Dense 3D Coordinate Code Prior Guidance for High-Fidelity Face Swapping and Face Reenactment

Anni Tang[1], Han Xue[1], Jun Ling[1], Rong Xie[1], Li Song[1,2] ✉
[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China
[2] MOE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
{memory97, xue_han, lingjun, xierong, song_li}@sjtu.edu.cn

(a) Face Swapping (b) Face Reenactment (c) Expression Editing (d) Pose Editing

Fig. 1: Our proposed method can perform: (a) Face swapping: swap source face onto the target; (b) Face reenactment: animate the source by the target; (c) Expression editing: edit the expression alone; (d) Pose editing: edit the pose alone.

***Abstract*— In face synthesis tasks, commonly used 2D face representations (e.g. 2D landmarks, segmentation maps, etc.) are usually sparse and discontinuous. To combat these shortcomings, we utilize a dense and continuous representation, named Projected Normalized Coordinate Code (PNCC), as the guidance and develop a PNCC-Spatio-Normalization (PSN) method to achieve face synthesis regarding arbitrary head poses and expressions. Based on PSN, we provide an effective framework for face reenactment and face swapping task. To ensure a harmonious and seamless face swapping, a simple yet effective Appearance-Blending Module (ABM) is proposed to fit the synthesized face to the target face. Our method is subject-agnostic and can be applied to any pair of faces without extra fine-tuning. Both qualitative and quantitative experiments are conducted to demonstrate the superiority of the proposed method in comparisons to existing state-of-the-art systems.**

I. INTRODUCTION

Photo-realistic face synthesis is an emerging research topic in the field of computer vision and graphics, in which face swapping and face reenactment are two promising subtasks. Face swapping aims at transferring the identity from a source face to a target face, while face reenactment utilizes the pose and expression of a target face to animate the source face. More and more studies have been investigated due to their promising applications in entertainment, privacy, virtual reality and video dubbing, etc.

In the task of face swapping/reenactment, 2D face representations (e.g. facial landmarks and face segmentation maps) are widely adopted as the guidance. For example, FSGAN [14] utilizes 2D facial landmarks as condition to guide the face synthesis, but it suffers from over-smooth results because facial landmarks are too sparse to accurately guide the synthesis of faces. Furthermore, these 2D representations cannot effectively disentangle the facial attributes (e.g. identity, pose and expression). Stuck by this problem, FOM [17] tends to generate severely distorted results when performing cross-identity face reenactment. Instead of using these 2D representations, some methods try to fuse the source and target information in the latent space to synthesize the desired faces. For example, FaceShifter [10] encodes the identity of the source image and the attribute of the target image into latent space and then merges these two encodings to synthesize the final result. Despite the impressive results, these encodings may not accurately represent the desired information and these methods lack controllability and flexibility in generation.

To tackle the aforementioned problems, we propose to utilize a denser representation of human faces, called Projected Normalized Coordinate Code (PNCC) [25], to guide the synthesis of vivid human faces and effectively achieve the disentanglement of identity and other attributes based on 3DMM [18]. PNCC is a kind of 2D face representation based on 3D face reconstruction, which has the characteristics of density and continuous variations in adjacent regions. Disappointingly, in previous works, PNCC was mostly used as an auxiliary representation for frame generation and how to use it to provide better guidance was not fully explored. For example, in Deep Video Portraits [8], PNCC and face texture images are cascaded together as condition to synthesize portrait faces, which does not give full play to the expression ability of PNCC.

To make full use of PNCC, we utilize PNCC as the only representation of human faces and develop a PNCC-Spatio-Normalization (PSN) method for facial image generation network to realize face synthesis under arbitrary poses and expressions. Based on the PSN method, a coarse-to-fine Vivid-Face-Rendering-Network (VFRN) is proposed to synthesize vivid faces which plays an important role in both face swapping and reenactment. In face swapping, the synthesized face in VFRN possesses not only the source identity information but also its skin color, leading to inconsistency between the target actor and the generated face. To tackle

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

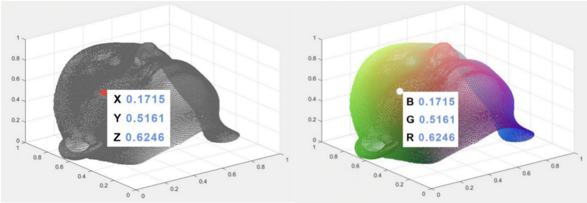

Fig. 2: *The Normalized Coordinate Code (NCC).* Let NCC be the texture of the normalized mean face (NCCx = B, NCCy = G, NCCz = R).

this problem, a simple yet effective Appearance-Blending Module (ABM) is proposed to adaptively adjust the skin color of the generated face so that it can be seamlessly integrated into the face of the target actor.

In general, this paper proposes a holistic approach to accomplish the goal of both face swapping and face reenactment. We fully exploit the representation potential of PNCC and conduct qualitative and quantitative experiments to demonstrate the superiority of our method on both tasks. Besides, comprehensive ablation studies also demonstrate that PNCC is more informative than other 2D representations and can be used as an accurate semantic map of human faces.

The contributions of this work can be summarized as: (1) We propose a holistic pipeline to jointly implement face swapping and face reenactment and achieve the state-of-the-art performance. (2) We propose to utilize the Projected Normalized Coordinate Code (PNCC) as the face representation and develop a PNCC-Spatio-Normalization (PSN) method to achieve face synthesis under arbitrary head poses and expressions. (3) A simple yet effective Appearance-Blending Module (ABM) is proposed to seamlessly fit the synthesized face into the target frame.

II. Related Work

A. 3D Morphable Model (3DMM)

Blanz et al. [18] proposed the 3D morphable model (3DMM) which describes the 3D face space with PCA:

$$\mathbf{S} = \overline{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}, \tag{1}$$

where $\mathbf{S}$ is a 3D face, $\overline{\mathbf{S}}$ is the mean face, $\mathbf{A}_{id}$ is the principle axis extracted from the 3D face scans with neutral expression and α_{id} is the corresponding shape parameter. $\mathbf{A}_{exp}$ is the principle axis extracted from the offsets between expression scans and neutral scans and α_{exp} is the corresponding expression parameter. We obtain $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ from BFM [7] and FaceWarehouse [2] respectively. The 3D face is projected onto the 2D plane through Perspective Projection:

$$V(\mathbf{p}) = f * \mathbf{Pr} * \mathbf{R} * \left(\overline{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{\mathrm{exp}}\alpha_{exp}\right) + \mathbf{t}_{2d}, \tag{2}$$

where $V(\mathbf{p})$ is the reconstruction and projection function, f is the scale factor, $\mathbf{Pr}$ is a known and fixed orthographic projection matrix, $\mathbf{R}$ is the rotation matrix computed from three rotation angles *pitch, yaw, roll* and $\mathbf{t}_{2d}$ is the translation vector. All parameters to be predicted include shape parameters α_{id}, expression parameters α_{exp}, and pose parameters f, $\mathbf{t}_{2d}$, *pitch, yaw, roll*.

Projected Normalized Coordinate Code (PNCC) Many facial feature maps have been designed since 3DMM was proposed. 3DDFA [25] applied a network to regress 3DMM parameters to realize dense face alignment, and proposed the Projected Normalized Coordinate Code (PNCC) which is a dense semantic representation of human face. Following [25], as shown in Fig. 2, we normalize the vertex coordinates of the 3D mean face, allowing the 3D coordinates of each vertex to be represented uniquely within the interval of [0,0,0] and [1,1,1]. By treating the normalized vertex coordinates (x,y,z) as the (B,G,R) color value of that vertex, each vertex in the 3D face model corresponds to a defined color which is called Normalized Coordinate Code (**NCC**). According to (3), the 3D face can be reconstructed from the estimated parameters $\mathbf{p}$ and projected to 2D plane with the Z-Buffer algorithm to render the PNCC image based on NCC color.

$$\text{PNCC} = Z\text{- Buffer}\left(V_{3d}(\mathbf{p}), \text{NCC}\right). \tag{3}$$

B. Face Swapping

GAN-based methods are the mainstream of face swapping algorithms. Subject-specefic methods [1], [9] need to be retrained each time a new actor is encountered. This limitation has been addressed by subject-agnostic face swapping methods such as FSNet [13], FSGAN [14] and FaceShifter [10]. FSGAN [14] performs face reenactment and face swapping together, but it hardly preserves the face shape and identity of the source actor. FaceShifter [10] is able to synthesize high fidelity results, but it sometimes fails to conduct face swapping due to the failure of face detection and tends to generate striped artifacts. In this paper, we propose a novel subject-agnostic method based on PNCC for face swapping.

C. Face Reenactment

Early methods consider the face reenactment task as a mapping problem between the source and target image domains. For example, Xu et al. [21] achieves face reenactment between two persons by adopting CycleGAN [24]. These methods require adequate images of source and target persons to learn a mapping network, which greatly limits their practical application. Some methods utilize face landmarks as the representation of target pose and expression to perform face reenactment [14], [16], [20]. Recently, methods utilizing 3D face reconstruction and neural rendering to implement face reenactment [3], [8], [22] have become prominent because of its high disentanglement and controllability. Our method is based on 3D face reconstruction to achieve flexible face reenactment.

III. Approach

To perform face swapping and reenactment simultaneously, we decompose our method into three stages: 3D face reconstruction, face rendering and blending. In the first stage, we perform 3D face reconstruction and feature rendering to prepare data. In the second stage, a Vivid-Face-Rendering-Network (VFRN) is utilized to synthesize the high-fidelity faces in both tasks, and the VFRN utilized in both tasks shares weights. Then in the third stage, the two tasks apply different blending modules to obtain the final results.

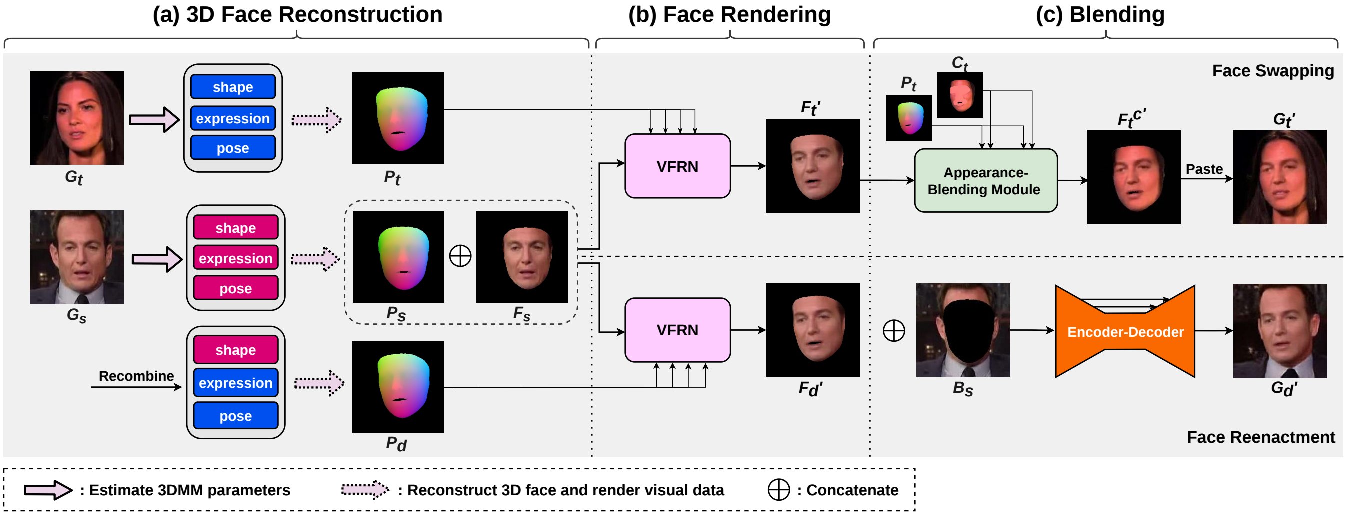

Fig. 3: *Method Overview.* The pipeline consists of three stages: (a) 3D face reconstruction: estimate 3DMM parameters, reconstruct 3D face and render visual data for the next stages; (b) Face rendering: apply VFRN to synthesize the high-fidelity face according to the conditional PNCC (P_t or P_d); (c) Blending: fuse F_t' to G_t seamlessly in face swapping / make up the background region of F_d' in face reenactment.

A. Overall Pipeline

As depicted in Fig. 3, generally, given a target frame G_t and a source frame G_s as input, the face swapping result G_t' and the face reenactment result G_d' can be achieved through the proposed pipeline. Here, we give the definition of source frame and target frame in different tasks. In face swapping, our purpose is to transfer the face details from the source frame to the target frame, letting the swapping results hold the identity of the source frame. In face reenactment, we aim to let the actor in target frame drive the actor in source frame, ensuring the driving results preserve the identity of source frame and the pose and expression of target frame.

In the first stage, we perform parametric 3D face reconstruction and feature rendering to obtain visual data for the next stages such as the target PNCC P_t, the source PNCC P_s and the source face F_s. We recombine the parameters of the source actor and the target actor to reconstruct the driven 3D face and render the corresponding PNCC P_d, preventing the identity leakage when performing face reenactment.

In the second stage, we feed the concatenation of F_s and P_s and a conditional PNCC (P_t or P_d) to VFRN to obtain the corresponding synthesized face, which is characterized by the face shape, pose and expression of the conditional PNCC and the identity of the source actor. In face swapping, the conditional PNCC is P_t and the synthesized face is F_t'. In face reenactment, the conditional PNCC is P_d and the synthesized face is F_d'.

The tasks of face swapping and reenactment utilize different blending modules. To realize photo-realistic face swapping, an Appearance-Blending Module (ABM) is designed to achieve color adaptation between the synthesized face F_t' and the target frame G_t, that is, adjust the color tone of F_t' to $F_t^{c\prime}$ which is compatible with G_t. The final face swapping result G_t' is obtained by pasting $F_t^{c\prime}$ to G_t automatically.

To implement face reenactment, after obtaining the driven face F_d' synthesized by VFRN, we use a blending network to make up the background region. The input of the blending network is the concatenation of the background region of the source image B_s and the synthesized face F_d', and the output is the final driven result G_d'.

B. 3D Face Reconstruction

In this part, we apply a pretrained model released by [4] to regress the shape, expression and pose parameters of the source actor and the target actor. Then, we reconstruct the 3D face according to the estimated parameters and render the PNCC image of each actor, denoted as P_t and P_s. To realize face reenactment, we recombine the shape parameters of the source actor and the pose and expression parameters of the target actor to reconstruct the driven source face and render the corresponding PNCC P_d, which plays an important role as the conditional PNCC in VFRN when performing face reenactment. It possesses the target pose and expression while maintaining the source identity. The disentanglement of shape, expression and pose ensures the avoidance of identity leakage when performing cross-identity face reenactment.

Moreover, BFM [7] database provides semantic label of each vertex in 3DMM, based on which we divide all vertices into four regions: eye, nose, mouth and the rest, and then render the face segment image like the image in the last row, third column of Fig. 9. With the face segment image, we can easily fetch the facial mask image (a 0-1 matrix, 1:face area, 0:other areas), and then multiply this mask by the original image to get a face image without background like F_s in Fig. 3. These data is useful in the following stages.

C. Vivid-Face-Rendering-Network (VFRN)

To synthesize photo-realistic faces with source identity and target attributes, we propose a Vivid-Face-Rendering-Network (VFRN). Without the loss of generality, we take face swapping as an example to introduce our VFRN. As shown in Fig. 4, the input of the whole network is the

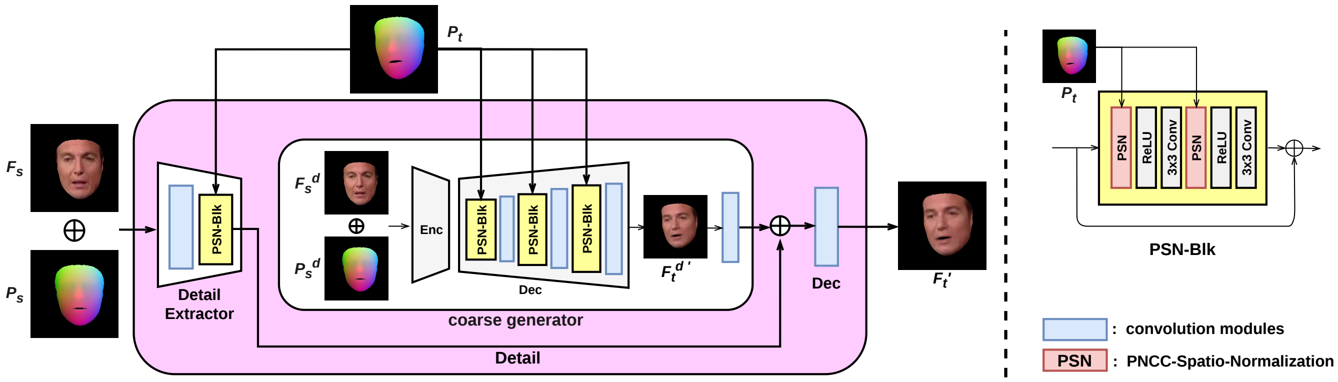

Fig. 4: *Vivid-Face-Rendering-Network (VFRN).* VFRN applies a coarse-to-fine architecture trained in two stages: in the first stage, a coarse generator is trained to get the coarse result $F_t^{d\prime}$ and in the second stage, a detail branch is trained to supplement the details and get the final result F_t'.

concatenation of F_s and P_s and the target PNCC P_t. In brief, VFRN applies a coarse-to-fine architecture to ensure stable and fine-grained synthesized results. Accordingly, the model is optimized in two stages.

In the first stage, we train a coarse generator with an encoder-decoder architecture to synthesize a coarse-grained face $F_t^{d\prime}$ which generally possesses the correct head pose and skin color but lacks details. We use the downsampling of F_s and P_s denoted as F_s^d and P_s^d as input to encode the information of the source face. In the decoder, we design a PSN-Blk, a residual block, whose detailed structure is shown on the right of Fig. 4. The core of PSN-Blk is the PNCC-Spatio-Normalization (PSN) method which embeds the PNCC with target pose and expression information and guides the synthesis of corresponding source face. The detailed structure of PSN is shown in Fig. 6. We use the target PNCC P_t to predict γ and β in normalization to realize the spatial adjustment of feature maps. Compared with facial landmarks or face semantic segments, PNCC representation exhibits denseness and continuity, thereby providing more accurate semantic information of the target face.

In the second stage, we train a detail branch to supplement details to synthesize fine-grained and photo-realistic faces. The detail extractor takes F_s and P_s as input and uses a PSN-Blk which takes P_t as input to provide target attribute information for the encoded feature maps. We concatenate these features with those feature maps encoded from $F_t^{d\prime}$ and use a transpose convolution layer to get the final output F_t'. Jointly combining the advantages of PSN with the two-stage training strategy, we synthesize the high-fidelity faces F_t' with target attributes and source facial details. In face reenactment task, we replace P_t with P_d as the conditional PNCC and obtain the corresponding coarse face $F_d^{d\prime}$ and vivid face F_d'.

D. Appearance-Blending Module (ABM)

In face swapping, the face synthesized by VFRN (F_t') is not color-consistent with the target face (G_t). To solve the problem of unmatched skin color between F_t' and G_t, we propose a simple yet effective Appearance-Blending Module (ABM) (shown in Fig. 5) to achieve the adaptive adjustment of skin color. In particular, we use a color map C_t to provide

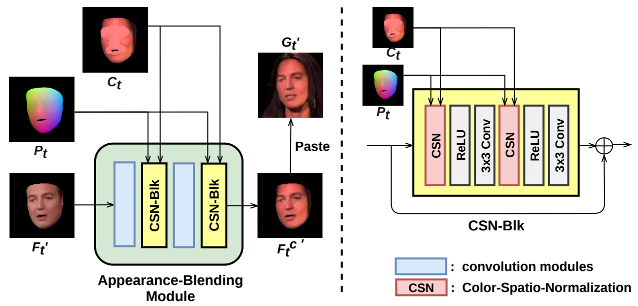

Fig. 5: *The framework of Appearance-Blending Module (ABM).*

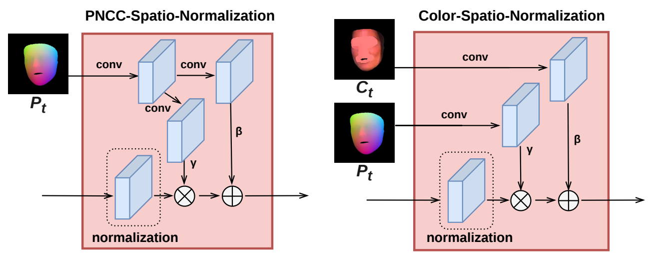

Fig. 6: *The detailed structure of PNCC-Spatio-Normalization method (PSN) and Color-Spatio-Normalization method (CSN).*

the information of target skin color. To generate the color map C_t, we first set the pixel values of the eyes, nose, and mouth regions to 0 in the target face image F_t and then use *dilation* operation in morphological filtering to fill the region of eye, nose and mouth.

The core of the Appearance-Blending Module (ABM) is the Color-Spatio-Normalization (CSN) method, as shown in Fig. 6. Different from PSN, CSN takes two different inputs to predict the bias parameters β and scaling parameters γ, respectively. Specifically, it takes a PNCC image P_t to predict the scaling parameter γ and a color map C_t to predict the bias parameter β to achieve the purpose of adjusting the overall skin color through the color map. In this way, we can adjust the skin color of F_t' to $F_t^{c\prime}$ while retaining the spatial outlines in F_t'. The experimental results shown in Sec. IV-B demonstrate the effectiveness of this method.

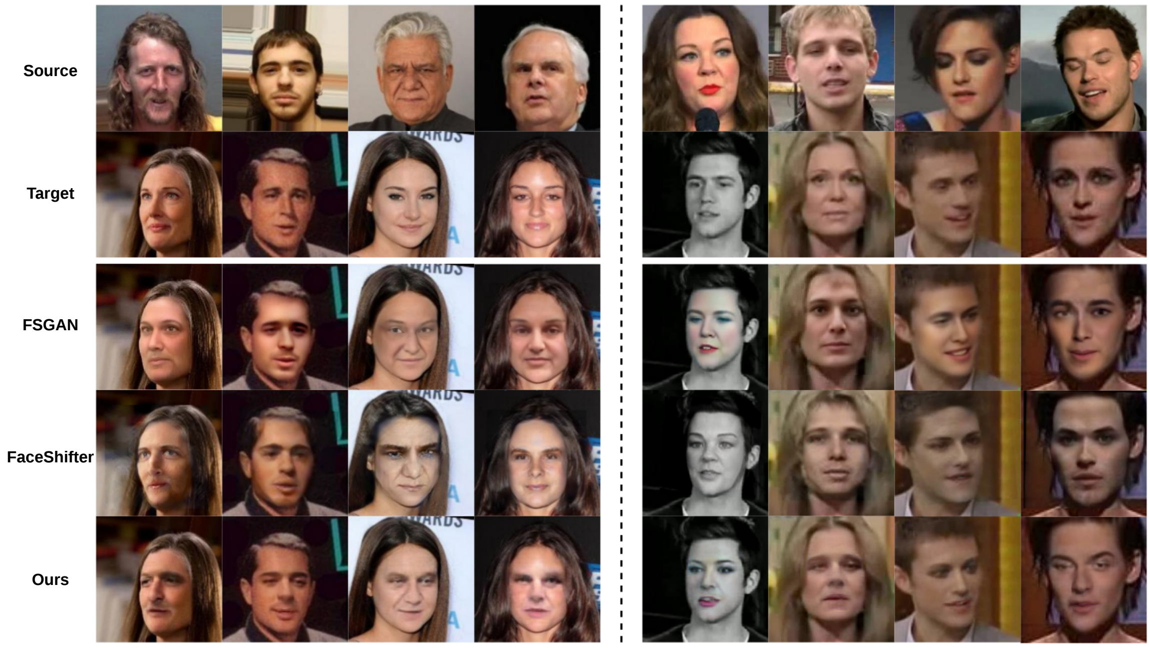

Fig. 7: *Qualitative face swapping results compared with FSGAN [14] and FaceShifter [10] on different datasets.* Left: CelebA [11], Right: VoxCeleb1 [12].

TABLE I: *Evaluation of face swapping on CelebA [11].*

Method	FID↓	CSIM↑	User Study		
			id.	attr.	real.(%)
FSGAN [14]	65.70	0.3733	49.30	**93.15**	35.21
F.S. [10]	54.51	**0.5545**	62.80	83.15	17.60
Ours	**36.77**	0.4468	**66.52**	89.15	**47.19**

E. Loss Functions

VFRN We use an L1 loss (4) to train the coarse generator to synthesize a relatively smooth and coarse face image which contains correct pose and skin color:

$$L_{crs} = \|F_t^d - F_t^{d\prime}\|_1. \tag{4}$$

To optimize the detail branch, we use a region-aware L1 loss (5) to pay more attention to the key organs in a face than to the rest, a perceptual loss (6) to recover more details and an adversarial loss (7) to improve the visual reality:

$$L_{rgn} = M \cdot \|F_t - F_t'\|_1, \tag{5}$$

where M refers to the weight mask in which we set the weights of eye, nose, mouth and the rest to 4, 3, 3, 2, and face segment image can work as the weight mask.

$$L_{perc}^{VFRN} = \sum_{i=1}^{n} \frac{1}{C_i H_i W_i} \left\| \mathrm{VGG}_i\left(F_t\right) - \mathrm{VGG}_i(F_t') \right\|, \tag{6}$$

where i is the selected layer indexes of VGGFace.

$$L_{adv}^{VFRN} = \|1 - D_1(P_t, F_t')\|_2^2. \tag{7}$$

In the detail branch, the total loss for the generator is (8) and loss for the discriminator is (9):

$$L_G^{VFRN} = \lambda_1 \cdot L_{rgn} + \lambda_2 \cdot L_{perc}^{VFRN} + \lambda_3 \cdot L_{adv}^{VFRN}, \tag{8}$$

$$L_D^{VFRN} = \lambda_4 \cdot (\|D_1(P_t, F_t')\|_2^2 + \|1 - D_1\left(P_t, F_t\right)\|_2^2), \tag{9}$$

where λ_1=3, λ_2=2, λ_3=λ_4=0.2.

Face Swapping To train ABM, we use an L1 loss to ensure the synthesized face to have the skin color and edge texture of the target face:

$$L_{color} = \|F_t^{c\prime} - F_t\|_1. \tag{10}$$

To maintain the same facial features before and after the skin color modification, we employ a perceptual loss in the organ regions using the same VGGFace network as (6):

$$L_{perc}^{ABM} = \sum_{i=1}^{n} \frac{1}{C_i H_i W_i} \left\| \mathrm{VGG}_i\left(O_t'\right) - \mathrm{VGG}_i(O_t^{c\prime}) \right\|, \tag{11}$$

where O indicates the organ regions in human faces. We also use an adversarial loss (12) to ensure the reality of the synthesized face:

$$L_{adv}^{ABM} = \|1 - D_2(P_t, F_t^{c\prime})\|_2^2. \tag{12}$$

In ABM, the total loss for the generator is (13) and loss for the discriminator is (14):

$$L_G^{ABM} = \lambda_5 \cdot L_{color} + \lambda_6 \cdot L_{perc}^{ABM} + \lambda_7 \cdot L_{adv}^{ABM}, \tag{13}$$

$$L_D^{ABM} = \lambda_8 \cdot (\|D_2(P_t, F_t^{c\prime})\|_2^2 + \|1 - D_2\left(P_t, F_t\right)\|_2^2), \tag{14}$$

where λ_5=10, λ_6=1, λ_7=λ_8=0.3.

Face Renactment We take frames with the same identity but different poses in one video as paired data to perform supervised learning. The blending network of face reenactment is also trained with L1, perceptual and adversarial loss.

Fig. 8: *Qualitative face reenactment results compared with FSGAN [14], FOM [17] and Bilayer [23] on CelebA [11].*

TABLE II: *Evaluation of face reenactment on CelebA [11].*

Method	FID↓	CSIM↑	User Study		
			id.	attr.	real.(%)
FSGAN [14]	141.52	0.4780	38.49	**93.47**	2.41
FOM [17]	56.47	**0.6550**	55.18	79.71	13.55
Bilayer [23]	224.37	0.3164	79.72	85.06	**43.07**
Ours	**55.6**	0.5289	**81.94**	91.04	40.97

IV. Experiments

Due to the limited space, we illustrate the *experimental setups* including datasets, implementation details and evaluation metrics in the supplementary material[1].

A. Comparison Results

Face swapping. We compare our method with FSGAN [14] and FaceShifter [10]. Fig. 7 shows the qualitative results in which the left part is conducted on CelebA [11] dataset and the right part is conducted on VoxCeleb1 [12] dataset. We can see that FSGAN [14] tends to generate obviously blurred and over-smooth results which makes the results less realistic. FaceShifter [10] has the disadvantage of striped artifacts and it sometimes fails to detect the face in an image, resulting in the decline of robustness. In contrast, our method tends to synthesize more realistic images with stronger robustness. The quantitative results shown in Tab. I indicate that our method has a better performance on most metrics, especially ID retention and visual reality.

Face reenactment. We compare our method with FSGAN [14], First-Order-Model (FOM) [17] and Bilayer-model [23]. In Fig. 8, we can see that FSGAN suffers from over-smooth results and fails to well preserve the source identity. Despite the high CSIM score achieved by FOM, it tends to generate severely distorted results which limits its practical applications. Bilayer-model [23] tends to synthesize images with high reality, yet the ID retention is not that good as ours. Note that these methods directly adopts the target facial landmarks or face as the driving information, leading to the problem of identity leakage. Different from these methods, we utilize the PNCC to animate the source face, which is rendered based on the source identity and target pose and expression, achieving better identity preservation. Tab. II presents the quantitative performance of different methods, our method obtaining leading scores on all metrics.

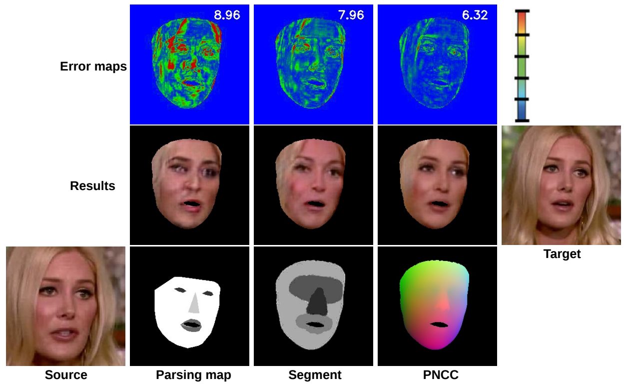

Fig. 9: *Ablation study for different face representations.* Col.2/3/4 corresponds to the result using the face representation of parsing map/face segment/PNCC. The error maps and generated results indicate the superiority of PNCC.

B. Ablation Studies

Why PNCC? To verify the superiority and irreplaceability of PNCC, we replace PNCC with some other representations: parsing maps and face segments, and show the qualitative comparison results in Fig. 9. We also calculate per-pixel Euclidean distance in color space between each synthesized image and the ground truth which is shown on the right-top of each error map in Fig. 9, and then visualize the error in RGB space. It can be observed from the error maps that adopting PNCC leads to the best performance both visually and quantitatively. The quantitative ablation results in Tab. III also suggest that the use of PNCC gives

[1]https://drive.google.com/file/d/1Fy7myvwdeP9dK7lU6B0FzvfJ0lYF4gYK/view?usp=sharing

Fig. 10: *Ablation studies for PSN & the coarse-to-fine design.* Col.3: apply AdaIN architecture in VFRN rather than PSN. Col.4: VFRN without the coarse-to-fine design. Col.5: our proposed VFRN (PSN + coarse-to-fine).

TABLE III: *Quantitative ablation results on VoxCeleb1 [12].* Mask-L1 loss is the region-aware L1 loss introduced in Sec. III-E. The last row shows the visual reality preference.

	Parsing map	Segment	AdaIN [5]	w/o c2f	Ours
L1↓	0.0462	0.0422	0.0465	0.0364	**0.0348**
Mask-L1↓	0.0810	0.0749	0.0818	0.0640	**0.0612**
real.(%)↑	1.69	8.47	5.08	20.34	**64.42**

rise to better performance compared to parsing maps and face segments both in objective and subjective experiments. Representations based on 2D face landmarks like parsing maps are too sparse to synthesize fine-grained images, suffering from performance degradation when dealing with large pose changes. Besides, methods using these representations rely on the accuracy of landmark detection, resulting in the relatively poor robustness. The face semantic representations based on 3D reconstruction do not rely on the detection of face landmarks and therefore are more robust in challenging scenarios. Compared with those sparse semantic representations like face segment image, PNCC is a dense and continuous semantic representation, which has superior performance on guiding high-quality face synthesis.

Why PSN? Using PNCC as the face representation, we applied an AdaIN [5] architecture in VFRN to synthesize faces to demonstrate the superiority of our PSN. The synthesized images using AdaIN are shown in Col.3 of Fig. 10, whose visual quality is obviously worse than ours. Quantitative metrics shown in Tab. III also indicate that our proposed PSN is more suitable for this face synthesis task.

Why coarse-to-fine? To verify the necessity of the coarse-to-fine architecture in VFRN, we remove the coarse-to-fine design, that is, use a generator with the architecture of the coarse generator, to synthesize the final face image. The qualitative comparison is shown in Col.4-5 of Fig. 10 and the quantitative comparison is shown in Tab. III. Without coarse-to-fine architecture, synthesized faces may appear grainy artifacts which impair the visual reality.

Why ABM (CSN)? In Sec. III-D, we designed an ABM to solve the problem of unmatched skin color between F_t' and G_t, that is, Col.2 and Col.1 of Fig. 11 respectively. To verify the rationality of Color-Spatio-Normalization (CSN), we replace PNCC with color map in CSN, obtaining the color adjustment results in Col.10 of Fig. 11. We also remove the estimation of γ, that is, only retain the β estimated by color map to add to the normalized feature maps, obtaining the final results in Col.11 of Fig. 11. In Col.10 and Col.11, the generated faces suffer from unreasonable textures and uneven skin tone, less realistic than ours. In fact, PNCC can provide face semantic information, indispensable for creating a realistic face when merely adjusting the skin color while keeping other features unchanged.

To further demonstrate the superiority of our proposed ABM, we compare our method to frequently used blending methods [1], [15], shown in Col.4-8 of Fig. 11. Deepfakes [1] has provided two blending methods: AdaIN method and histogram matching method, whose results are shown in Col.4-5 of Fig. 11 respectively with poor naturalness. Poisson blending [15] is also an effective method to seamlessly clone an object to another image, which is frequently used as a post-processing procedure in image synthesis tasks [6], [14], [19]. We utilized the *seamlessClone* function provided by OpenCV to implement poisson blending, and there are three modes to select: *MIXED_CLONE*, *MONOCHROME_TRANSFER* and *NORMAL_CLONE*. The corresponding results are shown in Col.6-8 of Fig. 11. The differences between the three modes are illustrated in the supplementary material.

It can be seen from Col.6 that the blending results of *MIXED_CLONE* suffer from poor ID retention because the background texture including the eye, nose and mouth in G_t (Col.1) influences the blending results (Col.6). The results of *MIXED_CLONE* (Col.6) are similar to G_t (Col.1) in terms of identity, which is unacceptable. Comparatively, the results of *MONOCHROME_TRANSFER* and *NORMAL_CLONE* (Col.7-8) are better at ID retention but suffer from poor visual reality, especially the unnatural transition around the edges of the face. Consequently, poisson blending method is not a satisfactory method to implement the seamless fusion of G_t (Col.1) and F_t' (Col.2).

In comparison, the proposed ABM (Col.9) performs well both in ID retention and visual reality. Tab. IV shows the user study results on different blending methods. We evaluate these methods in terms of both ID retention and visual reality, and our method achieves the most satisfactory performance in aggregate. More details about these comparative methods are illustrated in the supplementary material.

V. Conclusion

In this paper, we propose a novel pipeline to achieve face swapping and reenactment. We utilize the Projected Normalized Coordinate Code (PNCC) [25] as the face representation and develop a PNCC-Spatio-Normalization (PSN) method to implement face synthesis under arbitrary head poses and expressions with a high degree of reality and ID retention, exploring the strong representation ability of PNCC. Moreover, our proposed simple yet effective Appearance-Blending Module (ABM) has excellent performance on skin color adjustment, which is very useful for the seamless fusion

Fig. 11: *Ablation study for ABM.* Col.3: directly paste F_t' (Col.2) to G_t (Col.1). Col.4-5: blending methods released by [1]. Col.6-8: poisson blending under different modes. Col.9-11: ablation studies for the design of CSN.

TABLE IV: *A user study to evaluate the blending performance of different methods.* The id score indicates the ID retention between F_t' and the blending result. The reality score indicates the visual reality of the blending result. The last row shows the average score of ID retention and visual reality.

	Paste	AdaIN [1]	Hist. [1]	P.mixed [15]	P.mono [15]	P.normal [15]	w/o PNCC	w/o γ	Ours
id.	-	**95.61**	55.26	14.91	76.63	77.58	36.84	65.79	85.95
real.	2.63	11.84	10.53	81.58	40.21	44.74	40.79	68.42	**92.11**
avg.	2.63	53.73	32.90	48.25	58.42	61.16	38.82	67.11	**89.03**

of face and target background. A series of qualitative and quantitative experiments as well as user studies have been conducted to demonstrate the superiority of our method over existing methods in the field of face swapping and reenactment. We hope this work will inspire more relevant research in the future.

ACKNOWLEDGMENTS

This work was supported in part by MoE-China Mobile Research Fund Project under Grant MCM20180702, National Key R&D Project of China under Grant 2019YFB1802701, and Shanghai Key Laboratory of Digital Media Processing and Transmissions.

REFERENCES

[1] Deepfakes. https://github.com/ondyari/FaceForensics/tree/master/dataset/DeepFakes, 2019. Accessed: 2021-04-20.

[2] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, Mar. 2014.

[3] M. Cao, H. Huang, H. Wang, X. Wang, L. Shen, S. Wang, L. Bao, Z. Li, and J. Luo. Task-agnostic temporally consistent facial video editing, 2020.

[4] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020.

[5] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[6] M. Huh, R. Zhang, J.-Y. Zhu, S. Paris, and A. Hertzmann. Transforming and projecting images to class-conditional generative networks. In *ECCV*, 2020.

[7] IEEE. *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009.

[8] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[9] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *ICCV*, pages 3697–3705, 2017.

[10] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. 12 2019.

[11] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015.

[12] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[13] R. Natsume, T. Yatagawa, and S. Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In C. Jawahar, H. Li, G. Mori, and K. Schindler, editors, *Computer Vision – ACCV 2018*, pages 117–132, Cham, 2019. Springer International Publishing.

[14] Y. Nirkin, Y. Keller, and T. Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE international conference on computer vision*, pages 7184–7193, 2019.

[15] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM SIGGRAPH 2003 Papers*, 2003.

[16] E. Sanchez and M. Valstar. Triple consistency loss for pairing distributions in gan-based face synthesis. *arXiv:1811.03492*, 2018.

[17] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

[18] B. Volker and V. Thomas. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 187–194, 1999.

[19] O. Wiles, A. Sophia Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, pages 670–686, 2018.

[20] W. Wu, Y. Zhang, C. Li, C. Qian, and C. Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, pages 603–619, 2018.

[21] R. Xu, Z. Zhou, W. Zhang, and Y. Yu. Face transfer with generative adversarial network. *arXiv preprint arXiv:1710.06090*, 2017.

[22] G. Yao, Y. Yuan, T. Shao, and K. Zhou. Mesh guided one-shot face reenactment using graph convolutional networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1773–1781. Association for Computing Machinery, 2020.

[23] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, August 2020.

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[25] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 2017.