AP-GAN: Improving Attribute Preservation in Video Face Swapping

Longhao Zhang, Huihua Yang, *Member, IEEE*, Tian Qiu, and Lingqiao Li

***Abstract*— Face swapping is a popular subject in face manipulation, which aims to replace the identity of the target face with that of the source face. Existing methods cannot well preserve facial attributes (e.g., pose, expression, skin color, illumination, make-up, occlusion, etc.) of the target face, causing noticeable temporal discontinuity and instability artifacts for video face swapping. In this paper, we propose a lightweight Generative Adversarial Networks based framework named AP-GAN, which can precisely control the attribute of the generated face to be consistent with that of the target face, achieving efficient and high-fidelity video face swapping. Specifically, we derive a U-Net based generator with ID blocks to translate identity and PE blocks to correct pose and expression. Besides, a PE-aware discriminator is designed to help supervise pose and expression of the synthetic face. Furthermore, we propose a discriminator based perceptual loss leveraging multi-scale features of the discriminator to preserve facial attributes like skin color, illumination, make-up and occlusion. AP-GAN is trained on Flickr-Faces-HQ, CelebA-HQ and VGGFace2 and evaluated on FaceForensics++. Extensive experiments and comparisons to the existing state-of-the-art face swapping methods demonstrate the efficacy of our framework. Comprehensive ablation studies are also carried out to isolate the validity of each proposed component and to contrast with other face manipulation approaches.**

***Index Terms*— Generative Adversarial Networks, video face swapping, facial attributes, U-Net, perceptual loss.**

I. Introduction

IN THE last few years, face swapping has become a hotspot technology and attracted wide attention due to its numerous real-world applications. It aims to replace the identity of the face in the target image or video with that of the source face. Early methods [1], [2] simply replace the pixels in the target face area with the pixels in the source face area. Therefore, the performance is poor when the facial poses or illuminations differ greatly. Recently, with the development of Generative Adversarial Networks (GAN) [17], researches have achieved promising progress [3]–[7], [50]. However, it still remains challenges for video face swapping to synthesize temporally stable results, since algorithms are applied independently to each video frame, any inconsistency of facial attributes (e.g., pose, expression, skin color, illumination, make-up, occlusion, etc.) between synthetic face and target face will cause noticeable discontinuity artifacts, which are described in Fig. 1. In this paper, we propose a lightweight yet effective video face swapping framework called AP-GAN, which can well preserve facial attributes of the target face in each frame, significantly improving the temporal continuity and stability.

Resent state-of-the-art methods, such as IP-GAN [6] and FaceShifter [7], employ an extra attribute encoder to disentangle facial attributes and identity of the target face and integrate the extracted attribute vector with the identity vector of the source face, causing significant attribute loss, especially for pose and expression. We argue that identity, pose and expression are all expressed by facial features (e.g., eyes, nose and mouth), thus they are highly coupled and hard to be completely disentangled. Besides, due to the lack of constraint or supervision on pose and expression, identity tends to play a leading role in integration. Unlike these methods, we do not explicitly disentangle the identity and attributes of the target face, but derive a U-Net based generator to conduct face reconstruction. During reconstruction, we translate the identity with ID blocks, which take as input the multi-scale identity embedding of the source face. In addition, PE blocks are introduced to correct pose and expression, which take as input the feature-wise boundary map of the target face. Note that with the help of the skip connection of U-Net, multi-scale attribute features and structure information can be transmitted from encoder to decoder intact, effectively reducing attribute loss.

In this work, we carefully design the discriminator and take full advantage of its features to help preserve facial attributes. In particular, a PE-aware discriminator is proposed to have a weak supervision on pose and expression. It takes as input the concatenation of face image and the feature-wise boundary map, so as to incorporate the consistency of position and profile of facial features between the target face and the synthetic face into the criterion of distinguishing real or fake. Besides, we propose a novel discriminator based perceptual loss, which contains an occlusion-perceptual part and a style-perceptual part, leveraging multi-scale features of the discriminator to preserve facial attributes like skin color, illumination, make-up and occlusion. We note that handling occlusions is always

Manuscript received February 22, 2021; revised April 26, 2021; accepted June 5, 2021. Date of publication June 16, 2021; date of current version April 5, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61906050 and in part by the Key Research and Development Program under Grant Guike AB16380293. This article was recommended by Associate Editor Y. Qin. *(Corresponding author: Huihua Yang.)*

Longhao Zhang and Huihua Yang are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: yhh@bupt.edu.cn).

Tian Qiu is with the Department of Electrical and Computer Engineering, University of California at San Diego (UCSD), San Diego, CA 92161 USA.

Lingqiao Li is with the School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3089724.

Digital Object Identifier 10.1109/TCSVT.2021.3089724

Fig. 1. Comparisons with other video face swapping methods. From left to right we show the source images, the target images, the results of DeepFakes [4], the results of FaceSwap [3], the results of FS-GAN [50] and the results of ours. It can be seen that existing methods cannot well preserve facial attributes of the target face, such as pose, expression, skin color, illumination and make-up.

challenging in face swapping. [5] and [50] train a face segmentation network with a number of manual annotations to obtain occlusion-aware face masks. [7] trains the Heuristic Error Acknowledging Refinement Network (HEAR-Net) in a self-supervised way to restore occlusions. Compared with these methods, our approach is more simple and efficient, avoiding additional networks or manual annotations.

In practice, to maintain background and hair, some methods [3], [4], [50] only synthesize a small facial region rather than the whole head. However, there are inevitable face shape and style mismatches when the synthetic region is stitched together with the target face, causing temporal discontinuity and instability in video face swapping, despite the use of segmentation [50], blending [3], [4], [50] and color correction [3], [4] technologies. Our AP-GAN can synthesize the whole head while keeping hair and background intact, which allows us to simply adjust the synthetic image to superpose it on the corresponding area of the target frame with little artifacts.

To conclude, we make the following contributions:

- **A lightweight and efficient framework**. Our proposed AP-GAN only uses a U-Net based generator and an identity encoder for inference. To the best of our knowledge, it is the most lightweight and compact subject-agnostic and occlusion-aware face swapping framework, which is revealed in Table I.
- **Constraint and supervision on pose and expression**. We propose a novel PE block to correct pose and expression of the synthetic face with the help of the feature-wise boundary map and design a PE-aware discriminator to have a weak supervision on pose and expression.
- **A discriminator based perceptual loss for attribute preservation**. We introduce a perceptual loss which contains an occlusion-perceptual part and a style-perceptual part. It leverages multi-scale features of the discriminator to help preserve facial attributes like skin color, illumination, make-up and occlusion.

TABLE I

COMPARISONS WITH STATE-OF-THE-ART SUBJECT-AGNOSTIC AND OCCLUSION-AWARE FACE SWAPPING METHODS [7], [50] IN FRAMEWORK COMPOSITION

	ID	*Att*	*G*	*Seg*	*Occ*	*I&B*
FaceShifter [7]	√	√	√	-	√	-
FS-GAN [50]	-	-	√	√	-	√
Ours	√	-	√	-	-	-

√ means the framework contains the corresponding component. Components include: identity encoder (*ID*), attribute encoder (*Att*), generator (*G*), segmentation network (*Seg*), occlusion network (*Occ*), inpainting network (*I*) and blending network (*B*). It can be concluded that our framework is the most lightweight and compact.

AP-GAN is trained on Flickr-Faces-HQ [10], CelebA-HQ [15] and VGGFace2 [49] and evaluated on FaceForensics++ [16]. Extensive experiments and comparisons to the existing state-of-the-art methods demonstrate its efficacy: it can achieve efficient, high-fidelity and temporally continuous video face swapping.

II. RELATED WORKS

A. Generative Adversarial Networks

Generative adversarial networks (GANs) [17] aim to model the natural image distribution by forcing the generated samples to be indistinguishable from natural images. GANs enable a wide variety of applications such as image generation [18], [19], domain transfer [20], [21], super resolution [22] and image inpainting [23]. However, GANs are hard to converge in the training stage and the samples generated from GANs

are often far from natural. Plenty of efforts have been made to improve the training and performance of GANs. W-GAN [24] uses Earth Mover Distance as an objective for training GAN. It improves the stability of learning and get rid of problems like mode collapse. LS-GAN [25] changes the sigmoid cross entropy loss function 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. PG-GAN [15] proposed to grow both the generator and discriminator networks to increase the resolution of generated images. In [27], a multi-scale generator and discriminator architecture is proposed to synthesize photo-realistic images, which are more visually appealing than those computed by previous methods. In this work, we also adopt the multi-scale discriminators, which can guide the generator to generate globally consistent images and to produce finer details.

B. U-Net Based Architecture

U-Net [28] or U-Net based architecture is widely used in semantic segmentation, especially in medical image segmentation tasks [29], [30]. It contains an encoder to embed the input image to a low-dimension hidden space, a decoder to reconstruct the hidden vector to the output image which has the same resolution with the input image, and the skip connections between them. Res-UNet [31] introduce the residual module into the U-Net structure for high-quality retina vessel segmentation, while FD-UNet [32] takes advantage of the dense connection to remove artifacts from 2D photoacoustic tomography images reconstructed from sparse data. [33] propose a novel attention gate model for U-Net to automatically learns to focus on target structures of varying shapes and sizes, suppressing irrelevant regions in an input image while highlighting salient features useful for a specific task. In this work, we derive a U-Net based generator with ID blocks and PE blocks for video face swapping. The architecture is simple yet effective, whose skip connections can help preserve face attributes.

C. Adaptive Denormalization

Normalization is vital in neural network training regarding both discriminative or generative tasks. It makes the input features approach independent and identical distribution by a shared mean and variance. Adaptive denormalization was first proposed for style transfer [8]. Specifically, each feature map is normalized to zero mean and unit deviation. Then the normalized feature maps are denormalized by modulating the activation using an affine transformation whose parameters are learned upon other features or conditions. Based on this, [9] proposes the SPADE block, which applies a spatially varying affine transformation, making it suitable for image synthesis from spatially-varying semantic mask. [10] introduces the AdaIN block to control the generation of face attributes. Later, FaceShifter [7] combines SPADE and AdaIN and proposes the Adaptive Attentional Denormalization (AAD) layer to adaptively embed identity and attributes for face swapping. Besides, [11] adds the layer normalization (LN) to AdaIN to make up for the deficiency of the instance normalization (IN) for domain transfer. Motivated by [7]–[11], we propose the adaptive denormalization based ID block and PE block. The former takes as input the multi-scale identity embedding to achieve identity translation, and the latter takes as input the feature-wise boundary map to correct pose and expression of the synthetic face.

D. Face Manipulation

Generative adversarial networks (GANs) have produced impressive results in generating realistic images, especially in face manipulation. [54] addresses the issue of facial attribute transfer with unpaired data. While [12], [58], [59] focus on aesthetically re-rendering the aging face at any future age for an individual face. [13], [44], [62] aim to solve extreme or varying pose problem. [56], [75] are proposed for face super-resolution. [57] introduces a discriminator with dilated convolutions to achieve cross-domain face translation. [72] could generate a talking face video with accurate lip synchronization given an arbitrary speech clip or text information as input. [73]–[80] improve face recognition and reconstruction assisted by attribute or 2D and 3D image. Instead of swapping a face on to an existing image like we do, face hallucination approaches [70], [71] try to repair local details for the restoration of both identity and appearance characteristics. Moreover, [55] proposes a multi-scale GAN model to hallucinate realistic context (forehead, hair, neck, clothes) and background pixels automatically from a single input face mask.

E. Face Swapping

Face swapping and face swapping detection technologies [51]–[53] have attracted wide attention. Early face swapping methods [1], [2] are limited by pose and perspective, thus they could only be applied under certain conditions. 3D based methods [3], [5], [36] used the 3D model to handle the pose and perspective difference. However, the accuracy and realism of 3D reconstruction of faces are unsatisfactory. Recently, with the development of Generative Adversarial Networks (GANs), GAN based methods have achieved superior results. RS-GAN [37] independently handles face and hair appearances in the latent spaces, and then, face swapping is achieved by replacing the latent-space representations of the faces, and reconstructing the entire face image with them. IP-GAN [6] disentangles the identity and attributes of faces and conveniently recombines different identities and attributes for face swapping in open domains. Whereas due to the information loss caused by the compressed representation, these methods are incapable of achieving high-fidelity face swapping. FaceShifter [7] proposes the Adaptive Attentional Denormalization (AAD) layer to adaptively embed identity and attributes. FS-GAN [50] derives a recurrent neural network (RNN) based approach which adjusts for both pose and expression variations. Although these methods have achieved promising progress, they cannot well preserve attributes (e.g., pose, expression, skin color, illumination, make-up, occlusion, etc.) of the target face, causing noticeable temporal discontinuity and instability artifacts for video face swapping.

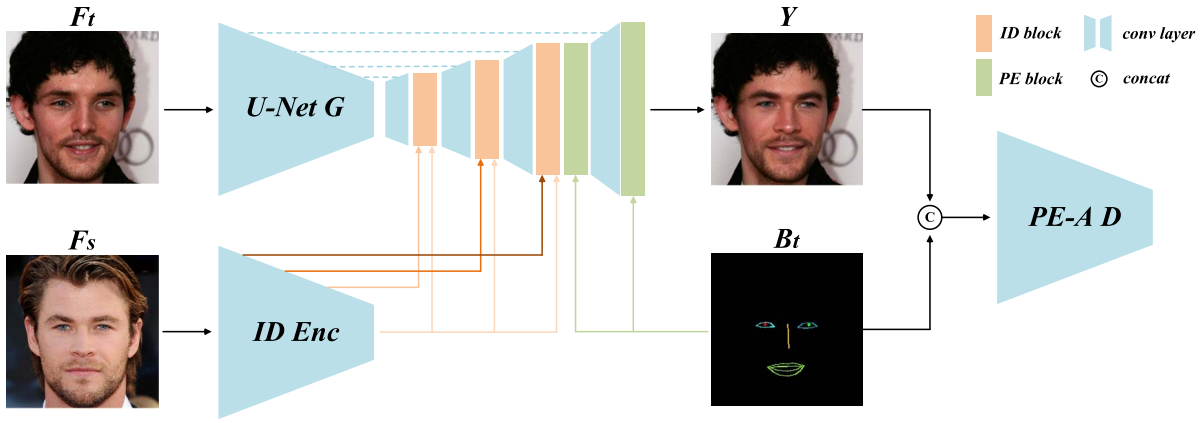


Fig. 2. Overview of the proposed AP-GAN. It contains three components: 1) a U-Net based generator (*U-Net G*) to reconstruct the target face F_t, which uses ID blocks to translate identity and PE blocks to correct pose and expression; 2) an identity encoder (*ID Enc*), which encodes the source face F_s to multi-scale id embedding and sends them to ID blocks of each stage (the color of the arrow from dark orange to light orange represents the scale of id vector from large to small); 3) a pose- and expression-aware discriminator (*PE-A D*), which takes as input the concatenation of face (F_t or Y) and the feature-wise boundary map B_t to help supervise pose and expression while distinguishing real and fake samples.

III. AP-GAN

In this section, we introduce the proposed AP-GAN, which is shown in Fig. 2. It contains three parts: 1) the identity encoder (*ID Enc*), which is a pre-trained state-of-the-art face recognition network [38]. It extracts multi-scale identity embedding from the source face F_s for identity translation. 2) The U-Net based generator (*U-Net G*), which takes as input the target face F_t and generate face Y. It has ID blocks to achieve the identity translation and PE blocks to ensure the pose and expression consistency of F_t and Y. 3) The PE-aware discriminator (*PE-A D*), whose input is the concatenation of F_t and its feature-wise boundary map B_t (as real sample), or Y and B_t (as fake sample).

A. U-Net Based Generator (U-Net G)

In this work, we derive a lightweight face generator with U-Net architecture. It takes as input the target face F_t to generate Y, which has the identity of the source face F_s and the facial attributes of F_t. To this end, the adaptive denormalization based ID block and PE block are proposed, which are described in Fig. 3. In particular, we use the former to modify the identity and the latter to correct pose and expression.

Let $M \in \mathbb{R}^{H\times W\times C}$ denotes the feature fed into a block, where H and W are height and width, respectively, and C is the number of channels. We first perform instance normalization (IN) and layer normalization (LN) on M respectively to get M_{IN} and M_{LN}:

$$\begin{aligned} M_{IN} &= \frac{M-\mu_{IN}}{\sigma_{IN}} \\ M_{LN} &= \frac{M-\mu_{LN}}{\sigma_{LN}} \\ \mu^c_{IN} &= \frac{1}{HW}\sum_{x,y} M^{c,x,y}_{IN}, \quad \sigma^c_{IN} \\ &= \sqrt{\frac{1}{HW}\sum_{x,y}(M^{c,x,y}_{IN})^2 - (\mu^c_{IN})^2} \end{aligned} \tag{1}$$

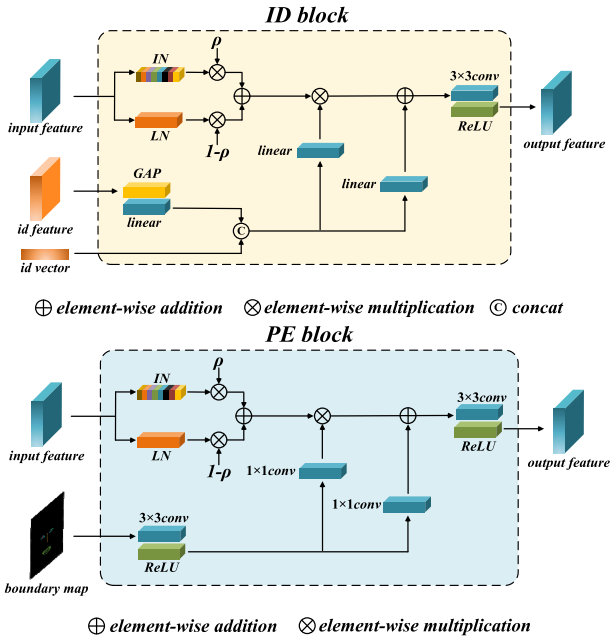


Fig. 3. Details of the ID block and the PE block. Note that ρ is a learnable modulation factor to adaptively combine features from the instance normalization layer (IN) and the layer normalization layer (LN). The affine transformation parameters of ID block are learned upon the multi-scale id embedding (id vector + id feature), and the affine transformation parameters of PE block are learned upon the feature-wise boundary map.

$$\begin{aligned} \mu_{LN} &= \frac{1}{CHW}\sum_{c,x,y} M^{c,x,y}_{LN}, \quad \sigma_{LN} \\ &= \sqrt{\frac{1}{CHW}\sum_{c,x,y}(M^{c,x,y}_{LN})^2 - (\mu^c_{LN})^2} \end{aligned} \tag{2}$$

where μ_{IN} and σ_{IN} are the channel-wise mean and standard deviation for IN, μ_{LN} and σ_{LN} are the layer-wise mean and standard deviation for LN. Then, we combine them using a learnable modulation factor ρ, and apply denormalizations with affine transformation parameters γ and β:

$$\overline{M} = \gamma \cdot (\rho \cdot M_{IN} + (1-\rho)\cdot M_{LN}) + \beta \tag{3}$$

Note that the affine transformation parameters γ_{ID} and β_{ID} for ID block are learned upon the multi-scale id embedding,

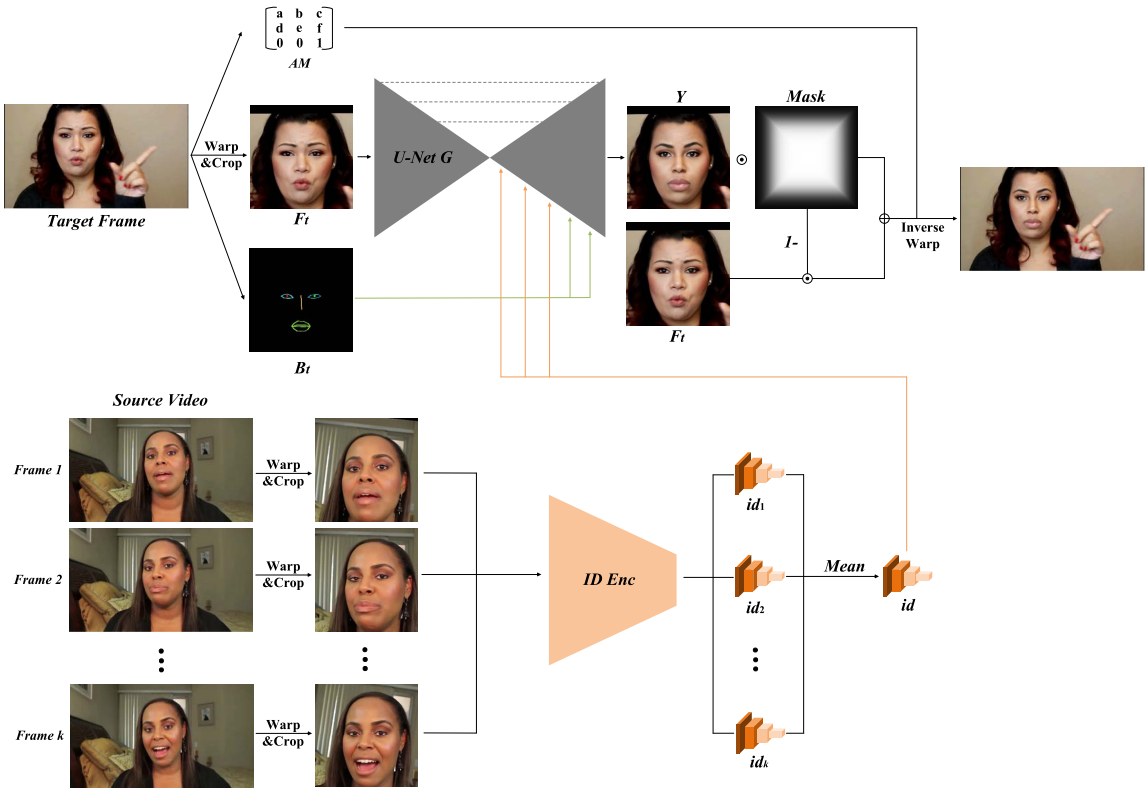


Fig. 4. The inference pipeline of our AP-GAN. For each frame in the target video, we find the target face by face detection. Then we warp and crop the target frame to implement face alignment to obtain the aligned face F_t. In this process, we get the affine matrix AM and the feature-wise boundary map B_t. Afterwards, F_t, B_t and multi-scale id embedding extracted by the identity encoder (*ID Enc*) are fed into the U-Net based generator (*U-Net G*) to generate Y. Finally, we blend Y and F_t with mirrored-sigmoid mask and carry out inverse warp with help of AM to stitch the blended output onto the target frame.

while γ_{PE} and β_{PE} for PE block are learned upon the feature-wise boundary map. Specifically, we feed the source face F_s into the *ID Enc* to obtain the multi-scale id embedding, which contains the 256-dimensional *id vector*, defined to be the last vector before the fully connected layer, and the *id feature*, defined to be the output feature of the last convolution layer of the stage with the same resolution as the input feature M. The *id feature* will go through a global average pooling layer (GAP) and a fully connected layer to remove the spatial information and then concatenate with the *id vector*. The feature-wise boundary map $B_t \in \mathbb{R}^{H\times W\times 3}$ is obtained by connecting the facial landmarks of F_t and scaling it to reach the resolution of M. The boundary line of each facial feature has a different pixel value, thus B_t is feature-wise. We note that, in order to minimize the impact on identity, only a few landmarks closely related to pose and expression are involved. In particular, we only leverage the points of eyes, eyeballs, mouth and nose bridge, since we argue that pose and expression are mainly described by the opening and closing of eyes, the position of eyeballs, the orientation of nose bridge and the shape of mouth.

As illustrated in Fig. 2, the ID block is applied in high and middle stages and the PE block in low stages. We argue that the id embedding contains more abstract and richer semantic information hence should work on higher feature levels with coarser resolutions to mold the identity, while landmarks and boundaries are more likely to work on lower feature levels with finer resolutions to correct profile details.

B. PE-Aware Discriminator (PE-A D)

To help preserve pose and expression, we propose a PE-aware discriminator (*PE-A D*), whose input is the concatenation of the face image and the feature-wise boundary map. Specifically, we concatenate $F_t \in \mathbb{R}^{H\times W\times C}$ with its feature-wise boundary map $B_t \in \mathbb{R}^{H\times W\times 3}$, or $Y \in \mathbb{R}^{H\times W\times C}$ with the B_t in the channel dimension to formulate the real or fake sample $\in \mathbb{R}^{H\times W\times (C+3)}$. In this way, the discriminator would also take note of whether the state of each facial feature of the synthetic face consistent with that of the target face, which is an explicit supervision on pose and expression.

Besides, to generate photo-realistic images, we design our PE-aware discriminator in a multi-scale way [27]. In particular, we use n discriminators with the same network architecture but operate at different image scales. The discriminator that operates at the coarser scale has a larger receptive field and a more global view of the image, guiding the generator to generate globally consistent images. On the other hand, the discriminator operating at the finer scale is specialized in guiding the generator to produce finer details. In this work, we set n to 3. Therefore, we down sample F_t, Y and B_t by a factor of 2 and 4 to create the pyramid of 3 scales as the input for the *PE-A D*.

C. Training Loss Functions

1) Adversarial Loss: We adopt adversarial training for the proposed framework. The generator competes in the following

two-player minimax game with the multi-scale discriminators:

$$\underset{G}{Min}\ \underset{D_1,D_2\cdots D_n}{Max}\ \sum_{i=1}^{n}\mathcal{L}_{adv}\left(G,D_i\right) \tag{4}$$

where $\mathcal{L}_{adv}$ is the adversarial loss, which is given by:

$$\begin{aligned}\mathcal{L}_{adv} = {} & \mathbb{E}_{F,B}\left[\log\left(D\left(F_t,B_t\right)\right)\right] \\ & +\mathbb{E}_{F,B}\left[\log\left(1-D\left(G\left(F_s\right),B_t\right)\right)\right]\end{aligned} \tag{5}$$

2) Identity Loss: $\mathcal{L}_{id}$ is minimized to accurately translate the identity from the source face F_s to the synthetic face Y. It is formulated as:

$$\mathcal{L}_{id} = 1 - cos\left(IDEnc\left(Y\right), IDEnc\left(F_s\right)\right) \tag{6}$$

where $cos(\cdot,\cdot)$ represents the cosine similarity of two id vectors extracted from the identity encoder *ID Enc*.

3) Reconstruction Loss: We use a reconstruction loss $\mathcal{L}_{rec}$ to help restore hair, background and some occlusion, defined as the squared Euclidean distance between F_t and Y:

$$\mathcal{L}_{rec} = \begin{cases} \left\|F_t - Y\right\|_2^2 & F_t = F_s \\ \lambda_1\left\|F_t - Y\right\|_2^2 & F_t \neq F_s \end{cases} \tag{7}$$

When F_s is the same as F_t, output Y must to be the same as F_t or F_s. When F_s and F_t are not the same, only a small part of the content needs to be completely restored, so we multiply the loss by a small weight λ_1. In this work, we set λ_1 to 0.1.

4) Perceptual Loss: In this work, we introduce a discriminator based perceptual loss $\mathcal{L}_{per}$ for facial attribute preservation. Unlike the widely used version [39], which leverages features of a pre-trained VGG network to construct the loss function, we take advantage of features from the discriminator. We note that the discriminator hardly captures identity information, because the identities of real and fake samples are bound to be different and cannot be used as criteria for face swapping. However, features extracted by the discriminator carry information about identity-irrelevant content, which we can exploit to restore facial occlusion. While the correlations of these features encode face styles, which can help us restore skin color, illumination and make-up. Therefore, $\mathcal{L}_{per}$ can be used to preserve facial attributes without affecting identity translation.

$\mathcal{L}_{per}$ consists of two parts: an occlusion-perceptual loss $\mathcal{L}_{op}$ and a style-perceptual loss $\mathcal{L}_{sp}$. Specifically, we get the multi-scale features of F_t and Y from different stages {1, 2, …, m} of the discriminator as $\{F^1, F^2, \ldots, F^m\}$ and $\{Y^1, Y^2, \ldots, Y^m\}$. The occlusion-perceptual loss $\mathcal{L}_{op}$ is defined as the squared Euclidean distance between per-stage features:

$$\mathcal{L}_{op} = \sum_{s=1}^{m}\frac{1}{C^s H^s W^s}\left\|F^s - Y^s\right\|_2^2 \tag{8}$$

where F^s or Y^s is the feature of F_t or Y of stage s, C^s, H^s and W^s denote the number of channels, height and width of stage s, respectively. While the style-perceptual loss $\mathcal{L}_{sp}$ is defined as the squared Frobenius norm of the difference between the per-stage Gram matrices:

$$\mathcal{L}_{sp} = \sum_{s=1}^{m}\frac{1}{\left(C^s H^s W^s\right)^2}\left\|FG^s - YG^s\right\|_F^2 \tag{9}$$

where FG^s and YG^s are Gram matrices [26] of F^s and Y^s, which are computed as:

$$\begin{aligned} FG_{ij}^s &= \sum_{k=1}^{H^s\times W^s} F_{ik}^s F_{jk}^s \\ YG_{ij}^s &= \sum_{k=1}^{H^s\times W^s} Y_{ik}^s Y_{jk}^s \end{aligned} \tag{10}$$

Gram matric is proportional to the uncentered covariance of the features, thus captures information about which features tend to activate together.

5) Overall Loss: The overall training loss function is a weighted sum of above losses as:

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{id}\mathcal{L}_{id} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{op}\mathcal{L}_{op} + \lambda_{sp}\mathcal{L}_{sp} \tag{11}$$

where we set $\lambda_{adv} = 2$, $\lambda_{id} = \lambda_{rec} = 10$, $\lambda_{op} = 1$ and $\lambda_{sp} = 3$ in this work.

D. Inference Pipeline

The inference pipeline is described in Fig. 4. For each frame in the target video, we first obtain the facial box and landmarks through the face detection model [60] and the landmark detection model [61]. Then we calculate the affine transformation matrix *AM* for alignment, which warps the facial landmarks in the box to be consistent with the reference landmarks (average facial landmarks). We implement face alignment and cropping using *OpenCV* function *warpAffine()* to get F_t. Besides, the feature-wise boundary map B_t is made by picking out the points of eyes, mouth and nose bridge from the aligned landmarks and connecting them. After that, the warped and cropped face F_t, feature-wise boundary map B_t and the multi-scale id embedding *id* which is extracted from the source video through the identity encoder (*ID Enc*) are fed into the U-Net based generator (*U-Net G*) to generate Y. In this work, we adopt a fast and effective blending method to adjust the output Y since it might not quite stitch together well when superimposed on top of the target frame. In particular, we apply the mirrored-sigmoid mask to weight the pixel values so that the outer edges of the output image contain pixels that match F_t, and towards the center, only pixels of Y are kept. The mirrored-sigmoid mask is formulated as:

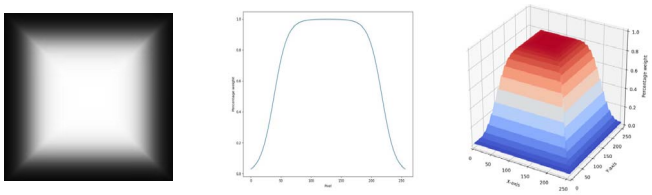
$$Mask\left(x,y\right) = \frac{1}{1+e^{\alpha\cdot\left(max\left(abs\left(x-mid\right),abs\left(y-mid\right)\right)-\gamma\cdot mid\right)}} \tag{12}$$

where *mid* is half the image size, α controls the slope of the function and is set to 0.08, and the inflection factor γ is set to 0.7. Its visualization of 2D and 3D is shown in Fig. 5. The blending process can be written as:

$$output = Mask\cdot Y + \left(1 - Mask\right)\cdot F_t \tag{13}$$

Finally, we call *warpAffine()* again setting the parameter *flags* to "*WARP_INVERSE_MAP*" to implement the inverse warping to stitch the output onto the target frame.

We note that a k-shot inference paradigm is introduced to further enhance the performance for video face swapping.



(a) (b) (c)

Fig. 5. The mirrored-sigmoid blending mask (a) and its visualization of 2D (along x or y dimension) (b) and 3D (c).

In particular, we evenly sample k frames from the source video and send k source faces into *ID Enc* to obtain the mean id embedding by averaging the output. k-shot strategy can effectively avoid the interference caused by some extreme conditions of the face from a single source frame on the extraction of *id* and make use of more abundant face information.

IV. Experiments

A. Experiments Setup

For each image, we use a face detection network [60] to locate the face and a landmark detection network [61] to obtain the facial landmarks. The face is aligned and cropped using common 5-point average landmarks, and resized to 256×256. The *ID Enc* is pre-trained on DeepGlint-Face (including MS1M-DeepGlint and Asian-DeepGlint) [41]. ID blocks are employed at high and middle reconstruction stages where the resolutions of feature maps are 4×4, 8×8, 16×16, 32×32, 64×64 and 128×128, and PE blocks are only employed at the last two stages where the resolutions are 128×128 and 256×256. We also try PE blocks at higher stages, but there is no obvious benefit.

The proposed AP-GAN is trained on Flickr-Faces-HQ [10], CelebA-HQ [15] and VGGFace2 [49], and evaluated on FaceForensics++ [16]. The Adam optimizer [42] is adopted. Learning rate for the generator and discriminator is 1×10^{-4} and 4×10^{-4}, respectively. $\beta_1 = 0$ and $\beta_2 = 0$. Also, we apply spectral normalization [43] to both the generator and discriminator to stabilize the training procedure.

B. Qualitative Comparison With Other Methods

We compare the face swapping quality of our proposed framework with other state-of-the-art video face swapping methods (FaceSwap [3], DeepFakes [4] and FS-GAN [50]) on FaceForensics++. Results are shown in Fig. 6. We note that since FaceSwap and DeepFakes all follow the strategy that first synthesizing the inner face region then blending it into the target face, as expected, they suffer from the blending inconsistency. However, our framework can well synthesize the whole face region, hair and background, so it can be easily blended into the target frame with little artifacts. Besides, existing methods cannot well respect facial attributes (e.g., pose, expression, skin color, illumination, make, etc.) from the target face. The inconsistency of pose and expression with the target face results in poor temporal continuity, while the deviation of light, skin color and make-up significantly reduce the temporal stability and realness. Yet we achieve higher fidelity by well preserving the facial attributes.

TABLE II
User Study Results

	id	*attr*	*temp*	*real*
DeepFakes [4]	10	4	14	8
FaceSwap [3]	12	6	4	4
FS-GAN [50]	22	26	22	16
Ours	**56**	**64**	**60**	**72**

We show the averaged selection percentages (%) for each method on identity (*id*), attributes (*attr*), temporal continuity and stability (*temp*), and realism (*real*).

To further reveal the visual and perceptual superiority of our method, we conduct user studies. In particular, we let 50 human evaluators watch source videos, target videos and reshuffled face-swapped videos generated by DeepFakes [3], FaceSwap [4], FS-GAN [50] and ours, and ask them to select: 1) the one that has the most similar identity with the source video; 2) the one that shares the most similar facial attributes (including pose, expression, skin color, illumination, make-up and occlusion) with the target video; 3) the one with the best temporal continuity and stability; and 4) the most realistic one. The averaged selection percentage for each method on each study is presented in Table II. Results show that our proposed framework surpasses other face swapping methods by large margins.

C. Quantitative Comparison With Other Methods

Quantitative experiments are carried out on FaceForensics++. We use the following metrics to conduct the quantitative comparison: identity retrieval (*id*), domain-invariant perceptual distance (*DIPD*), pose distance (*pose*), expression distance (*exp*), structural similarity (*SSIM*), average endpoint error (*AEE*) and flow warping error (*FWE*).

1) Identity Translation: The identity retrieval is adopted to measure identity translation accuracy. We extract the identity vector using a state-of-the-art face recognition network [38] and use the cosine similarity to estimate the identity distance. For each generated face from the test set, we search the nearest face in FaceForensics++ frames and check whether it belongs to the correct source video. The averaged accuracy of all such retrievals is reported as the identity retrieval (*id*).

2) Attribute and Content Preservation: Attribute and content preservation is measured by a variant of perceptual distance [45] called domain-invariant perceptual distance (*DIPD*) [46], which is given by Euclidean distance between the normalized features of the target face F_t and the synthetic face Y. To compute the *DIPD*, features are extracted from the *conv5* layer of VGGNet. Then, the instance normalization is applied to remove their means and variances, which can filter out most identity-specific information and retain attributes and contents like skin color, illumination, occlusion, hair and background, according to [8].

3) Pose and Expression Preservation: We utilize the open-source pose estimator [47] and 3D face model [48] to estimate pose and expression preservation as [7] does. We report the Euclidean distance of pose and expression vectors between F_t and Y.

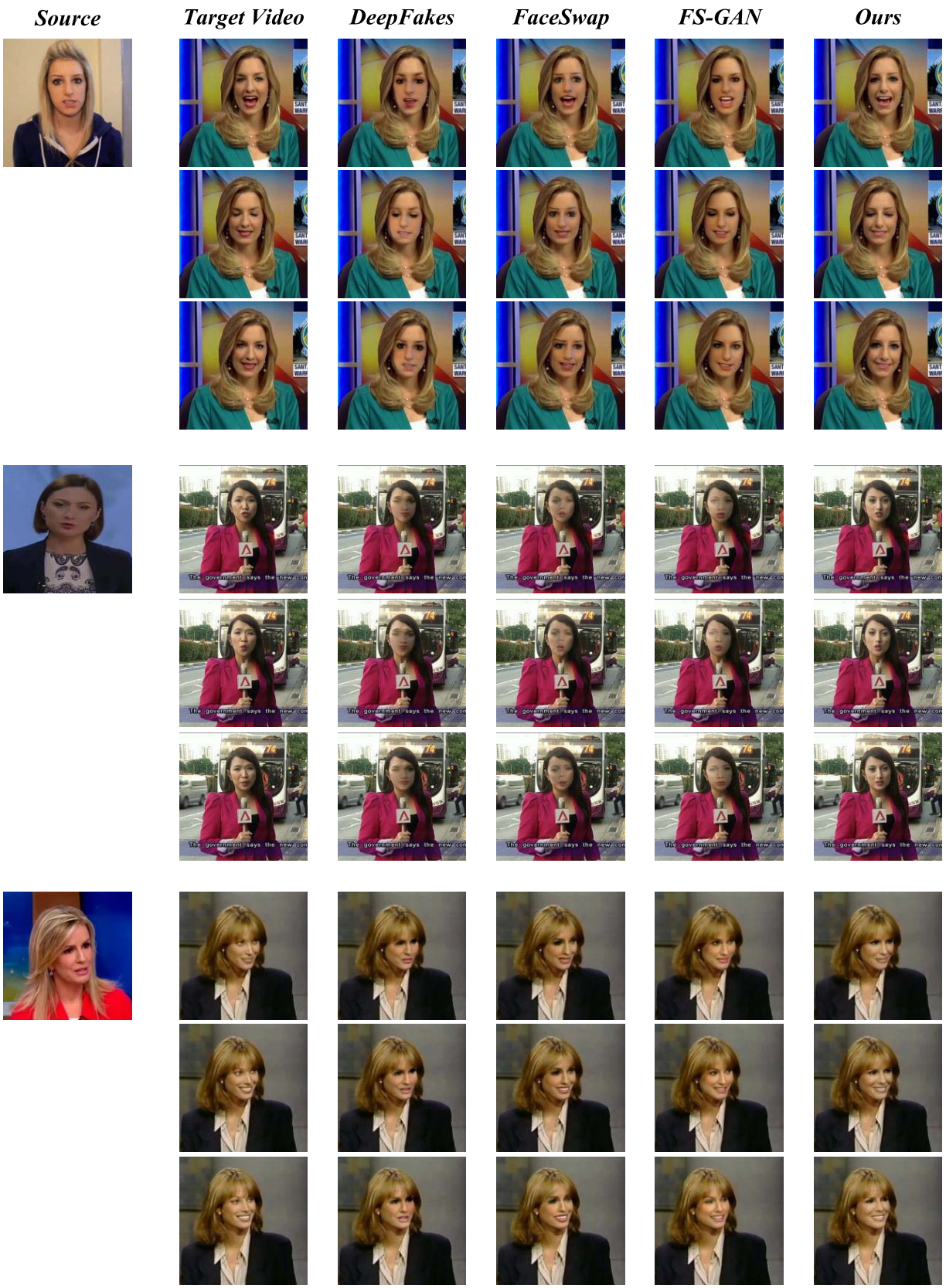


Fig. 6. Qualitative comparison with DeepFakes [4], FaceSwap [3] and FS-GAN [50]. Our results have less artifacts and blending inconsistency, and are also more faithful to the target face attributes and styles.

4) Realism: To quantify the photorealism of the synthetic image, the structural similarity (*SSIM*) is adopted. *SSIM* measures the similarity between Y and F_t in terms of brightness, contrast and structure. The higher the *SSIM* value, the lower the distortion of the synthetic image.

TABLE III
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART FACE SWAPPING METHODS

	id↑	*DIPD*↓	*pose*↓	*exp*↓	*SSIM*↑	*AEE*↓	*FWE*↓
Nirkin *et al.* [5]	0.76	-	3.29	2.33	-	-	-
IP-GAN [6]	0.82	-	4.04	2.50	-	-	-
FaceShifter [7]	0.97	-	2.96	**2.06**	-	-	-
DeepFakes [4]	0.76 / 0.81	1.33	4.87 / 4.14	3.02 / 2.57	0.38	1.89	0.27
FaceSwap [3]	0.60 / 0.54	0.78	2.90 / 2.51	2.73 / 2.14	0.44	1.95	0.28
FS-GAN [50]	0.81 / 0.62	0.55	2.72	2.59	0.42 / 0.51	1.81	0.25
Ours	**0.98**	**0.28**	**2.43**	**2.19**	**0.77**	**1.50**	**0.22**

↑ means larger numbers are better, ↓ means smaller numbers are better. Result with _ indicates that it is reported in [7] or [50].

5) Temporal Continuity and Stability: We introduce the average endpoint error (*AEE*) [40] and the flow warping error (*FWE*) to measure the temporal continuity and stability. In particular, for each original video in FaceForensics++ and its face swapping result, we cut and crop them to the original face sequence and the resulting face sequence. Specifically, we enlarge the face box detected in the first frame by 1.2 times as the fixed area and track the face in this area until the *IOU* of the detected face box and this area in a certain frame is less than 0.25. Then, we use FlowNet2.0 [40] to estimate the optical flow for every sequence. *AEE* is calculated as the average Euclidean distance between the flow fields of each pair of adjacent frames in the original sequences and the corresponding flow fields in the resulting sequences.

The flow warping error between frame m and frame n of a resulting sequence is calculated as:

$$E_{warp}(Y_m, Y_n) = \frac{1}{\sum_{i=1}^{HWC} M_{m,n}^{(i)}} \sum_{i=1}^{HWC} M_{m,n}^{(i)} \left\| Y_m^{(i)} - W(Y_n^{(i)}) \right\|_1 \tag{14}$$

where $M_{m,n} \in \{0, 1\}$ is the occlusion map for frame m and frame n. We use the occlusion detection method in [14] to obtain the mask $M_{m,n}$. W is backward warping with optical flow of the original sequence. Therefore, the flow warping error (*FWE*) of a resulting sequence is:

$$FWE(\{Y_t\}_{t=1}^{T}) = \frac{1}{T-1} \sum_{t=2}^{T} \{E_{warp}(Y_t, Y_1) + E_{warp}(Y_t, Y_{t-1})\} \tag{15}$$

where $E_{warp}(Y_t, Y_1)$ considers the long-term stability and $E_{warp}(Y_t, Y_{t-1})$ considers the short-term stability.

Table III shows the quantitative comparisons with other state-of-the-art methods (Nirkin *et al.* [5], IP-GAN [6], FaceShifter [7], DeepFakes [4], FaceSwap [3] and FS-GAN [50]). As can be seen from the table, our proposed framework is superior to others in terms of identity translation, attributes preservation, realism and temporal continuity and stability.

D. Ablation Study

To verify the necessity of the discriminator based perceptual loss, we select target images with facial occlusions, complex lightning or make-up from CelebA-HQ [15] to conduct the comparative experiments. Results are shown in Fig. 7 (a). From the results we can see that the perceptual loss does help preserve make-up (eye-shadow in the 1st and 2nd columns of target images) and facial occlusions (bubble, finger, hat and sunglasses in the 3rd, 4th and last columns of target images).

To verify the necessity of the PE block and the PE-aware discriminator, we choose target images with some extreme facial postures and expressions from CelebA-HQ for comparative experiments. Results in Fig. 7 (b) indicate that eye-gaze, mouth shape and some micro-expression of the synthetic faces are highly consistent with those of the target faces when the pose and expression constraints and supervisions are applied. We note that the pose and expression consistency is crucial for the temporal continuity of the resulting video.

Adaptive instance denormalization (AdaIN) [10] normalizes each feature map independently thus ignores the correlation between feature channels. In this work, we reveal that AdaIN based block [7], [9] is sub-optimal for the face swapping task. As displayed in the 2nd row of Fig. 7 (c), some structural artifacts (e.g., binocular asymmetry) emerge in the synthetic face when we change our ID block structure to AdaIN structure. This happens because the left eye of the source face is closed, yet the generator can only generate it by guessing, not by referring to the other eye. However, by combining instance normalization (IN) with layer normalization (LN), we successfully eliminate asymmetry artifacts and significantly improves the realism of the synthetic faces, which is shown in the 3rd row of Fig. 7 (c).

Quantitative comparison results for these proposed components are shown in Table IV. It turns out that replacing AdaIN with AdaLIN (IN + LN) in our ID blocks can improve identity translation (*id*) and realism (*SSIM*). PE block and PE-aware discriminator can significantly improve the preservation of facial attributes (*DIPD*), especially for pose and expression (*pose* & exp), thus enhance the temporal continuity and stability (*AEE* & FWE). The discriminator based perceptual loss ($\mathcal{L}_{op}+\mathcal{L}_{sp}$) increases *DIPD* to a certain extent and greatly increases *SSIM*.

In this work, we propose the k-shot inference paradigm to further enhance the performance for video face swapping, which can effectively avoid the interference caused by some extreme conditions of the face from a single source frame on the extraction of the id embedding and make use of more abundant face information. In order to inspect its influence on the evaluation metrics, we test with multiple k and show the results in Table V. In practice, we adopt the 10-shot paradigm to balance the effect and time consumption.

TABLE IV
QUANTITATIVE COMPARISON WITH EACH PROPOSED COMPONENT

PE blk	*PE-A D*	*AdaLIN*	Λ_{op}	Λ_{sp}	*id*↑	*DIPD*↓	*pose*↓	*exp*↓	*SSIM*↑	*AEE*↓	*FWE*↓
					0.95	0.48	2.71	2.80	0.65	1.62	0.24
√					0.95	0.39	2.56	2.52	0.66	1.56	0.24
√	√				0.95	0.32	2.47	**2.23**	0.66	1.52	0.23
√	√	√			**0.96**	0.32	2.47	**2.23**	0.67	1.52	0.23
√	√	√	√		**0.96**	0.31	**2.46**	**2.23**	0.71	**1.51**	**0.22**
√	√	√	√	√	**0.96**	**0.30**	**2.46**	**2.23**	**0.74**	**1.51**	**0.22**

TABLE V
QUANTITATIVE COMPARISON WITH DIFFERENT k VALUES IN K-SHOT INFERENCE PARADIGM

k	*id*↑	*DIPD*↓	*pose*↓	*exp*↓	*SSIM*↑	*AEE*↓	*FWE*↓
1	0.96	0.30	2.46	2.23	0.74	1.51	**0.22**
5	0.97	0.29	**2.43**	2.24	**0.77**	**1.50**	**0.22**
10	**0.98**	**0.28**	**2.43**	**2.19**	**0.77**	**1.50**	**0.22**
50	**0.98**	**0.28**	2.44	**2.19**	**0.77**	**1.50**	**0.22**

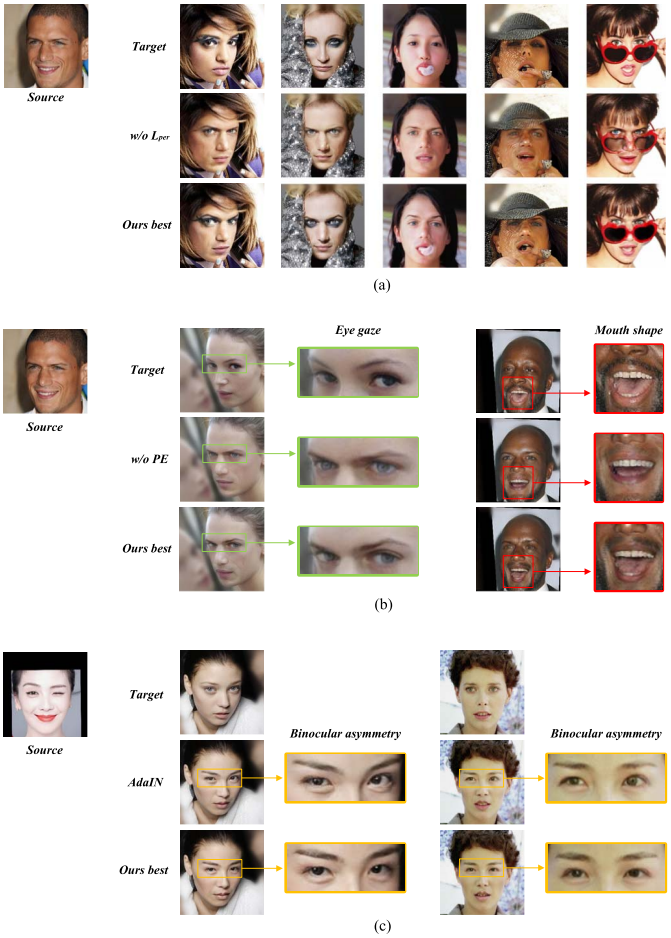


Fig. 7. Qualitative comparison results for discriminator based perceptual loss, pose and expression constraints and supervisions (PE), and AdaIN block.

In addition to achieving high-fidelity video face swapping, our proposed AP-GAN is also the most advanced in terms of synthesis quality and identity translation compared with other face manipulation methods. To prove this, we carry out the comparative experiments with some state-of-the-art approaches [54], [55], [64]–[69] on CelebA-HQ [15] and LFW [63] using common evaluation metrics like the Frechet inception distance score (*FID*), the structural similarity (*SSIM*), the mean match score (*MMS*) [55] and the identity discrepancy (*ID*) [65]. Results are shown in Table VI and VII.

TABLE VI
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART FACE MANIPULATION METHODS ON LFW [63] DATASET

	FID ↓	*SSIM* ↑	*MMS* ↑
GenFace [67]	177.06	0.49	0.54
DeepFillv1 [69]	241.69	0.32	0.48
EdgeConnect [68]	141.69	0.18	0.45
MS-GAN [55]	46.12	0.75	0.72
DeepFakes [4]	43.03	0.45	0.46
Ours	**30.31**	**0.78**	**0.90**

TABLE VII
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART FACE MANIPULATION METHODS ON CELEBA-HQ [15] DATASET

	FID ↓	*SSIM* ↑	*ID* ↓
GA-GAN [54]	23.71	0.42	-
MG-GAN [64]	**14.81***	0.76*	0.27*
InterFace-GAN [65]	16.15*	**0.81***	0.50*
SCFE-GAN [66]	22.03*	0.61*	0.35*
Ours	19.72	**0.81**	**0.08**

To be fair, for [65] and [66], we only edit facial attributes such as expression, age and glasses, so as not to destroy the identity. Result with * means it is reproduced by us using the official implementation.

V. CONCLUSION

Existing approaches cannot well preserve facial attributes (e.g., pose, expression, skin color, illumination, make-up, occlusion, etc.) of the target face, causing noticeable temporal discontinuity and instability artifacts for video face swapping. To achieve efficient and high-fidelity video face swapping, we propose a lightweight Generative Adversarial Networks based framework named AP-GAN, which can precisely control the attributes of the synthetic face in line with the target face's. Specifically, we derive a U-Net based generator with

ID blocks to translate identity and PE blocks to correct pose and expression. Besides, a PE-aware discriminator is designed to help supervise the pose and expression of the synthetic face. Moreover, we propose a discriminator based perceptual loss leveraging multi-scale features of the discriminator to preserve facial attributes like skin color, illumination, make-up and occlusion. Extensive experiments and comparisons to the existing state-of-the-art methods demonstrate the efficacy of our framework.

REFERENCES

[1] D. Bitouk, "Face swapping: Automatically replacing faces in photographs," in *Proc. SIGGRAPH*, 2008, vol. 27, no. 2, pp. 1–8.

[2] H. Wang, C. Pan, and H. Gong, "Facial image composition based on active appearance model," in *Proc. IEEE Int. Conf. Acoust.*, Mar. 2008, pp. 893–896.

[3] *FaceSwap*. Accessed: Oct. 8, 2020. [Online]. Available: https://github.com/MarekKowalski/FaceSwap

[4] *Deepfakes*. Accessed: Oct. 8, 2020. [Online]. Available: https://github.com/deepfakes/faceswap

[5] Y. Nirkin *et al.*, "On face segmentation, face swapping, and face perception," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2017, pp. 98–105.

[6] J. Bao, D. Chen, and F. Wen, "Towards open-set identity preserving face synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6713–6722.

[7] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards high fidelity and occlusion aware face swapping," 2019, *arXiv:1912.13457*. [Online]. Available: https://arxiv.org/abs/1912.13457

[8] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[9] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.

[10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019.

[11] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2019, *arXiv:1907.10830*. [Online]. Available: https://arxiv.org/abs/1907.10830

[12] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, Jul. 2003.

[13] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3677–3685.

[14] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proc. German Conf. Pattern Recognit.*, 2016, pp. 26–36.

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: https://arxiv.org/abs/1710.10196

[16] R. Andreas *et al.*, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1–11.

[17] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, 2014, pp. 2672–2680.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," Nov. 2015, *arXiv:1511.06434*. [Online]. Available: https://arxiv.org/abs/1511.06434

[19] J. Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2223–2232.

[20] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1125–1134.

[21] Y. Choi, M. Choi, and M. Kim, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun.2018, pp. 8789–8797.

[22] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[23] G. Liu *et al.*, "Image inpainting for irregular holes using partial convolutions," in *Proc. IEEE/CVF Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 85–100.

[24] "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2017, pp. 214–223.

[25] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[26] L. Gatys, A. Ecker, and M. Bethge, "A neural algorithm of artistic style," *J. Vis.*, vol. 16, no. 12, p. 326, Sep. 2016.

[27] T. C. Wang, M. Y. Liu, and J. Y. Zhu, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[29] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*. [Online]. Available: https://arxiv.org/abs/1802.06955

[30] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 424–432.

[31] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-UNet for high-quality retina vessel segmentation," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Oct. 2018, pp. 327–331.

[32] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 568–576, Feb. 2020.

[33] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999

[34] L. B. Jimmy, R. K. Jamie, and E. H. Geoffrey, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: https://arxiv.org/abs/1607.06450

[35] U. Dmitry, V. Andrea, and L. Victor, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: https://arxiv.org/abs/1607.08022

[36] Y. Lin, S. Wang, Q. Lin, and F. Tang, "Face swapping under large pose variations: A 3D model based approach," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 333–338.

[37] R. Natsume, T. Yatagawa, and S. Morishima, "RSGAN: Face swapping and editing using face and hair representation in latent spaces," 2018, *arXiv:1804.03447*. [Online]. Available: https://arxiv.org/abs/1804.03447

[38] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[39] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for realtime style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[40] E. Ilg *et al.*, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2462–2470.

[41] *DeepGlint-Face*. Accessed: Feb. 1, 2020. [Online]. Available: http://trillionpairs.deepglint.com/overview

[42] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[43] T. Miyato *et al.*, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: https://arxiv.org/abs/1802.05957

[44] J. Zhao *et al.*, "Dual-agent GANs for photorealistic and identity preserving profile face synthesis," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 31, 2017.

[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[46] X. Huang *et al.*, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.

[47] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2074–2083.

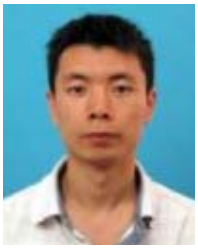
[48] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 285–295.

[49] Q. Cao *et al.*, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2018, pp. 67–74.
[50] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7184–7193.
[51] T. T. Nguyen *et al.*, "Deep learning for deepfakes creation and detection: A survey," 2019, *arXiv:1909.11573*. [Online]. Available: https://arxiv.org/abs/1909.11573
[52] X. Ding, Z. Raziei, E. C. Larson, E. V. Olinick, P. Krueger, and M. Hahsler, "Swapped face detection using deep learning and subjective assessment," *EURASIP J. Inf. Secur.*, vol. 2020, no. 1, pp. 1–12, Dec. 2020.
[53] S. Akhil, "Deepfakes generation using LSTM based generation adversarial networks," Doctoral dissertation, Rochester Inst. Technol., Rochester, NY, USA, 2020.
[54] D. Huang, X. Tao, J. Lu, and M. N. Do, "Geometry-aware GAN for face attribute transfer," *IEEE Access*, vol. 7, pp. 145953–145969, 2019.
[55] S. Banerjee, W. J. Scheirer, K. W. Bowyer, and P. J. Flynn, "On hallucinating context and background pixels from a face mask using multi-scale GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 300–309.
[56] J. Cai, H. Hu, S. Shan, and X. Chen, "FCSR-GAN: End-to-end learning for joint face completion and super-resolution," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.
[57] A. Gokaslan *et al.*, "Improving shape deformation in unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 649–665.
[58] J. Tang, Z. Li, H. Lai, L. Zhang, and S. Yan, "Personalized age progression with bi-level aging dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 905–917, Apr. 2018.
[59] X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan, "Personalized age progression with aging dictionary," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3970–3978.
[60] S. Zhang *et al.*, "SFD: Single shot scale-invariant face detector," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 192–201.
[61] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial Landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.
[62] C. Fu *et al.*, "High fidelity face manipulation with extreme poses and expressions," *IEEE Trans. Inf. Forensics and Security*, vol. 16, pp. 2218–2231, 2021.
[63] G. B. Huang *et al.*, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, 2008.
[64] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, "Mask-guided portrait editing with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3436–3445.
[65] Y. Shen *et al.*, "InterFaceGAN: Interpreting the disentangled face representation learned by GANs," *IEEE Trans. Pattern Anal. Mach. Intell.*
[66] Y. Jo and J. Park, "SC-FEGAN: Face editing generative adversarial network with user's sketch and color," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1745–1753.
[67] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3911–3919.
[68] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," 2019, *arXiv:1901.00212*. [Online]. Available: https://arxiv.org/abs/1901.00212
[69] R. A. Yeh, C. Chen, and T. Y. Lim, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5485–5493.
[70] L. Liu, S. Li, and C. L. P. Chen, "Iterative relaxed collaborative representation with adaptive weights learning for noise robust face hallucination," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1284–1295, May 2019.
[71] X. Cheng, J. Lu, B. Yuan, and J. Zhou, "Identity-preserving face hallucination via deep reinforcement learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4796–4809, Dec. 2020.
[72] L. Yu *et al.*, "Multimodal inputs driven talking face generation with spatial-temporal dependency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 203–216, Jan. 2021.
[73] X. Tu *et al.*, "3D face reconstruction from a single image assisted by 2D face images in the wild," *IEEE Trans. Multimedia*, vol. 23, pp. 1160–1172, 2021.
[74] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-aided dual-agent GANs for unconstrained face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, Oct. 2019.
[75] J. Xin *et al.*, "Residual attribute attention network for face image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 9054–9061.
[76] D. Liu, X. Gao, N. Wang, C. Peng, and J. Li, "Iterative local re-ranking with attribute guided synthesis for face sketch recognition," *Pattern Recognit.*, vol. 109, Jan. 2021, Art. no. 107579.
[77] D. Liu, X. Gao, N. Wang, J. Li, and C. Peng, "Coupled attribute learning for heterogeneous face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4699–4712, Nov. 2020.
[78] D. Liu, N. Wang, and C. Peng, "Deep attribute guided representation for heterogeneous face recognition," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 835–841.
[79] J. Zhao, J. Xing, L. Xiong, S. Yan, and J. Feng, "Recognizing profile faces by imagining frontal view," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 460–478, Feb. 2020.
[80] J. Zhao *et al.*, "3D-aided deep pose-invariant face recognition," in *Proc. IJCAI*, 2018, vol. 2, no. 3, p. 11.

Longhao Zhang is currently pursuing the Ph.D. degree in pattern recognition with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. He is particularly interested in the areas of generative model and weakly supervised object localization. His research interests include computer vision, machine learning, and deep learning.

Huihua Yang (Member, IEEE) received the Ph.D. degree from the East China University of Science and Technology, China, in 2005. From 2005 to 2007, he was a Post-Doctoral Research Fellow with Tsinghua University. He is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. He has published more than 40 articles. His research interests include machine learning, spectrum analysis, and optimization. He is a Senior Member of CCF and a member of ACM. He serves as the Director for the China Instrument and Control Society (CICS) and the Vice Director for the NIR Division of CICS.

Tian Qiu received the B.E. degree in opto-electronics engineering from the Beijing Institute of Technology, Beijing, China, in 2019. He is currently pursuing the M.S. degree in electrical and computer engineering with the University of California at San Diego. His research interests include computer vision, image and video processing and compression, and machine learning.



Lingqiao Li received the Ph.D. degree from the School of Automation, Beijing University of Posts and Telecommunications, China, in 2020. He is currently an Assistant Researcher with the Guilin University of Electronic and Technology, China. His research interests include pattern recognition and large spectral data analysis.