to the non-location-aware one. The results are shown in Fig. 8. From which we can see that removing $L_{PI}$ will cause the framework unable to transform the identity accurately. Removing $L_{PD}$ will lose some face attributes and details of source images like skin tone, texture and illumination during face transformation. Not using the weighted location-aware $L_R$ will cause the framework unable to preserve facial expressions due to the loss of eyeballs, mouth and other details. Lastly, if $L_G$ is removed, the generated images will be blurry.

TABLE I
QUANTITATIVE COMPARISON WITH OTHER OPEN-SOURCE FRAMEWORKS

| | Top-1 acc↑ | ID retrieval↑ | DIPD↓ | Head pose↓ | Expression↓ | SSIM↑ |
|---|---|---|---|---|---|---|
| CVAE-GAN [13] | **0.86** | 0.46 | 1.45 | 5.08 | 3.13 | 0.47 |
| DeepFakes [11] | 0.53 | 0.77 | 0.61 | 3.79 | 2.70 | 0.59 |
| Ours | 0.83 | **0.89** | **0.34** | **2.20** | **2.11** | **0.68** |

↑ means larger numbers are better, ↓ means smaller numbers are better.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3044187, IEEE Access



TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT TRAINING STRATEGIES

| Training strategy | Top-1 acc↑ | ID retrieval↑ | DIPD↓ | Head pose↓ | Expression↓ | SSIM↑ |
|---|---|---|---|---|---|---|
| many-to-one | 0.83 | 0.89 | 0.34 | 2.20 | **2.11** | 0.68 |
| many-to-one + fine-tune | 0.82 | 0.89 | 0.36 | 2.27 | 2.29 | 0.61 |
| many-to-k | **0.85** | **0.91** | **0.32** | **2.11** | 2.13 | **0.72** |
| many-to-one + unfreeze P | 0.82 | 0.89 | 0.35 | 2.21 | 2.13 | 0.68 |

In this work, we propose a *many-to-one* face identity transformation framework, which means all source identities share the same encoder but each target identity corresponds to a decoder. Although it is not necessary to train a pair of encoders and decoders independently for each pair of source and target identities, as in a *one-to-one* framework like DeepFakes, it is still less practical when facing the multiple target identities. Therefore, we propose the following strategies that can speed up and simplify training for multi-target task. The quantitative results can be seen in Tab. 2.

**Train one, fine-tune the others.** First, we select a target identity to train the framework for 100 epochs. Then, for other target identities, we only need to fine-tune the decoder on each corresponding training set for a couple of epochs with the parameters of encoder and discriminator frozen. This training strategy can significantly improve the training speed but the quantitative result is slightly lower than our baseline.

**Single-encoder, multi-decoder training.** We change our framework from *many-to-one* to *many-to-k*. Specifically, we train $k$ decoders at once, as mentioned in [44]. The architecture is shown in Fig. 9. In our experiments, we set $k$ to 4 and reduce the batch size and network complexity

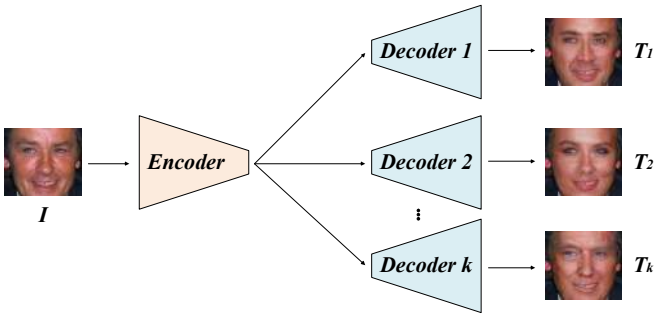


FIGURE 9. **Single-encoder, multi-decoder training.**

accordingly. Experiment results reveal that it not only speeds up the training, but also boosts the network performance from all aspects.

Moreover, we also try to unfreeze the perceptual network $P$ and update its parameters during training. Through experiments we find that when its learning rate is set very low, the result will be slightly lower than baseline, which is shown in Tab. 2. As the learning rate increases, performance continues to deteriorate.

## V. CONCLUSION

In this paper, a novel framework called FIT-GAN is proposed for open-set, high-fidelity and target-specific face identity transformation. In particular, an encoder-decoder network is trained to learn the transformation mapping between the source face and the transformed face with the target identity. For better preservation of attributes, the face parsing pyramid is applied to different levels of the decoder as the prior. Besides, a novel perceptual constraint is proposed to guarantee the correct transformation of the desired identity and to help retrieve lost image details during image reconstruction. Extensive experiments and comparisons to several existing methods demonstrate the efficacy of FIT-GAN: it can achieve more realistic identity transformation while better preserving attributes and details. However, there is still a flaw in our work, that is, although the identity in the source image can come from the open domain, the model can only transform it to the specific target rather than an arbitrary identity. Our future work will explore how to transform the source identity to the target identity which is also in the open domain.

## APPENDIX

### A. NETWORK ARCHITECTURE

The generator network $G$ is an auto-encoder, consisting of a down-sampling encoder network and an up-sampling decoder network. In this paper, the self-attention block and SPADE Resblock are introduced to $G$ to improve the quality of the generated image and better preserve face attributes, which are shown in Fig. A.1. Details of the architecture of $G$ is shown in Fig. A.2 (a), where ↑ means 2-times up-sampling, ↓ means 2-times down-sampling, and IN is Instance Normalization. The discriminative network $D$ is a binary classification network, whose input is the concatenation of the transformed face and the face parsing mask, which is shown in Fig. A.2 (b). For the perceptual network $P$, we adopt the ResNet-101 [30] pre-trained on VGG-Face2. The architecture is illustrated via the following chain of operations:

*Conv-64* → *Max Pool* → *Resblock-256* → *Resblock-512* → *Resblock-1024* → *Resblock-2048* → *Average Pool* → *FC-N* → *Softmax*

where $N$ is the number of classes.

### B. MORE FACE IDENTITY TRANSFORMATION RESULTS

The proposed FIT-GAN can achieve high-fidelity face identity transformation, which means face attributes can be well preserved while editing the identity. The visualization results in Fig. B reveal that our method can successfully

achieve identity transformation in the face of different age, gender, skin tone and even extreme facial expressions, while preserving face attributes and details.

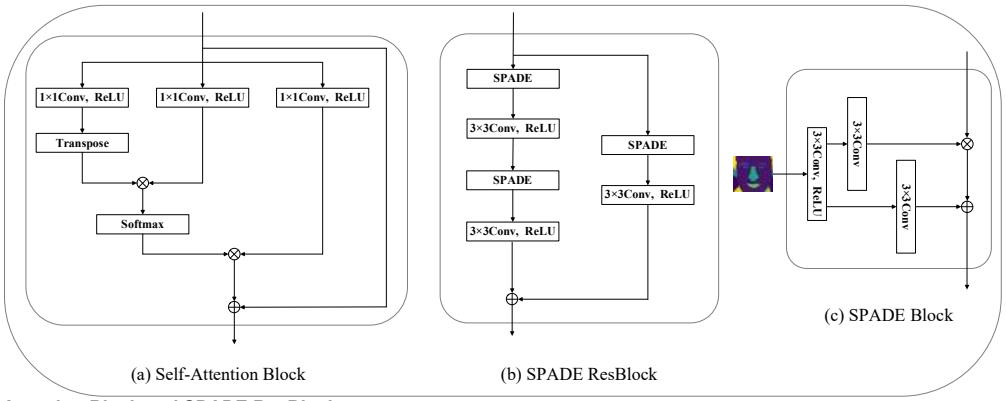


**FIGURE A.1.** **Self-Attention Block and SPADE ResBlock.**

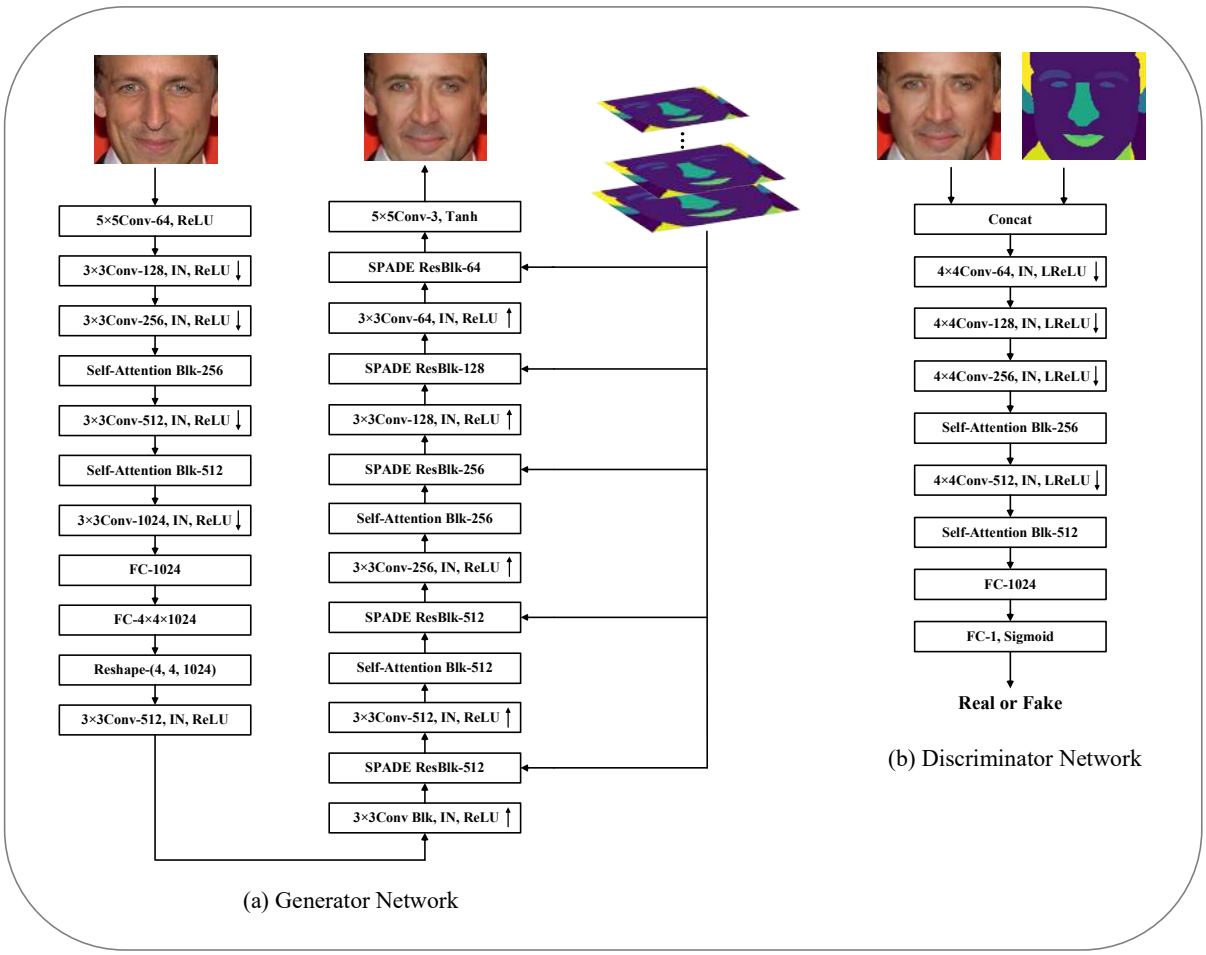


**FIGURE A.2.** **Network Architecture.**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3044187, IEEE Access



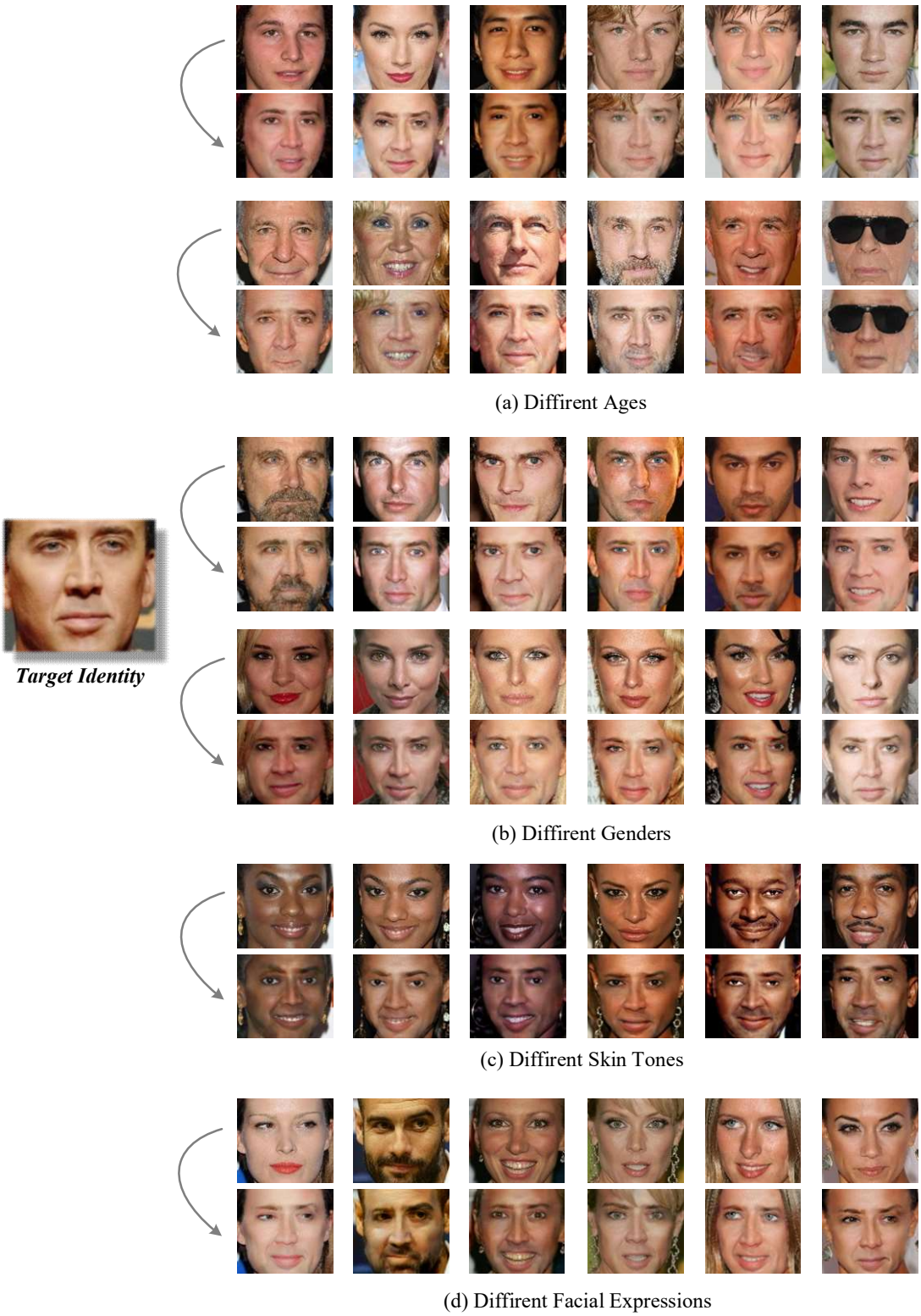


**FIGURE B.** **High-Fidelity Transformation Results.**

## REFERENCES


[1] Goodfellow I, Poujet-Abadie J, Mirza M, et al, Generative Adversarial Nets, Advances in neural information processing systems, 2672-2680 (2014)

[2] Kingma D P, Welling M, Auto-Encoding Variational Bayes. (2013)

[3] Rezende D J, Mohamed S, Wierstra D, Stochastic Backpropagation and Variational Inference in Deep Latent Gaussian Models, arxiv. (2014)

[4] Huang R, Zhang S, Li T, et al, Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis, IEEE International Conference on Computer Vision. (2017)

[5] Yin X, Yu X, Sohn K, et al, Towards large-pose face frontalization in the wild. In Proc. ICCV, 1–10. (2017)

[6] Tran L, Yin X, Liu X, Disentangled Representation Learning GAN for Pose-Invariant Face Recognition, IEEE Computer Vision and Pattern Recognition. (2017)

[7] Choi Y, Choi M, Kim M, et al, StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. (2017)

[8] He Z, Zuo W, Kan M, et al, Attgan: Facial attribute editing by only changing what you want, IEEE Transactions on Image Processing, 28(11), 5464-5478 (2019)

[9] Perarnau G, Van De Weijer J, Raducanu B, et al, Invertible conditional gans for image editing, arXiv preprint arXiv:1611.06355 (2016)

[10] Lample G, Zeghidour N, Usunier N, et al, Fader networks: Manipulating images by sliding attributes, Advances in Neural Information Processing Systems, 5967-5976 (2017)

[11] https://github.com/deepfakes/faceswap.

[12] Pérez P, Gangnet M, Blake A, Poisson image editing, ACM SIGGRAPH 2003 Papers, 313-318. (2003)

[13] Bao J, Chen D, Wen F, et al, CVAE-GAN: fine-grained image generation through asymmetric training, Proceedings of the IEEE International Conference on Computer Vision, 2745-2754 (2017)

[14] Liu Z, Luo P, Wang X, et al, Deep learning face attributes in the wild, Proceedings of the IEEE international conference on computer vision, 3730-3738 (2015)

[15] Ng H W, Winkler S, A data-driven approach to cleaning large face datasets, 2014 IEEE international conference on image processing, 343-347 (2014)

[16] Ackley D H, Hinton G E, Sejnowski T J, A learning algorithm for Boltzmann machines, Cognitive science, 9(1), 147-169 (1985)

[17] Larsen A B L, Sønderby S K, Larochelle H, et al, Autoencoding beyond pixels using a learned similarity metric, arXiv preprint arXiv:1512.09300. (2015)

[18] [18] Sohn K, Lee H, Yan X, Learning structured output representation using deep conditional generative models, Advances in neural information processing systems, 3483-3491 (2015)

[19] Arjovsky M, Chintala S, Bottou L, Wasserstein gan, arXiv preprint arXiv:1701.07875. (2017)

[20] Mao X, Li Q, Xie H, et al, Least squares generative adversarial networks, Proceedings of the IEEE International Conference on Computer Vision, 2794-2802 (2017)

[21] Simonyan K, Vedaldi A, Zisserman A, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034. (2013)

[22] Yosinski J, Clune J, Nguyen A, et al, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579. (2015)

[23] Johnson J, Alahi A, Fei-Fei L, Perceptual losses for real-time style transfer and super-resolution, European conference on computer vision, 694-711 (2016)

[24] Park T, Liu M Y, Wang T C, et al, Semantic image synthesis with spatially-adaptive normalization, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2337-2346 (2019)

[25] Zhang H, Goodfellow I, Metaxas D, et al, Self-attention generative adversarial networks, arXiv preprint arXiv:1805.08318. (2018)

[26] Cao Q, Shen L, Xie W, et al, Vggface2: A dataset for recognising faces across pose and age, 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 67-74 (2018)

[27] Gatys L A, Ecker A S, Bethge M, A neural algorithm of artistic style, arXiv preprint arXiv:1508.06576. (2015)

[28] Kingma D P, Ba J, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980. (2014)

[29] Zhang K, Zhang Z, Li Z, et al, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters, 23(10), 1499-1503 (2016)

[30] He K, Zhang X, Ren S, et al, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778 (2016)

[31] Szegedy C, Vanhoucke V, Ioffe S, et al, Rethinking the Inception Architecture for Computer Vision, IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2818-2826 (2016)

[32] Zhang R, Isola P, Efros A A, et al, The unreasonable effectiveness of deep features as a perceptual metric, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 586-595 (2018)

[33] Huang X, Liu M Y, Belongie S, et al, Multimodal unsupervised image-to-image translation, Proceedings of the European Conference on Computer Vision (ECCV), 172-189 (2018)

[34] Huang X, Belongie S, Arbitrary style transfer in real-time with adaptive instance normalization, Proceedings of the IEEE International Conference on Computer Vision, 1501-1510 (2017)

[35] Wang X , Girshick R , Gupta A , et al, Non-local Neural Networks. (2017)

[36] Fu J , Liu J , Tian H , et al, Dual Attention Network for Scene Segmentation. (2018)

[37] Xu W , Keshmiri S , Wang G, et al, Adversarially Approximated Autoencoder for Image Generation and Manipulation, IEEE Transactions on Multimedia. (2019)

[38] Xu W , Keshmiri S and Wang G, Toward Learning a Unified Many-to-Many Mapping for Diverse Image Translation, Patten Recognition. (2019)

[39] Li L, Bao J, Yang H, et al, FaceShifter: Towards High Fidelity and Occlusion Aware Face Swapping. (2019)

[40] Yujun S, Ping L, Junjie Y, et al, FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 821-830 (2018)

[41] Huang G, Liu Z, Maaten L V D, et al, Densely Connected Convolutional Networks, Computer Era. (2017)

[42] Ruiz N, Chong E, and Rehg J M, Fine-Grained Head Pose Estimation Without Keypoints. (2018)

[43] Deng Y, Yang J, Xu S, et al, Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. (2019)

[44] J Naruniec, L Helminger, C Schroers, et al, High-Resolution Neural Face Swapping for Visual Effects. (2020)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3044187, IEEE Access



**Longhao Zhang** is a Ph.D. of Automation School of Beijing University of Posts and Telecommunications. His major is Pattern Recognition. His research interests are in computer vision, machine learning, and deep learning. He is particularly interested in the areas of generative model and weakly supervised object localization.

**Xipeng Pan** received his Ph.D. degree from School of Automation, Beijing University of Posts and Telecommunications, China in 2019. Currently, he is a postdoctoral research fellow of Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, China. His research interests include machine learning, medical image processing. He is a member of CCF, and a member of ISAIR.

**Huihua Yang** received his Ph.D. degree from East China University of Science and Technology, China in 2005. He was a postdoctoral research fellow of Tsinghua University from 2005 to 2007. Currently, he is a professor of School of Automation, Beijing University of Posts and Telecommunications, China. His research interests include machine learning, spectrum analysis, and optimization. Dr. Yang has published more than 40 papers and serves as Director of China Instrument and Control Society (CICS), Vice Director of NIR Division of CICS, and is a senior member of CCF, and a member of ACM.

**Lingqiao Li** received his Ph.D. degree from School of Automation, Beijing University of Posts and Telecommunications, China in 2020. Currently, he is assistant researcher at the Guilin University of Electronic and Technology, China and his research interests include pattern recognition, large spectral data analysis.