HR-Net: A Landmark Based High Realistic Face Reenactment Network

Qiuyu Ren, Zhiying Lu, Haopeng Wu, Jianfeng Zhang, and Zijian Dong

***Abstract*— In the past, GAN-based face reenactment methods concentrated mostly on transferring the facial expressions and positions of the source. However, the generated results were susceptible to blurring in some minute details of the face, such as teeth and hair, and their backgrounds were also not guaranteed to be consistent with the manipulated images in terms of light and shadow. Because of these issues, the generated results could be distinguishable as fakes. In this paper, we proposed a landmark based method named HR-Net, which can render source facial expressions and postures on any identity and simultaneously generate realistic face details. Firstly, a lightweight landmark identity conversion module (LIC) was designed to address the identity leakage problem, and it represented facial expressions and poses with only 68 2D landmarks. On this basis, a boundary-guided face reenactment module (BFR) was presented to only learn the background of the reference images; thus, the results generated by BFR can be consistent with the reference images' light and shadow. Moreover, a novel local perceptual loss function was implemented to support the BFR module in generating more realistic details. Extensive experiments demonstrated that our method achieved the state of the art.**

***Index Terms*— Face reenactment, landmark based, image synthesis.**

I. Introduction

FACE reenactment is the task of transferring the pose and expression of one face to another to generate the same stance and emotion. This technology can be extremely beneficial to various industries such as film, virtual reality, teleconferencing, and more. Over the past few decades, numerous research methodologies for face reenactment have been developed. In earlier times, facial manipulation was achieved by leveraging computer graphics. chien, etc [1] uses the optical flow to describe facial action and perform facial animation. The dense optical flow model is capable of mapping both the source and reference structures. However, the computer graphics technology is costly to employ because it requires complex and customized face models.

Manuscript received 4 January 2023; revised 15 March 2023; accepted 5 April 2023. Date of publication 18 April 2023; date of current version 30 October 2023. This work was supported by the Natural Science Foundation of Shandong Province under Grant ZR2021MF020. This article was recommended by Associate Editor J. Gui. *(Corresponding author: Zhiying Lu.)*

Qiuyu Ren, Zhiying Lu, Haopeng Wu, and Zijian Dong are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: qiuyu_ren01@tju.edu.cn; luzy@tju.edu.cn; wuhpfree@sina.com; dongwisdom2000@tju.edu.cn).

Jianfeng Zhang is with the School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China (e-mail: zhjf@tju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3268062.

Digital Object Identifier 10.1109/TCSVT.2023.3268062

Deep learning methods represented by feature extraction networks reduce the high price of CG methods. Monkey-net [2] proposed a motion field to represent the action in the drive frame and extract this motion field using the U-Net architecture. Subsequently, FOMM [3] performs a first-order Taylor expansion of the motion field, considering the motion to be related to the detection point and its local. Then the motion fields play a role in wrapping reference frames to obtain the animated targets. The emergence of GAN [4] networks in deep learning has greatly improved the quality of processed images, and a number of face reenactment methods depending on generative models have been proposed. Early GAN-based methods mostly work on the RGB space of images to learn the actions of source faces and the identities of reference faces. References [5], [6], and [7], all of which supply images or video frames directly into generative models for reenactment. Xu et al. [5] leverage CycleGAN [8] to generate the target character's face image by treating the source and reference images directly as network input. Since the action and identity are not decoupled in their method, the identity of the target generated in this way tends to be an average of the source and reference. This flaw we call "identity leakage."

Later, some people [9], [10] started to combine the latent space with GAN networks to manipulate expressions and identities in the latent space. Liu et al. [11] argues depicting face appearance and action in the latent space allows complete decoupling and controlling both of them separately. Tripathy et al. [12], [13], [14] introduce AU(Action Unit) space to correspond to facial muscle and angles of source images, which decouples the gestures from the source identities. Other researchers [15], [16], [17], [18] establish three-dimensional facial latent representations to tackle the identity leakage problem based on 3DMM (3D Morphable Model), which can reconstruct the expressions on the face. According to these methodologies, the reenacted heads are allowed for larger and more free posture changes. The superior results of these methods come at the cost of the complex overhead of 3DMM. Additionally, [16], [17] do not address transferring the pose orientation of the source face to the target face.

Due to the widespread availability of open-source landmark detection tools, many GAN-based face reenactment methods utilize landmarks' latent space to extract expressions and head poses from the source image. Methods [19], [20], [21] have successfully represented human head action features using landmarks and used GAN networks to generate face images

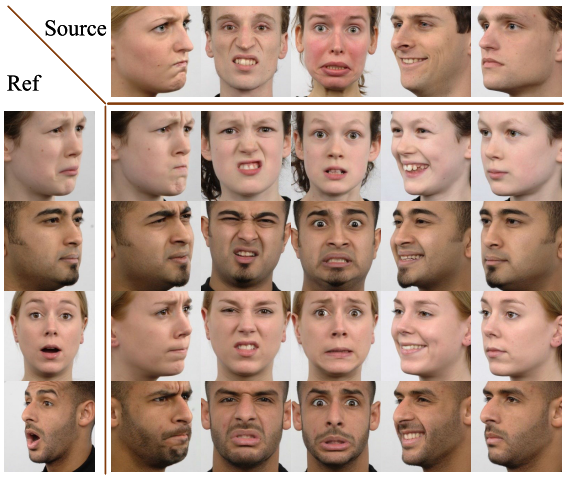

Fig. 1. The top horizontal row of the images are the source face images, and the leftmost column are the reference face images. Our method is able to render any expressions and postures of the source face on the reference face realistically, no matter the original expressions and postures of the sources.

of desired identities from corresponding landmark images. However, landmarks are not fully decoupled from the face identity [16], [17], leading to identity leakage if used directly for generating target faces with different identities. Additionally, because of the sparsity of landmarks, the landmark-based approach generates face images that are challenging to maintain both the local detail of the face and the background of the image.

In conclusion, most current GAN-based reenactment methods use latent representations for better visualization of the target faces, but they struggle to transfer the pose of the source face to the target face, and some of them can be complex and challenging to train. Furthermore, the current approach of using landmarks primarily addresses the identity leakage problem but ignores detailed rendering, resulting in generated images with inferior local detail. For example, the face appears blurred with hair and gaps in the teeth, or the background appears distorted. This can lead to decreased audience interest in the generated images [22], as per the Valley of Terror theory [23], [24].

To address the issues mentioned above, we propose a novel landmark based face reenactment method for freely transferring the pose and expression of a source face to a target face with reference identity. Firstly, we pass the landmarks extracted from the source faces through a lightweight landmark converter to remove their source identities, and introduce a landmark loss function to accelerate successful module training. Secondly, to address the issue of sparse spatial landmarks, we implement a data augmentation measure for modified landmarks to obtain target boundary images. We then introduce a boundary-guided face generator that uses boundary and reference images in an adversarial generative manner to obtain realistic reenacted images. Finally, with the purpose of addressing the lack of generated details in the complex structure, we propose a novel local perceptual loss function to optimize the local facial detail quality in the generated images. Fig. 1 shows that our approach can effectively generate realistic, detailed architecture while reenacting arbitrary facial expressions and poses.

In summary, the main contributions of HR-Net are as follows:

- We implement a light landmark identity conversion (LIC) module, which can keep both the expression and pose of the source face. It addresses the problem of "identity leakage", where different identities are reenacted to each other.
- The newly introduced boundary-guided face reenactment (BFR) module combines boundary images and reference images at multiple scales to generate target images. Thus the target image has the same light and shadow as the reference image and is consistent with the boundary image gestures.
- The modified landmark loss function is adapted to raise the performance of the landmark identity conversion (LIC) module. And also, we propose a novel local perceptual loss that helps the method to generate face images with excellent background and foreground details maintained directly without segmenting the images.
- Extensive experiments on experimental datasets and field datasets demonstrate HR-Net reaches the state-of-the-art in recent years and preserves the light and shadow of the reference images nicely.

II. Related Work

A. Landmarks Based Reenactment

The majority of the methods [10], [19], [20], [25], [26], [27] use 2D landmarks for representation, while the rest [21], [28] use 3D landmarks. ReenactGAN [29] was the first generative reenactment method for extracting human facial expressions based on landmarks. However, its target-specific decoder component limits its ability to retrain a specific boundary-to-image decoder each time a new reference face is selected. Ha et al. [28] noticed that the residual identity from source landmarks would still lead to identity leakage issues, especially when performing between different identities. To combat identity leakage, Sungjoo's work relies on a vast amount of prior knowledge, which is built on its few-shot learning. While [28] can accurately mimic the mouth movements of the characters using predicted landmarks, it cannot replicate large gestures in action.

FReeNet [19] is a method that transfers expressions represented in landmark latent space from source faces to different identities, and it proposes a landmark converter to overcome the identity leakage problem. However, the method cannot render the source head direction on the target face. Li-Net [20] achieves pose transfer by adding a face image rotation module. However, the landmark converter in Li-Net fails to preserve the poses of source images, and the additional rotation module needs to be trained separately, thus increasing the complexity of the method. DualGAN [21] induces a recognition pre-training network in the landmark converter, which removes the softmax layer and uses the output features of the last layer as the identification characteristics of the reference face. It employs a discriminator and a classifier to assist

the converter in stripping source identities while retaining both expressions and poses of source images. Huang et al. [10] does not output the modified target landmark via the decoder. Instead, the method encodes the source landmark and the reference landmark separately into latent codes and then utilizes them as two inputs for the subsequent module. Fu et al. [30] obtain a pose vector by computing the source landmark and an expression vector by computing the source AU. Using two independent representations, they describe all the motions of the source face and ultimately reproduce the entire source face.

Our work is closely related to the methods proposed in [19] and [25]. Furthermore, our method can manipulate the direction of the target head to be in line with the source, which is advanced.

B. Image Synthesis

The emergence of Generative Adversarial Networks (GAN) [4] has provided a new approach for image synthesis. However, the original GAN network was difficult to train and caused pattern collapse. This issue was addressed by Wasserstein GAN (wGAN) [31], which used the Wasserstein distance to calculate the distance between the generated data distribution and the real data distribution. To enhance the controllability of generated images, conditional GAN (cGAN) [32] was proposed, which added conditional restrictions in addition to noisy vectors.

StyleGAN [33], [34] proposed a style-based generator that can decouple the code controlling the image attributes in an intermediate latent space. This method demonstrated that the code can determine the different styles of the generated results. Inspired by cGAN and StyleGAN, the face reenactment topic can also be considered as a conditional generation problem. Our idea to solve this task is to decouple the expression and pose features to control them.

In [35], the conditional GAN's generator was applied to image translation. Different styles of faces were treated as different domains, and conditional GAN was used to transform the faces into different domains to achieve good performance. In image translation, a series of methods [8], [36], [37], [38], [39], [40] perform superior. For example, Pix2Pix [36] achieves impressive results in paired image datasets by using L1 and adversarial losses. Wang et al. [37] create high-resolution images by employing a progressive training method with the same utilization of paired datasets. CycleGAN [8] achieves translation between different styles of images without paired images by using two mirror generators and the proposed consistency loss.

StarGAN-1 [39], which has a loop shared generative network structure, can transform different domains with just one generator by giving different labels. StarGAN-2 [40] designs a style extractor instead of labels like those employed in StarGAN-1. The extractor in StarGAN-2 can learn the style of a domain of images and then generate images with the same style. In our study, we separate RGB images and boundary images into two distinct style domains, considering the RGB image format and the boundary image format of the same action face as two different styles.

III. Proposed Framework

This section introduces the modules of HR-Net and the loss functions of each module. The overall framework and the full reenactment's progression are shown in Fig. 2. As shown in Fig. 2, the model consists of two modules, the landmark identity conversion(LIC) module and the boundary-guided face reenactment(BFR) module. These two modules are trained sequentially. For the sake of convenience, we note the source image, which provides expression and pose, as I_s ($\in \mathcal{R}^{3\times256\times256}$), and we note the reference image, which provides identity information, as I_r ($\in \mathcal{R}^{3\times256\times256}$). First, using the current state-of-the-art open source face detection toolkit [41] FNet, we extract from I_s and I_r to obtain their landmarks, which are denoted as l_s ($\in \mathcal{R}^{2\times68}$) and l_r ($\in \mathcal{R}^{2\times68}$) respectively. The majority of facial identity content on I_s is eliminated because of FNet processing, and the desired expression and pose from I_s is embedded in landmark latent space. To address identity leakage in the LIC module, the l_s and l_r are input to a landmark generator G_L, then the G_L outputs the modified target landmarks $\hat{l}_t$ ($\in \mathcal{R}^{2\times68}$) with l_r identity and l_s gesture. Second, $\hat{l}_t$ is subsequently processed to the target boundary image $\hat{b}_t$ ($\in \mathcal{R}^{3\times256\times256}$), with the goal of allowing $\hat{b}_t$ to carry more of the same identity as the I_r than I_s. The obtained $\hat{b}_t$ is stored as RGB image format like I_r. In the BFR module, $\hat{b}_t$ and I_r are sent to a novel boundary-guided face generator G_F in image format, so the G_F could extract the action features of $\hat{b}_t$ and the style features of I_r. After that, the two different features are multi-scale fusion by adaptive instance normalization (AdIN). The fused feature map is then passed through an image activation layer, yielding the target image $\hat{I}_t$ ($\in \mathcal{R}^{3\times256\times256}$) with I_s expression and pose and I_r identity. Moreover, in order to solve the problem of detailed textures such as teeth and hair sticking in the generated images, an improved local perceptual loss is proposed to optimize this existing problem.

A. Landmark Identity Conversion Module

Despite the fact that l_s has removed most of the identity content of I_s in landmark latent space, the coordinate distribution of l_s still reflects the source's facial geometry, such as the contour of the face or the bridge of the nose, which can reflect the identity of the source. In other words, the source's action and identity are not fully decoupled. If we can not obtain independent Gaussian distributions of action features in the landmark latent space, the entangled action and identity features can lead to a low upper bound on final model performance. That is, some generated $\hat{I}_t$ may not achieve the desired results. Since it is impossible for landmark coordinates to carry no identity information at all, one effective strategy to accomplish this objective is to devise a converter that enables $\hat{I}_t$ to preserve the same actions as the source, while also adopting the geometric identity of the landmark l_r. Depending on this, the problem of "identity leakage" would be addressed.

The LIC module is depicted in the upper portion of Fig. 2. The module contains a landmark generator G_L and a discriminator D_L, and it uses generative adversarial training to enhance the quality of the target $\hat{l}_t$. Inspired by the design of mapping

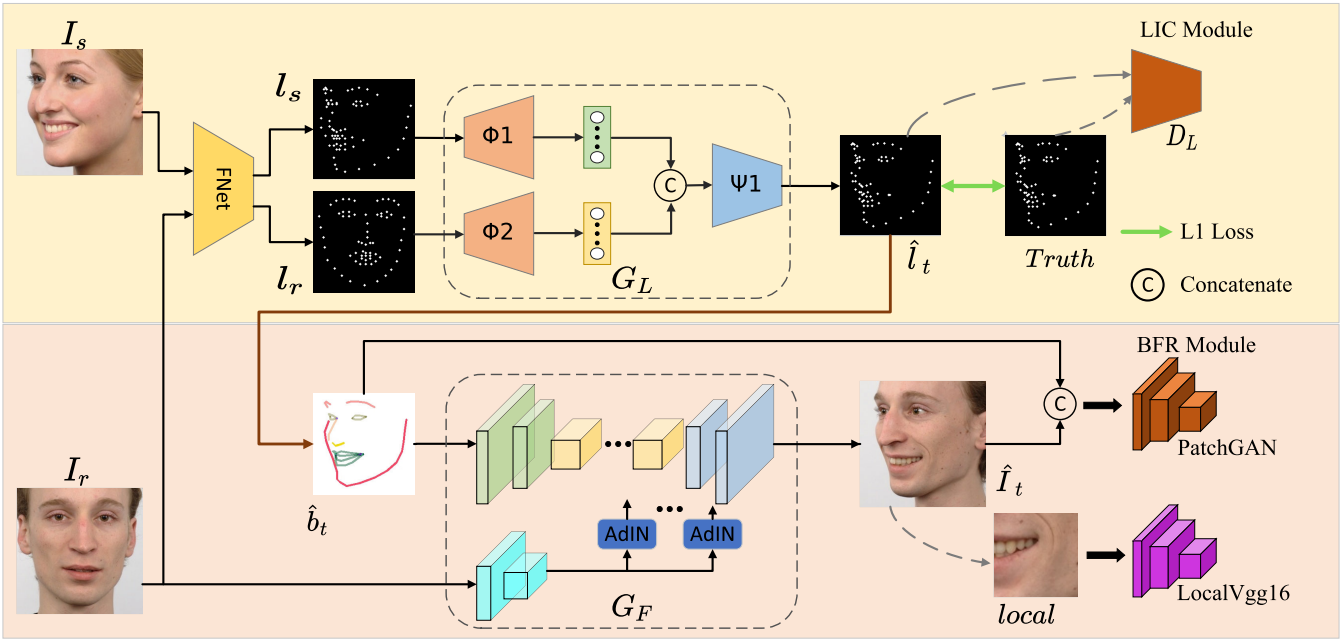

Fig. 2. Overview of the whole framework. The two modules of the framework, LIC and BFR, respectively, contain a landmark generator G_L and a face generator G_F. A face detector FNet would acquire both the source and reference landmarks, and then both landmarks would be sent to G_L to decouple the remaining source's identity on the target landmarks $\hat{l}_t$. After that, the target boundary image $\hat{b}_t$, which is generated from $\hat{l}_t$, is then joined with the reference image by G_F to produce the target image $\hat{I}_t$.

network decoupling in styleGAN [34], G_L consists of two encoders(ϕ_1, ϕ_2) and one decoder(ψ_1). ϕ_1 and ϕ_2 have the same structure, consisting of five fully-connected layers, each layer followed by a relu-leaky activation layer. The structure of ψ_1 is a mirror of the encoder. ϕ_1 is considered as an action encoder that decouples l_s into a one-dimensional action vector, and ϕ_2 is considered as an identity encoder that decouples l_r into a one-dimensional identity vector. Subsequently, the action vector and the identity vector are concatenated and decoded by ψ_1 to yield the target landmark $\hat{l}_t$. The structure of D_L is the same as that of ψ_1; the only difference is that ψ_1 outputs the landmark $\hat{l}_t$, while the former outputs a number between [0,1] for determining whether $\hat{l}_t$ is real or fake.

In order to obtain the optimal G_L, the loss function is designed as follows:

1) L1 Reconstruction Loss: We expect the closer the distance between the generated $\hat{l}_t$ and the ground truth l_t, the better. In actuality, we flatten the coordinate values of 68 landmarks into a vector before feeding them into the ϕ_1 and ϕ_2. Compared to the L1 loss, L2 will be more sensitive to anomalous coordinate values. So we select L1 reconstruction loss for improved robustness, as in equation (1).

$$L_1 = \|\hat{l}_t - l_t\|_1 \tag{1}$$

2) Adversarial Loss: Inspired by Zhang's et al. [19] approach to training, we introduce a discriminator D_L, treat G_L as a generator, and get the target output $\hat{l}_t$ after feeding l_s and l_r to G_L. Following that, $\hat{l}_t$ is fed to D_L to determine if its distribution is reasonable. the role of D_L is to facilitate G_L to generate $\hat{l}_t$ that conforms to the true landmark distribution space.

$$L_{D_L}^{adv} = E_{l_t \sim P_{data}(l_t)}[log(D_L(l_t))] + E_{\hat{l}_t \sim P_{data}(\hat{l}_t)}[1 - log(D_L(\hat{l}_t))] \tag{2}$$

where l_t is considered as the true target landmark estimate and $\hat{l}_t$ can be denoted as $\hat{l}_t = \psi_1(\phi_1(l_s), \phi_2(l_r))$.

3) Identity Consistency Loss: Because the identities of l_s and l_r can be arbitrary, when the identities of l_s and l_r are the same, there is a possibility that G_L only gets $\hat{l}_t$ from l_s through ψ_1, causing l_s and $\hat{l}_t$ to form a single shot. In order to prevent G_L learning without ϕ_2 participation, identity consistency loss (L_{idt}) is introduced. In particular, the L_{idt} is to ensure that G_L learns the action code only from ψ_1 and the identity code only from ψ_2.

$$L_{idt} = \|l_t - \psi_1(\phi_1(l_r), \phi_2(l_r))\|_1 \tag{3}$$

4) Triplet Loss: Landmarks with different identities but the same expression and posture are distributed close together in Euclidean space, while two landmarks with the same identity but different expressions and postures have a certain gap in their Euclidean distance [42]. The little distance between different identity classes and the high distance within the same identity class will lead to the difficulty of training G_L, and the generated results prone to pattern collapse.

With the addition of triple loss, as illustrated in Fig. 3, the objective loss function will penalize the distance between $\hat{l}_t$ and l_s in high-dimensional space while reducing the intra-class distance between $\hat{l}_t$ and l_t. With triple loss, $\hat{l}_t$ can better remove the identity information from l_s and carry only the identity on

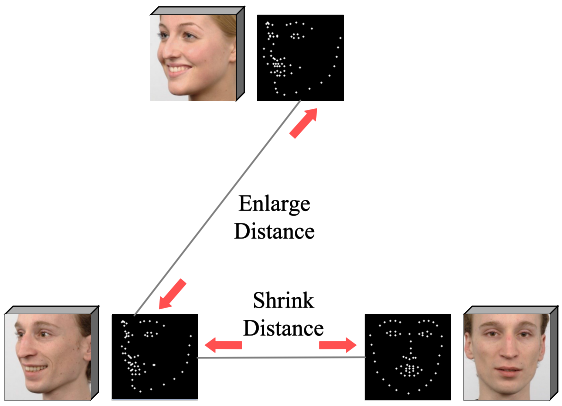

Fig. 3. The effect of triple loss is to enlarge the inter-class distance and shrink the intra-class distance.

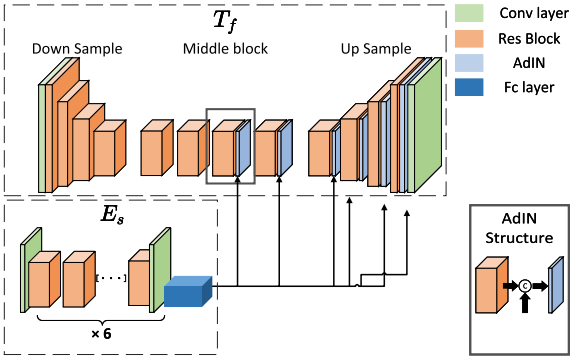

Fig. 4. The structure of the face generator G_F.

l_r, which helps the module solve the identity leakage problem.

$$L_{tri} = \max(dis(\hat{l}_t, l_t) - dis(\hat{l}_t, l_s) + margin, 0) \quad (4)$$

where margin is a constant that shrinks the distance between l_t and $\hat{l}_t$ and expands the distance between l_t and l_s.

Overall, the total loss function is expressed as:

$$L_{LIC} = \lambda_1 L_{D_L}^{adv} + \lambda_2 L_1 + \lambda_3 L_{idt} + \lambda_4 L_{tri} \quad (5)$$

where $\lambda_i (i = 1, 2, 3, 4)$ denotes the weight of the components of the loss function.

B. Boundary-Guided Face Reenactment Module

The face generator G_F uses StarGAN-2 as the backbone. As shown in Fig. 4, it consists of a boundary-to-face transformer T_f and a style extractor E_s. T_f is an hourglass-type network structure, with the downsampling block consisting of a convolutional layer and four decreasing-size residual blocks; the middle block consisting of four identical residual blocks; and the final upsampling block consisting of four increasing-size residual blocks followed by a deconvolutional layer. We add a tanh activation layer after the deconvolution layer as the image activation layer. The tanh layer constrains the output matrix threshold between $[-1, 1]$, which is very useful to prevent the G_F training gradient explosion and the color distortion of the output images. The network structure of the style extractor E_s is, in order, a convolutional layer, six residual blocks like the middle block of T_f, a convolutional layer, and finally a fully connected layer. The E_s style code output is then concatenated with the output features of each layer of T_f middle and up-sampling blocks. After that, the concatenated features will pass through an AdIN layer like Fig. 4 and flow to the next block.

Once we get the target landmarks, as mentioned above, the landmark coordinates are at a high distance within the identity intra-class and at a little distance inter-class. The solution is to extend the inter-class distance by increasing the identity information carried by $\hat{l}_t$. We divide the 68 landmark points of $\hat{l}_t$ into 13 parts (face contour, eyebrows, nose, mouth, etc.) according to semantics, and then connect the points of each part with a line. Wu et al. [29] argue that lines with distinct boundaries are not the real boundaries of a human face. Therefore, we apply Gaussian blur to each boundary line.

The $\hat{b}_t$ obtained from $\hat{l}_t$, is sent to G_F as a three-channel image together with I_r. T_f extracts multi-scale action features from $\hat{b}_t$ image, E_s extracts texture features that are the identity information of the reference face at all levels on I_r. The texture features are injected into the layers of the middle blocks and up-sampling blocks of T_f by adaptive instance normalization. Huang and Belongie [43] considers that the style of an image is determined by the mean and variance of the statistics of its feature map. Therefore, through removing the mean and variance of $\hat{b}_t$ feature map and injecting the statistics of I_r by multi-scale, the target face $\hat{I}_t$ generated by G_F has the identity of I_r face in all levels. That is to say, I_r provides multi-scale texture information for $\hat{I}_t$, which makes the identity information of $\hat{I}_t$ and I_r consistent. As for the discriminator, a 70×70 patchGAN [36] is used to penalize G_F. To ensure that $\hat{I}_t$ and I_s expressions and poses are as indistinguishable as possible, inspired by condition GAN [32], $\hat{b}_t$ is viewed as a condition and concatenated with output $\hat{I}_t$, then both of them are forwarded into the patchGAN. The results demonstrate the training paradigm of conditional GAN works well for head reenactment in a large pose.

The loss function of the face generation module consists of the following three components:

1) Conditional Adversarial Loss: Unlike typical general adversarial loss, G_F applies the conditional adversarial loss with the equation (6). We concatenate the generated image I_t and the boundary image $\hat{b}_t$ before feeding them into the patchGAN.

$$L_{D_B} = E_{I_t \sim P_{data}(I_t)}[log(D_B(I_t, b_t))] + E_{\hat{I}_t \sim P_{data}(\hat{I}_t)}[log(D_B(G_f(b_t), b_t))] \quad (6)$$

2) L1 Loss: We apply the L1 paradigm of the difference between the output $\hat{I}_t$ and the true value I_t like equation (7). Comparing with the L2 paradigm of both, L1 encourages less blurring [36].

$$L_1 = \|\hat{I}_t - I_t\|_1 \quad (7)$$

3) Local Perceptual Loss: L1 loss can reconstruct the color and luminous intensity of the global components of an image with good quality. However, it may not perform well in preserving complex details such as tooth gaps and hair lines on a face. This is because L1 loss tends to generate images

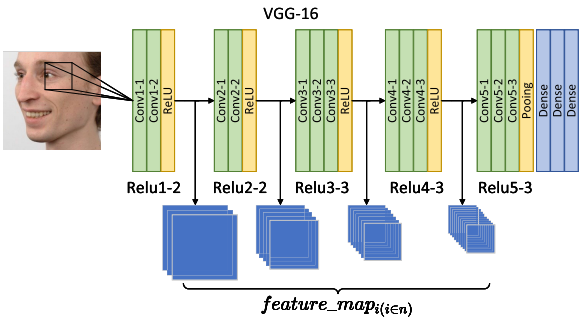

Fig. 5. Different Conv blocks of VGG-16.

with excellent global light and shadow quality while ignoring local details during the BFR module. The L1 distance is less concerned with outliers compared to the L2 distance, and thus, the high-frequency local components of the image are often ignored. To overcome this limitation, we propose a local perceptual loss.

Since the convolutional layer can only process information from adjacent pixels [44], [45], we introduce a pre-trained VGG-16 network and intercept its first few layers as a low-dimensional feature extractor. A completed VGG-16 consists of 13 convolutional layers followed by 3 fully connected layers, for a total of 16 layers. However, the image does not scale down after operating with each layer of convolution. We remove the last three fully connected layers and use the activation layer in the VGG-16 as a cutoff. As illustrated in Fig. 5, the 13-layer convolution can be divided into 5 blocks, denoted as Relu1-2, Relu2-2, Relu3-3, Relu4-3, and Relu5-3 (abbreviated as $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \mathcal{R}_4, \mathcal{R}_5$), respectively. If an image is input into the first layer of VGG-16, then the feature map output from the five cutoffs can be represented in order:

$$f_i = \mathcal{R}_{\rangle}(\mathcal{R}_{\rangle_{-\infty}}(\cdots \mathcal{R}_{\infty}(I_t)\cdots)) \tag{8}$$

$$\hat{f}_i = \mathcal{R}_{\rangle}(\mathcal{R}_{\rangle_{-\infty}}(\cdots \mathcal{R}_{\infty}(\hat{I}_t)\cdots)) \tag{9}$$

As in Fig. 5, f_i, $\hat{f}_i$ are represented as the feature maps of the output of the $\mathcal{R}_i$ block.

To focus on the high-frequency component, we calculate the L2 distance between the obtained feature maps. Each point on the feature map represents a local part of the image, and the relu3-3 block of the VGG net is finally selected as the output layer. The ablation experiments for the selection of the block of VGG net are illustrated in Section IV. The local perceptual loss is formulated as follows:

$$L_{per}^{local} = \frac{1}{k}\sum_{i=0}^{k} \| f_i - \hat{f}_i \|_2^k \tag{10}$$

Here k represents the feature map size.

Overall, the total loss function is expressed as:

$$L_{BFR} = \alpha_1 L_{D_F}^{adv} + \alpha_2 L_1 + \alpha_3 L_{per}^{local} \tag{11}$$

where $\alpha_i (i = 1, 2, 3)$ denotes the weight of the components of the whole loss function.

IV. Experiment

In this section, we first introduce the datasets and the implementation setting details elaborated in Section III. Metrics and comparative results with state-of-the-art methods will then be shown. Moreover, we also list the comparative results of various ablation experiments on HR-Net.

A. Datasets and Implementation Details

1) Implementation Detail: The LFG module is trained from scratch with the objective defined in equation (5) and the weights in it are settled as $\lambda_1 = 1$, $\lambda_2 = 100$, $\lambda_3 = 1$, and $\lambda_4 = 10$. Furthermore, the margin for triple loss in equation (4) is set to 1. The BFR module is trained from scratch with the objective defined in equation (11), and the weights in it are settled as $\alpha_1 = 1$, $\alpha_2 = 10$ and $\alpha_3 = 1$. Both modules of the HR-Net use Adam as the optimizer. The Adam optimizer parameters for G_L and D_L are jointly set to $\beta_1 = 0.99$, $\beta_2 = 0.999$; for G_F and D_F, the Adam optimizer parameters are jointly set to $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The LIC module is trained first and would be selected as a well-performing landmark generator G_L for the training of the other one. We train the LIC module for 2000 epochs with batch sizes of 128; the initial learning rate is 0.0002 and decays by forty percent every 600 epochs. The BFR module is trained for 300 epochs with batch sizes of 16 and an initial learning rate of 0.0002. The latter's learning rate remains constant at its initial value for the first 150 epochs and then decreases linearly to zero. All experiments are run on Ubuntu 18.04 with an NVIDIA RTX 3090 GPU.

2) RaFD: Dataset contains a total of 8,040 images from 67 identities, each identity with 8 expressions and 5 angles [46]. The identities with the same angle and expression contain three different gazes. We follow the same settings as in the FReeNet [19]. Namely, we use the 45°, 90°, and 135° angles of the dataset and crop the images. We align and resize the images to 256*256. To test the performance, 100 images are selected as source images for each identity to reenact, resulting in 6,700 generated images.

3) VoxCeleb-1: Dataset is a contains 1,251 celebrity video URLs collected from YouTube [47]. The dataset has been divided into a training set and a test set, and we follow the experimental protocol reported in the Few-shot [48]. One image per frame is taken from each video, and all images are cropped and scaled to 256 × 256. Moreover, all images are face-aligned. For the testing performance phase, 50 celebrities are randomly selected from the test set. Each of the test set's 32 hold-out frames has no overlap with the training images.

B. Compared to SOTA

1) Method: In this study, we comprehensively evaluated our proposed method by comparing it against five state-of-the-art methods on both the RaFD [46] and VoxCeleb-1 [47] datasets. The methods we compared our approach against were FReeNet [19], DualGAN [21], FOMM [3], Pix2Pix-HD [37], and X2Face [49] for the RaFD dataset, and Few-Shot [48], FIPLS [50], FOMM, Pix2Pix-HD, and X2Face for the VoxCeleb-1 dataset.

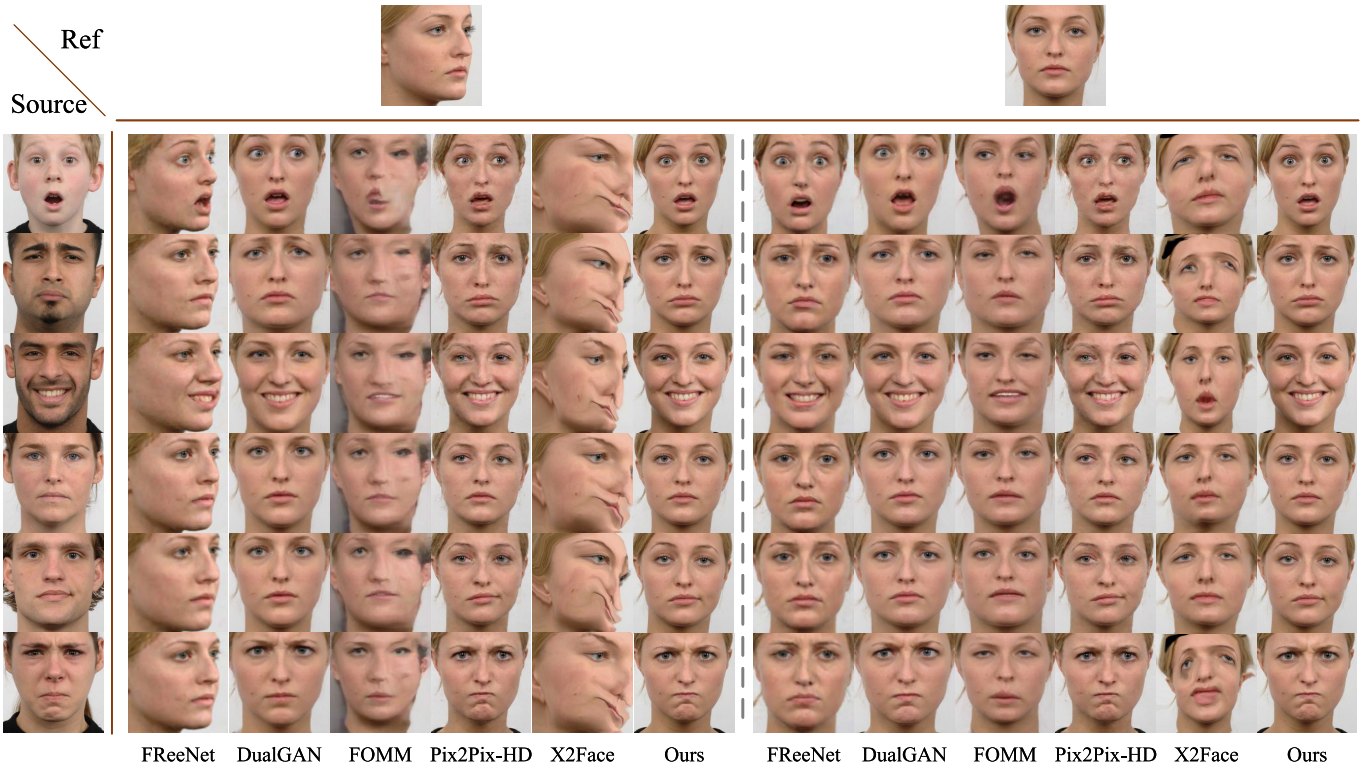

Fig. 6. Qualitative results on RaFD. The top row of two faces are reference images, and the leftmost column of six faces are source images.

FReeNet and X2Face fine-tuned their pre-trained networks on both datasets, and we evaluated their performance thereafter. Additionally, we also tested X2Face on both datasets using the original pre-trained network provided by the authors.

To the best of our knowledge, DualGAN, FewShot, and FIPLS do not have code available for reproducible execution, so we only copy the results and illustrations of their articles. We also trained Pix2Pix-HD using the same hyperparameters as reported by the original authors. To ensure a fair comparison, we used landmark-enhanced boundary maps as three-channel images stacked with the reference faces as six-channel tensors for training.

2) Metrics: In order to evaluate the quality of the generated images, we introduce FID [51], the specific method is to extract the features of the generated images using the trained Inception-V3 network and calculate the distance between the generated images and the real images in the feature space. SSIM [52](Structural Similarity) is introduced to compare the different structures between truth and target images at a low-level, especially the variance of the face details. We also make use of a state-of-the-art face recognition network [53] to get the identity features of a generation and calculate the CSIM (Cosine Similarity) with its ground truth. The objective of CSIM is to evaluate the identity match of the reenacted images.

3) Compared Result: Table I and Fig. 6 present the qualitative and quantitative results of our method compared to state-of-the-art approaches on the RaFD dataset. As shown in Fig. 6, FReeNet method can transfer the expression of the source face to the reference face; however, it cannot change the pose direction of the reference face. When the reference face pose direction is frontal, FOMM is also capable of generating clearer faces, but the resulting expressions are not perfectly reproduced. Conversely, when the reference face pose direction is sideways, FOMM fails to generate faces. The results of DualGAN in Fig. 6 demonstrate that it can achieve the transfer of the source face pose. However, it does not sufficiently address the issue of identity leakage. For example, the generated face contour in the second row is visibly too large and resembles more the male identity of the source.

TABLE I
QUANTITATIVE RESULTS ON THE RaFD

Method	FID↓	SSIM↑	CSIM↑
FReeNet [19]	12.17	0.717	0.812
DualGAN [21]	4.79	0.726	0.862
FOMM [3]	9.37	0.723	0.801
Pix2Pix-HD [37]	15.557	0.701	0.779
X2Face [49]	128.43	0.61	0.68
Ours	4.633	0.882	0.906

Regarding Pix2Pix-HD, a GAN-based method, as shown in Fig. 6, the pixels at the edges of the face generated by this method are always blurred or even drifted to the background. This phenomenon can be easily discerned by human eyes or machine algorithms as a m̂odified fakeîmage. Additionally, as seen in Table I, the metrics of X2Face on the RaFD dataset are significantly higher than all other methods, including our own. From Fig. 6, we can find that when the reference image is a side face, the generated result of X2Face is difficult to identify the human shape; when the reference image is a front face, the generated face is also greatly distorted. The possible

Fig. 7. Qualitative results on VoxCeleb-1. The first column Ref is the manipulated face, and the second column Truth is the desired expression and pose of the manipulated face.

reason for this issue could be that the space learned by the X2Face pre-trained network does not match the distribution of the RaFD dataset.

Compared to the above methods, our method can transfer the expression and pose of the leftmost column source face quite well, no matter the gestures of the reference faces. Furthermore, the image results in Fig. 6 illustrate HR-Net addresses the identity leakage problem more perfectly compared to DualGAN. For the quantitative analysis, HR-Net achieves a FID of 4.633 on the RaFD dataset, which exceeds all three states of the arts. Typically, the FID only indicates that the generations are more like the "real" images and does not reflect whether the generated results have identity leakage problems. That's why the FID of DualGAN is only 4.79, which is close to our FID value, but the face contours of the same identities generated by different sources are not the same. It is noticeable that our method pays attention to both the L1 parametric and local L2 parametric constraints. That is, both the SSIM metric and the CSIM metric of our method shown in Table I are significantly higher than any other. As presented in Table I, the SSIM result of our method is able to reach 0.882, and the CSIM reaches an amazing 0.906.

The qualitative and quantitative results of our method are compared to state-of-the-art methods, including Few-shot, on the VoxCeleb-1 dataset, as shown in Fig. 7 and Table II. We note that Few-shot is a few-shot learning model and chose their one-shot inference results as a comparison. Our method achieves outstanding results on the VoxCeleb-1 dataset. As illustrated in Fig. 7, our method can reconstruct the microphone in the background, which is a clear advantage and contribution of our BRF module, enabling the learning of all foreground-background information from the reference image simultaneously. In contrast, none of the other compared methods, including Pix2Pix-HD, can generate the microphone in the second row of images. Pix2Pix-HD generates a good foreground of the face but does not learn the color information of the reference background, leading to significant weakening of the image metrics. Our method also preserves the background shape of the reference image well compared to the wrapped methods, including FOMM and X2Face.

TABLE II
QUANTITATIVE RESULTS ON THE VOXCELEB-1

Method	FID↓	SSIM↑	CSIM↑
Few-shot [48]	43.0	0.670	0.15
FIPLS [50]	43.9	0.79	0.18
FOMM [3]	25.0	0.723	0.813
Pix2Pix-HD [37]	42.7	0.56	0.09
X2Fac [49]	45.8	0.68	0.16
Ours	10.34	0.843	0.881

Table II demonstrates that our method achieves the best results for each metric on the VoxCeleb-1 dataset. However, the background complexity of the VoxCeleb-1 dataset is higher than that of the laboratory dataset RaFD, which explains why all metrics of our method in Table II are inferior to Table I.

To summarize, our method achieves state-of-the-art performance on the VoxCeleb-1 dataset, as demonstrated by both qualitative and quantitative results. The BRF module is a clear advantage, enabling our method to learn all foreground-background.

C. Ablation Experiments

The experimental setup for our study includes several components. Firstly, we conduct an ablation study on the LIC module to determine its impact on the overall performance of our proposed HR-Net model. Secondly, We also explore the impact of different components of the L_{LIC} loss function on various aspects of the HR-Net model. Thirdly, we investigate the need for landmark enhancement measures to further improve the quality of generation. Finally, we discuss the optimal VGG-16 perceptual block for our model. In addition to the three metrics presented in the preceding section, we also employ PSNR [54] and LPIPS [55] for quantitative analysis,

Fig. 8. Qualitative results for the LIC module in the VoxCeleb-1 ablation experiment. The first column of the figure shows the reference face, and the second column shows the source face providing expressions and gestures. The last two columns demonstrate the role of the LIC module in addressing the "identity leakage" problem during cross-reenactment.

TABLE III
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON LIC MODULES

	FID↓	SSIM↑	CSIM↑	PSNR↑	LPIPS↓
without LIC module	23.46	0.712	0.763	19.991	0.1148
with LIC module	10.34	0.843	0.881	34.409	0.0218

with PSNR being an essential metric in the field of image hyper-segmentation, and LPIPS assessing the similarity of two images in a manner more congruent with human perception.

1) Ablation on LIC Module: Our experiments demonstrate the significance of the LIC module, as evidenced by both qualitative and quantitative analysis on the VoxCeleb-1 dataset. Fig. 8 highlights the impact of the LIC module on the generated images, where the contours of the face bear traces of the source face when the module is not involved. When the source image in the first line drives the reference front face with an open-mouthed male into a side face, the generated male face without the LIC module shows a smaller contour and bears female identity features of the source picture. In contrast, the generated face with the LIC module does not exhibit an unnatural contour, indicating that the involvement of the LIC module successfully addresses the identity leakage problem.

Furthermore, our findings demonstrate that the LIC module improves the performance of all five metrics of the generated images, as shown in Table III. By resolving the identity leakage problem, the generated faces are more consistent with the true values, resulting in improved metric scores for the pixel level and feature semantic level of the images.

2) Ablation on L_{LIC}: We conducted validation of each component of the L_{LIC} loss function on the RaFD dataset and evaluated their impact on the training process and final

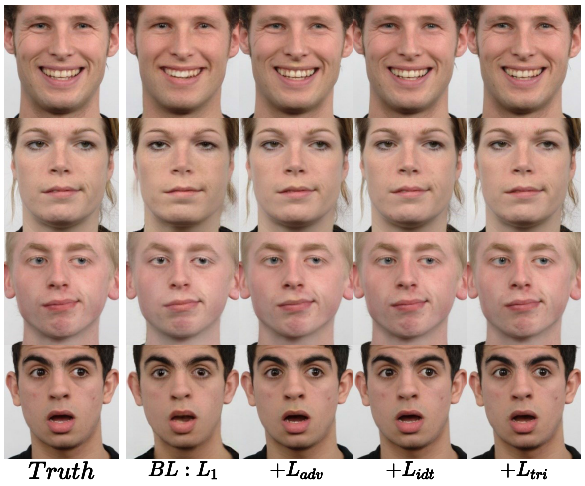

Fig. 9. Loss the effect of different components on the quality of face generation in the generation phase. baseline (BL) is set to L1 only, from left to right representing the Loss function adding the last component in turn.

TABLE IV
THE RESULTS OF THE METRICS OF EACH COMPONENT OF L_{LIC} ARE ADDED IN TURN. BASELINE (BL) REFERS TO THE MODEL WITH L1 LOSS ONLY. ACE REPRESENTS THE AVERAGE COORDINATE ERROR OF THE LANDMARKS AND THE OTHER METRICS INDICATE THE QUALITY OF THE GENERATED FACES

	ACE	FID↓	SSIM↑	CSIM↑	PSNR↑	LPIPS↓
BL:L_1	3.943	10.028	0.821	0.856	33.402	0.0548
$+L_{adv}$	3.713	7.319	0.839	0.861	32.015	0.0371
$+L_{idt}$	3.677	6.707	0.847	0.869	32.811	0.0354
$+L_{tri}$	1.742	4.633	0.882	0.906	33.919	0.0167

quality of the generated faces, as presented in Table IV and Fig. 9. In addition to the five metrics that measure the image quality, we introduced the Average Coordinate Error (ACE) to determine the difference between the modified $\hat{l}_t$ and the ground truth l_t. We computed ACE by generating 1000 random $\hat{l}_t$, calculating them with the corresponding l_t, and averaging the results.

We used L1 loss as the baseline for $Loss_{LIC}$ and then added L_{adv}, L_{idt}, and L_{tri} in turn. The notation $+L_*$ in Fig. 9 and Table IV indicates that the corresponding loss was added to the target loss, and the notation $+L_{tri}$ indicates that the loss at this point is consistent with equation (5).

Table IV shows that the baseline has the highest ACE, indicating that $\hat{l}t$ still contains some of the source contour information. Moreover, the LIC module obtained from the baseline loss training has the worst performance according to the five quantitative metrics in the final face generation test phase. As $Ladv$, L_{idt}, and other losses are introduced during training, the relative ACE starts to decrease. Notably, the introduction of L_{tri} results in a significant decrease in the ACE of $\hat{l}t$ to 1.742. While it may be difficult for the human eye to perceive the improvement in image quality resulting from the addition of $LossLIC$ components, Table IV shows that the addition of these components is indeed beneficial.

3) Ablation on Landmark Enhance: To highlight the superior performance of boundary images over landmark images, we use both landmark and boundary images as inputs to

Fig. 10. Landmark-based and boundary-based reconstructed images. The first and third rows are the target landmarks and target boundaries after the LIC module, and the second and fourth rows are the respective reconstructed generated images.

TABLE V
QUANTITATIVE RESULTS BASED ON LANDMARKS AND BOUNDARIES

	RaFD				
	FID↓	SSIM↑	CSIM↑	PSNR↑	LPIPS↓
$\hat{l}_t$ input	8.152	0.801	0.822	20.047	0.1383
$\hat{b}_t$ input	4.633	0.882	0.906	33.919	0.0167
	VoxCeleb-1				
	FID↓	SSIM↑	CSIM↑	PSNR↑	LPIPS↓
$\hat{l}_t$ input	14.773	0.813	0.848	28.944	0.0636
$\hat{b}_t$ input	10.34	0.843	0.881	34.409	0.0218

TABLE VI
QUANTITATIVE RESULTS OF VARIOUS BLOCKS OF VGG-16

Loss	FID↓	SSIM↑	CSIM↑	PSNR↑	LPIPS↓
without L_{per}^{local}	14.64	0.802	0.832	30.121	0.0581
L_{per}^{1-2}	12.79	0.814	0.845	33.377	0.0473
L_{per}^{2-2}	8.701	0.837	0.869	33.582	0.0408
L_{per}^{3-3}	4.633	0.882	0.906	33.919	0.0167
L_{per}^{4-3}	6.545	0.851	0.874	31.419	0.0311

G_F, and present the qualitative and quantitative results in Fig. 10 and Table V, respectively. According to the flow of the BFR module shown in Fig. 2, generating $\hat{I}_t$ is a matter of reconstructing the face from the boundary domain back into the image domain while keeping the face's actions in the boundary domain. We feed both $\hat{b}_t$ and $\hat{I}_t$ into the patchGAN with the goal of constraining $\hat{I}_t$'s actions to be consistent with those of $\hat{b}_t$. As a result, we assume that the greater the amount of action information carried in the input domain, the better. In Fig. 10, we can find that the reconstructed image in the fourth row and seventh column is more consistent with the latent space's pose than the reconstructed image in the second row and seventh column. And this phenomenon indicates that the ability to constrain action information from the boundary latent space is stronger than the landmark latent space.

We calculate the quantitative results of landmark based and boundary based in the same way as the paragraph of the compared result. The results in Table V also reflect that the boundary-based generation of faces performs better on the five metrics mentioned. The generator G_F with $\hat{l}_t$ as input, keeping the hyper-parameter settings consistent with those in the implementation details, and the generated results have a FID of 8.152 in the RaFD dataset and a FID of 14.773 in the VoxCeleb-1 dataset. The reasons why the metric performance still exceeds some previous works are that we use a better network framework, conditional GAN training, and a well-designed loss function.

4) Optimal Perceptual Block: Additionally, we conduct experiments with different blocks of VGG-16 to determine the optimal local perceptual feature extractor for our model, and report the results in Table VI and Fig. 11. We denote the baseline as the L_{BFR} without local perceptual loss. We only experiment with the output of the first four blocks separately as the perceptual matrix to participate in the calculation of L_{per}^{local} because the purpose of VGG-16 is to be implemented as a chunk-aware low-level extractor, i.e., a local convolution operator. These various blocks as L_{per}^{local} can be denoted as L_{per}^{1-2}, L_{per}^{2-2}, L_{per}^{3-3} and L_{per}^{4-3}. Our goal is to discover the optimal block of the network to improve the details.

When there is no local perceptual loss involved in the L_{BFR}, as shown in Fig. 11, the generated teeth are very blurred, and even the color of the teeth adheres to the lower lip. In the generated images of the second, third, and fifth rows, their

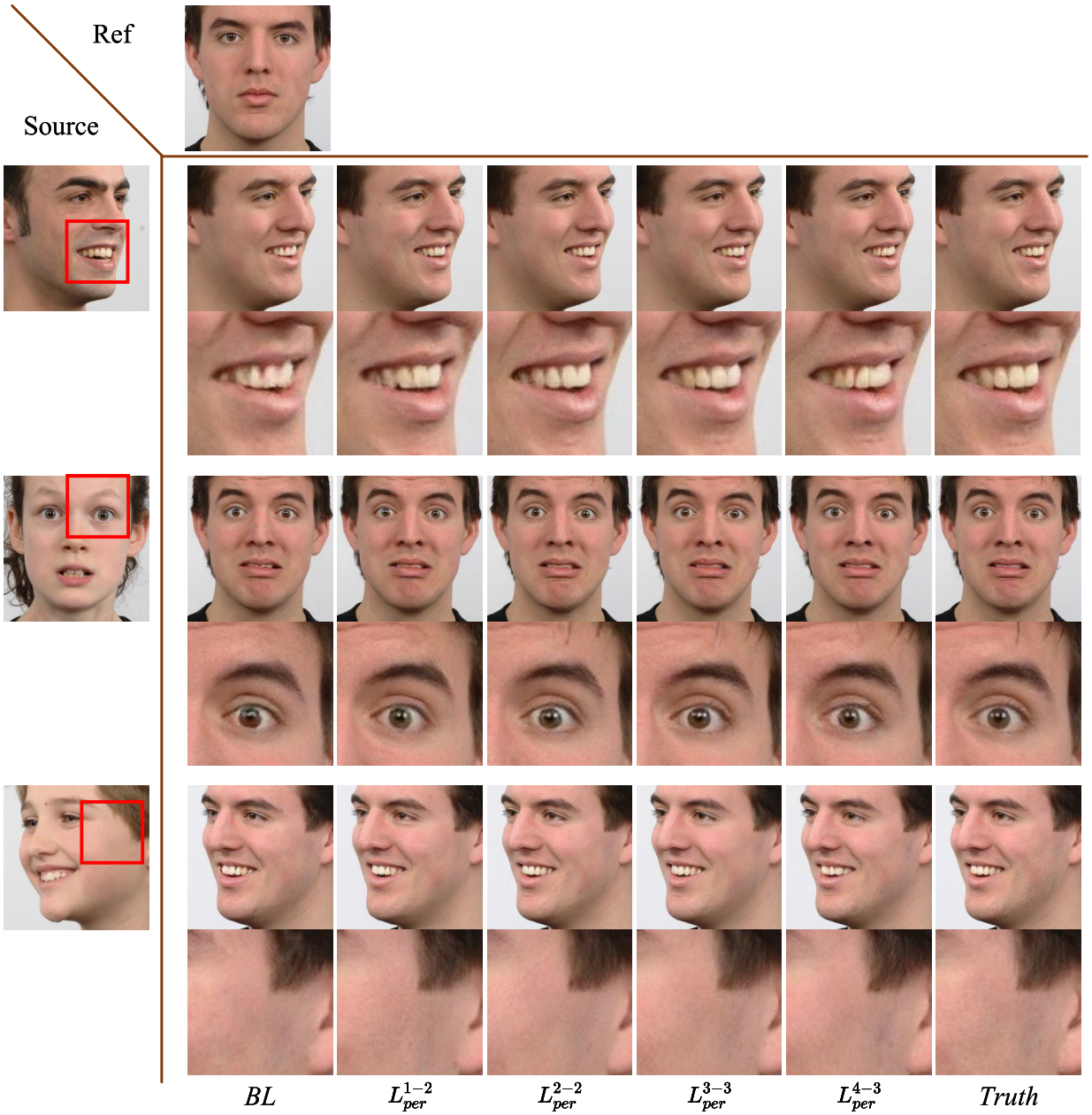

Fig. 11. In the second column from the left, the baseline (BL) denotes the L_{BFR} without local perceptual loss. Others are the quality outcomes of various L_{per}^{n}. The red box circles the places where face generation details are most likely to be blurred, The second and fourth columns illustrate the local generation results and their groundtruths.

teeth gaps are not clear, while the resulting images of L_{per}^{3-3} show clear gaps between each tooth, and the shape of the teeth resembles that of ground truth. As for the fourth column of images, when there is no local perceptual loss, the generation of wrinkles and hairlines fails. For other blocks involved, the generated hair is either too tiny in size or appears blurred, but overall, both the tooth and hair generation details demonstrate that Relu3-3 is the optimal block.

Notably, as seen in Table VI, without L_{per}, the FID and LPIS are 14.64 and 0.0581, however, after L_{per}^{3-3} is added, the FID and LPIS reduce to 4.63 and 0.0167, respectively. The introduction of local perceptual loss causes FID and LPIPS to decrease by an order of magnitude compared to what they would have been in the absence of L_{per}. This is because of how the two measures are measured; both FID and LPIPS compare the differences in feature space between the generated images and the ground truth using pre-trained recognition networks. Overall, both FID and LPIPS prefer to evaluate an image's "goodness" from a human perspective. FID and LPIPS provide advanced evidence that HR-Net is state-of-the-art.

V. Conclusion

We propose a novel method that can transfer one facial expression and pose to arbitrary identities. First, a light landmark identity converter is introduced to address the identity leakage. We combine both the source landmark and the reference landmark into the converter, and subsequently, we obtain the modified target landmark with the source action and the

reference identity from it. A triple loss is introduced to assist the converter train and prevent landmark pattern collapse. Second, a boundary-guided face generator is introduced after enhancing the target landmarks to a target boundary image. It is capable of learning action representations on the boundary images and identity representations on the reference images. We combine these two different representations into the face generator to generate more incredible facial images. Furthermore, a local perceptual loss function is implemented to facilitate the generator in generating results with increasing detail quality. Extensive ablation experiments and comparisons with state-of-the-arts on two different datasets demonstrate our method performs excellently. However, many previous landmark-based works, including our method, have trained the individual modules separately. Our future research will focus on enabling an overall training model framework. At the same time, we work on generating face images with higher resolution.

REFERENCES

[1] J.-T. Chien and S.-J. Huang, "Learning flow-based disentanglement," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 15, 2022, doi: 10.1109/TNNLS.2022.3190068.

[2] A. Siarohin, S. Lathuiliere, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2377–2386.

[3] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–15.

[4] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 1–16.

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

[6] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-GAN: Unsupervised video retargeting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 119–135.

[7] T.-C. Wang et al., "Video-to-video synthesis," 2018, *arXiv:1808.06601*.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[9] S. Bounareli, V. Argyriou, and G. Tzimiropoulos, "Finding directions in GAN's latent space for neural face reenactment," 2022, *arXiv:2202.00046*.

[10] P.-H. Huang, F.-E. Yang, and Y.-C.-F. Wang, "Learning identity-invariant motion representations for cross-ID face reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7084–7092.

[11] Y. Liu, Q. Li, Q. Deng, and Z. Sun, "Towards spatially disentangled manipulation of face images with pre-trained StyleGANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1725–1739, Apr. 2023.

[12] S. Tripathy, J. Kannala, and E. Rahtu, "ICface: Interpretable and controllable face reenactment using GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3385–3394.

[13] S. Tripathy, J. Kannala, and E. Rahtu, "FACEGAN: Facial attribute controllable rEenactment GAN," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1329–1338.

[14] F.-T. Hong, L. Zhang, L. Shen, and D. Xu, "Depth-aware generative adversarial network for talking head video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3397–3406.

[15] A. Tang, H. Xue, J. Ling, R. Xie, and L. Sang, "Dense 3D coordinate code prior guidance for high-fidelity face swapping and face reenactment," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Dec. 2021, pp. 1–8.

[16] B. Peng, H. Fan, W. Wang, J. Dong, and S. Lyu, "A unified framework for high fidelity face swap and expression reenactment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3673–3684, Jun. 2022.

[17] X. Tu et al., "Image-to-video generation via 3D facial dynamics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1805–1819, May 2021.

[18] C. Yang, S.-Y. Yao, Z.-W. Zhou, B. Ji, G.-T. Zhai, and W. Shen, "Poxture: Human posture imitation using neural texture," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8537–8549, Dec. 2022.

[19] J. Zhang et al., "FReeNet: Multi-identity face reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5326–5335.

[20] J. Liu et al., "Li-Net: Large-pose identity-preserving face reenactment network," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[21] G.-S. Hsu, C.-H. Tsai, and H.-Y. Wu, "Dual-generator face reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 642–650.

[22] X. Fu, X. Wang, J. Liu, W. Liu, J. Dai, and J. Han, "MakeItSmile: Detail-enhanced smiling face reenactment," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[23] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "VisemeNet: Audio-driven animator-centric speech animation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–10, Aug. 2018.

[24] K. F. MacDorman, R. D. Green, C.-C. Ho, and C. T. Koch, "Too real for comfort? Uncanny responses to computer generated faces," *Comput. Hum. Behav.*, vol. 25, no. 3, pp. 695–710, May 2009.

[25] L. Yu, J. Yu, M. Li, and Q. Ling, "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.

[26] P. Sun, Y. Li, H. Qi, and S. Lyu, "LandmarkGAN: Synthesizing faces from landmarks," *Pattern Recognit. Lett.*, vol. 161, pp. 90–98, Sep. 2022.

[27] S. Liang, Z.-Z. Zhou, Y.-D. Guo, X. Gao, J.-Y. Zhang, and H.-J. Bao, "Facial landmark disentangled network with variational autoencoder," *Appl. Math.-A, J. Chin. Universities*, vol. 37, no. 2, pp. 290–305, Jun. 2022.

[28] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "MarioNETte: Few-shot face reenactment preserving identity of unseen targets," in *Proc. AAAI*, vol. 34, 2020, pp. 10893–10900.

[29] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, "ReenactGAN: Learning to reenact faces via boundary transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 603–619.

[30] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High-fidelity face manipulation with extreme poses and expressions," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2218–2231, 2021.

[31] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[32] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[33] 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, pp. 4401–4410.

[34] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

[35] Z. Chen, Y. Wang, T. Guan, L. Xu, and W. Liu, "Transformer-based 3D face reconstruction with end-to-end shape-preserved domain transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8383–8393, Dec. 2022.

[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

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

[38] 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.

[39] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "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.

[40] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN V2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8188–8197.

[41] 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.
[42] H. Chen, Y. Lin, B. Li, and S. Tan, "Learning features of intra-consistency and inter-diversity: Keys toward generalizable deepfake detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1468–1480, Mar. 2023.
[43] 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.
[44] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2018, pp. 7794–7803.
[45] H. Zhang et al., "Self-attention generative adversarial network," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.
[46] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the Radboud faces database," *Cognit. Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.
[47] A. Nagrani, J. Son Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," 2017, *arXiv:1706.08612*.
[48] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9459–9468.
[49] O. Wiles, A. Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 670–686.
[50] H. Zhang, Y. Ben, W. Zhang, T. Chen, G. Yu, and B. Fu, "Fine-grained identity preserving landmark synthesis for face reenactment," 2021, *arXiv:2110.04708*.
[51] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–4.
[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[53] 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.
[54] A. Z. Shaikh and M. Ghanbari, "Comparison of predictive coding methods for image transmission," *IEEE Trans. Commun.*, vol. COM-22, no. 8, pp. 1046–1055, Apr. 1974.
[55] 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.

Qiuyu Ren received the B.S. degree from Tianjin University, China, in 2021, where he is currently pursuing the master's degree with the School of Electrical and Information Engineering. His current research interests include face recognition and image translation.

Zhiying Lu received the B.S. and M.S. degrees in industrial automation and the Ph.D. degree in electrical engineering from Tianjin University, Tianjin, China, in 1987, 1992, and 2010, respectively. She is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. She has more than 30 years of academic research and presided more than a number of national and provincial projects. Her research interests include image processing, biometrics recognition, and power systems.

Haopeng Wu received the M.S. degree from the Inner Mongolia University of Technology, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, China. His research interests include face recognition and facial expression recognition.

Jianfeng Zhang received the B.S. and M.S. degrees in automation from the Nanjing University of Science and Technology, Nanjing, China, in 2007 and 2010, respectively, and the Ph.D. degree from Tianjin University, Tianjin, China, in 2020. From 2010 to 2015, he was committed to the research of information security with the Shandong Academy of Sciences. He is currently an Associate Professor with Shandong Normal University, Jinan, China. His research interests include biometrics recognition and medical image.

Zijian Dong received the B.S. degree from the Hebei University of Technology in 2022. He is currently pursuing the master's degree with Tianjin University. His research interests include computer vision and micro-expression recognition.