JR2Net: Joint Monocular 3D Face Reconstruction and Reenactment

Jiaxiang Shang [1*], Yu Zeng [2*], Xin Qiao [3], Xin Wang [3], Runze Zhang [3], Guangyuan Sun [3], Vishal Patel [2], Hongbo Fu [4]

[1]Hong Kong University of Science and Technology,
[2]Johns Hopkins University,
[3]Tencent,
[4]City University of Hong Kong
jshang@cse.ust.hk, zengxianyu18@qq.com, {seanqiao,alexinwang,ryanrzzhang,gerrysun}@tencent.com,
vpatel36@jhu.edu, hongbofu@citvu.edu.hk

Abstract

Face reenactment and reconstruction benefit various applications in self-media, VR, etc. Recent face reenactment methods use 2D facial landmarks to implicitly retarget facial expressions and poses from driving videos to source images, while they suffer from pose and expression preservation issues for cross-identity scenarios, i.e., when the source and the driving subjects are different. Current self-supervised face reconstruction methods also demonstrate impressive results. However, these methods do not handle large expressions well, since their training data lacks samples of large expressions, and 2D facial attributes are inaccurate on such samples. To mitigate the above problems, we propose to explore the inner connection between the two tasks, i.e., using face reconstruction to provide sufficient 3D information for reenactment, and synthesizing videos paired with captured face model parameters through face reenactment to enhance the expression module of face reconstruction. In particular, we propose a novel cascade framework named JR2Net for **J**oint Face **R**econstruction and **R**eenactment, which begins with the training of a coarse reconstruction network, followed by a 3D-aware face reenactment network based on the coarse reconstruction results. In the end, we train an expression tracking network based on our synthesized videos composed by image-face model parameter pairs. Such an expression tracking network can further enhance the coarse face reconstruction. Extensive experiments show that our JR2Net outperforms the state-of-the-art methods on several face reconstruction and reenactment benchmarks.

Introduction

Face reenactment is the task of transferring the expressions and poses from driving video frames to a source image. Recent face reenactment methods usually adopt a deep generative approach (*e.g.*, Conditional Generative Adversarial Networks (CGANs)), which is driven by either 2D facial landmarks (Siarohin et al. 2019; Zhang et al. 2020; Ha et al. 2020; Zakharov et al. 2019, 2020; Burkov et al. 2020; Wu et al. 2018; Pumarola et al. 2018) or 3D face models (Kim et al. 2018; Yao et al. 2020; Thies et al. 2016). Such methods leverage facial landmark motion, rendered 3D Morphable Models (Paysan et al. 2009; Cao et al. 2013; Li et al. 2017; Ploumpis et al. 2020; Tewari et al. 2021; Mo et al. 2022; Zeng et al. 2020; Ha et al. 2020) (3DMMs) or optical flow as guidance for their generation modules. However, such representations are insufficient to preserve the expression, pose, and illumination of driving frames since they do not formulate the expression and pose individually, as illustrated in Figure 1 (see the results by FOM (Siarohin et al. 2019)). Besides, the entanglement of face shape, expression, and pose also limit the accuracy of face reenactment.

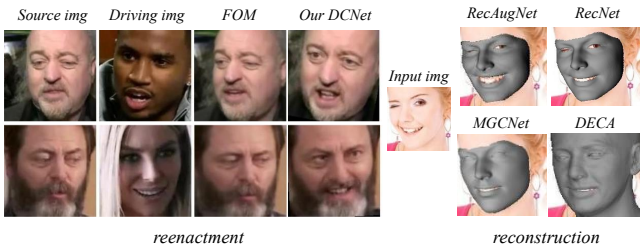

Figure 1: Left: Our reenactment pipeline (DCNet) achieves better pose and expression on the cross-identity setting than FOM (Siarohin et al. 2019). Right: Our final face reconstruction procedure (RecAugNet) reconstructs the facial expression in an input image more faithfully than our coarse method (RecNet), MGCNet (Shang et al. 2020) and DECA (Feng et al. 2021).

During the construction of face reenactment, face reconstruction methods are employed to recover 3D facial geometry and appearance from a monocular image (Feng et al. 2021; Shang et al. 2020; Deng et al. 2019; Tewari et al. 2019, 2018; Tran and Liu 2018; Tewari et al. 2017) or video frames (Wu et al. 2019; Tewari et al. 2019). In this task, 3DMMs are commonly used as a solid face prior and a linear expression basis parameterized by shape and expression coefficients. To benefit from the 3DMMs, the above self-supervised face reconstruction methods have been proposed to regress 3DMM coefficients by minimizing 2D-based losses, *e.g.*, landmark and rendering losses. However, such 2D-based losses suffer from depth ambiguity and the entangling of face shape and expression. Large expression

*These authors contributed equally.

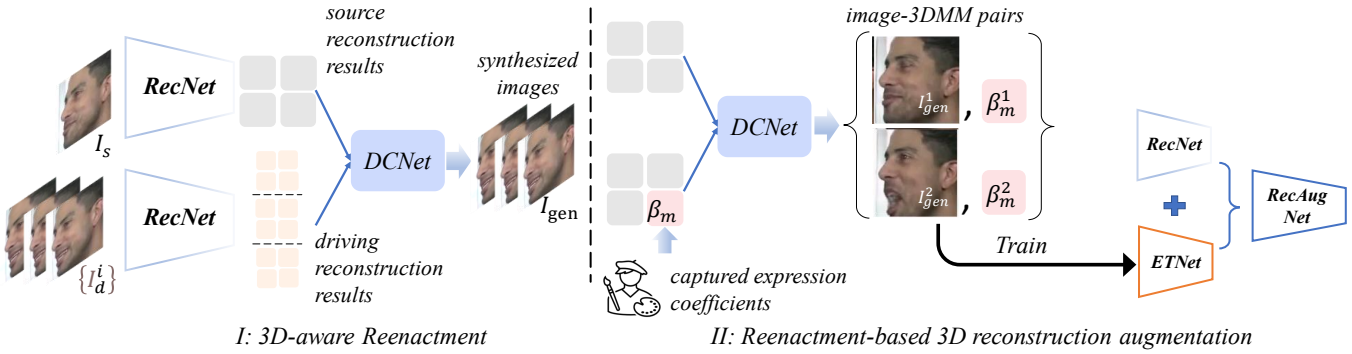

Figure 2: The pipeline of our method. We first reconstruct 3DMM shape and expression coefficients, poses, and SH coefficients as 3D results from a source image I_s and a driving video $\{I_d^i\}$ by our RecNet. Such results are used to train our DCNet. Furthermore, we use the expression coefficients β_m from animators as input to DCNet. Then, we synthesize image-expression coefficient pairs $\{I_{gen}^i,\ \beta_m^i\}$, and use them to train ETNet. Finally, we utilize the trained ETNet to enhance our RecNet to get RecAugNet.

cases are significantly affected (see the results by RecNet, MGCNet (Shang et al. 2020), and DECA (Feng et al. 2021) in Figure 1), since the detected facial landmarks are usually inaccurate on such cases and the proportion of large expression samples in the training dataset is small.

To alleviate these problems, we explore the inner connection between face reconstruction and reenactment to tackle both tasks together in a unified system. Hence, we propose a novel cascade framework named JR2Net for **J**oint Face **R**econstruction and **R**eenactment, which uses face reconstruction to provide sufficient 3D information for reenactment, and synthesizes videos paired with captured face model parameters through face reenactment to enhance the expression module of face reconstruction.

Firstly, we train a monocular face **Rec**onstruction Network called RecNet as a coarse reconstruction network following (Deng et al. 2019; Shang et al. 2020) (Figure 2). To provide more robust guidance/conditions for face reenactment, we reconstruct 3DMM coefficients, face poses and Spherical Harmonics (SH) coefficients for the source image and each frame of a driving video. Different from previous methods, we recombine and render these reconstruction results to normal maps, light maps, masks, and image features, since such dense pixel-level representations about face shape, texture, pose, and illumination provide more detailed and specific conditions for the face reenactment. In particular, we propose a view synthesis module to warp images/image features from source image space to driving frame space, making such pixel-level conditions aligned with driving face pose. According to such dense conditions, our 3D-aware face reenactment network is called **D**ense **C**onditional face reenactment **N**etwork (DCNet) (Figure 2-I).

Secondly, we use 3DMM expression parameters captured from 3D animators to guide the reenactment results. In this way, the synthesized video frames are paired with 3DMM expression coefficients; hence, we can use these pairs as the training data to train an **E**xpression **T**racking **N**etwork (ETNet) (Figure 2-II). Finally, we use such an ETNet to replace the facial expression module of our RecNet, which upgrades the RecNet to RecAugNet. In this way, we use face reconstruction results to push the improvement of face reenactment, and use reenactment results in turn to enhance our coarse reconstruction method on the facial expression.

In summary, our contributions are as follows: Firstly, we explore the inner connection of monocular 3D face reconstruction and face reenactment and propose a novel cascade training pipeline. Secondly, we build a face reenactment network DCNet, which benefits from dense pixel-level conditions. Thirdly, our RecAugNet upgrades current face reconstruction methods on facial expression (Figure 1-RecAugNet). Finally, our pipeline JR2Net shows obvious improvements over the state-of-the-art methods on several face reconstruction benchmarks (Cao et al. 2013; Phillips et al. 2005) and the challenging face reenactment datasets (Chung, Nagrani, and Zisserman 2018).

Related Work

Face Reconstruction

Various real-time self-supervised face reconstruction solutions have been proposed (Feng et al. 2021; Shang et al. 2020; Deng et al. 2019; Tewari et al. 2018; Tran and Liu 2018; Tewari et al. 2017; Wang et al. 2022; Daněček, Black, and Bolkart 2022). While they achieve impressive results, they often do not perform well in large expression cases (*e.g.*, mouth pucker). One reason is that these methods suffer from the geometry ambiguity problem, since they rely only on the supervision of 2D attributes, *e.g.*, landmark error (between the 2D landmarks projected from the predicted 3DMMs and the ground-truth 2D landmarks). Another reason is that the ratio of large expression images in training data is often low. To tackle this issue, Wang et al. (2022) propose an emotion refinement operation. The approach of (Daněček, Black, and Bolkart 2022) leverages expression features to supervise a self-supervised face reconstruction pipeline. Although the above two mentioned methods improve expression results, they still fail on large expression cases, since their training data mostly contain neutral expressions only. Hence, our cascade training pipeline JR2Net

leverages image-expression coefficient pairs from DCNet to provide solid expression supervision and enough train samples to facilitate the face reconstruction training strategy while preserving the robustness of the self-supervised training strategy.

Face Reenactment

Recent face reenactment methods (Pumarola et al. 2018; Siarohin et al. 2019; Hsu, Tsai, and Wu 2022; Peng et al. 2021; Wang, Mallya, and Liu 2021) usually leverage 2D facial landmarks or boundaries to represent transformations from source to driving. For example, Ganimation (Pumarola et al. 2018) introduces a GAN conditioning scheme based on Action Units (AUs) annotations to generate reenacted face images driven by predicted AUs of driving faces. FOM (Siarohin et al. 2019) feeds a source image and a frame of a driving video to a keypoint detector learned in an unsupervised manner to predict sparse keypoints and local transformations, and another dense motion network accepts these intermediate representations and outputs optical flow fields to generate reenacted images. Since such methods leverage facial landmarks to construct transformations from the source image and driving frames to reenacted images implicitly, they do not formulate the shape and expression of source and driving faces individually. Hence, these methods suffer from the cross-identity problem since the relationship learned from the same person cases is difficult to adapt to different person cases.

As same as facial landmarks, several works (Thies et al. 2016; Kim et al. 2018; Yao et al. 2020; Masi et al. 2019; Nirkin, Keller, and Hassner 2019) have resorted to predicting 3DMMs for source image and driving frames and recombining them. For example, Thies et al. (2016) propose an online reenactment setup. It estimates the 3DMM parameters of input faces by optimizing facial landmarks and photo-consistency losses, and re-renders the final faces with a mouth region retrieved from the driving face sequences. Kim et al. (2018) regress 3DMM coefficients for the source image and driving frames, and render the corporation of them as the conditional input to a rendering-to-video translation network. However, their simple conditions cannot preserve the source identity and texture detail.

Our JR2Net system also leverages 3DMMs as a bridge but differs from (Kim et al. 2018; Thies et al. 2016), which only provide driving view rendering images or re-renders as the guidance. We argue that our conditional input of warped image features, normal maps, and light maps plays an important role in neural synthesis, since such guidance forms essential parts of the physical rendering pipeline. In addition, in order to obtain a one-shot face reenactment module, the source image's information is very important. Hence, we design a view synthesis module (Zhou et al. 2017) to warp the guidance from the source image to the driving image to make the multi-modal conditional input pixel-aligned with generated images (Chen et al. 2019).

Methodology

As illustrated in Figure 2, we present a novel cascade system for joint reconstruction and reenactment. In the following sections, we first introduce the face, camera, and light models. Then we propose the face reconstruction, reenactment, and expression tracking networks respectively.

Model

Face Model. We use the 3D Morphable Model (3DMM) proposed by (Paysan et al. 2009) as face shape prior. In pursuit of a controllable expression, we apply Blendshape from Polywink as our expression basis, which is manually processed to the same topology as (Paysan et al. 2009). Specifically, the 3DMM encodes both face shape and texture as

$$v = \hat{s} + b_{id} \cdot \alpha + b_{exp} \cdot \beta, c = \hat{c} + b_c \cdot \gamma, \quad (1)$$

where $\hat{s}$ and $\hat{c}$ denote the mean shape and the mean albedo, respectively. b_{id}, b_{exp}, and b_c are the PCA basis of identity, expression, and albedo, respectively. $\alpha \in R^{80}$, $\beta \in R^{51}$, and $\gamma \in R^{64}$ correspond to the 3DMM coefficients to be estimated.

Camera Model. We employ the pinhole camera model to define the projected 2D vertex as $v_{proj} = K \cdot (R \cdot v + t)$, where the face pose is represented by an Euler angle rotation $R \in SO(3)$ and translation $t \in R^3$, and the K is the intrinsic matrix.

Illumination Model. To acquire realistically rendered face images, we model the scene illumination by Spherical Harmonics (SH) (Ramamoorthi and Hanrahan 2001a,b) as $SH(N_l, c|\theta_b) = c * \sum_{b=1}^{B^2} \theta_b H_b$, where N_l is the vertex normal of a face mesh, and $\theta_b \in R^{27}$ is the coefficient. $H_b : R^3 \rightarrow R$ are SH basis functions and $B^2 = 9$ (B = 3 is the number of bands) parameterizes the colored illumination in red, green and blue channels.

Face Reconstruction

As the beginning of our JR2Net, our coarse RecNet is trained in a self-supervised manner following (Shang et al. 2020; Deng et al. 2019). RecNet regresses the 3DMM coefficients, camera poses, and SH coefficients given an RGB image. We use the ResNet-50 network (He et al. 2016) as the backbone of our RecNet to regress the above coefficients. Then we calculate the projected 2D landmarks and rendered images from the above parameters and utilize several loss functions as previous methods (Shang et al. 2020; Deng et al. 2019), *e.g.*, the render loss $\mathcal{L}_{render}$, landmark loss $\mathcal{L}_K$, identity loss $\mathcal{L}_{id}$, and regularization loss $\mathcal{L}_{reg}$. Detailed losses can be found in the *suppl.* material.

Face Reenactment

Our face reenactment module consists of RecNet and DCNet. Given a source image I_s, the corresponding normal map N_s and light image L_s from RecNet are provided to DCNet, and we do the same for driving video frames $\{I_d\}$. DCNet aims to synthesize an output image I_{gen} with the same texture as the source image and the same pose and expression as the driving frames. Figure 3 shows the architecture of our DCNet. It consists of a face texture encoder, a context encoder, and a conditional generator. We briefly introduce each module as follows.

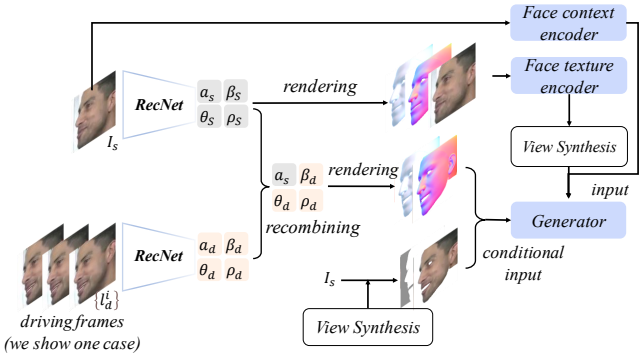

Figure 3: The training pipeline of DCNet. We first reconstruct 3DMM shape and expression coefficients α, β, 6-DoF poses p, and SH coefficients θ as 3D results from the source image I_s and driving video $\{I_d^i\}$ by our RecNet. To build the generator input, we input the source image to the context encoder, and input the source image, light map, and normal map to the texture encoder. The features from the encoders are then put into the generator as input. For the condition input, we recombine the source and driving reconstruction results and render them as the light and normal maps. Then we use view synthesis to warp the source image and mask to the driving view. Finally, we fuse these intermediate results together as the condition input.

Encoders. Our DCNet has two parallel encoders: a face texture encoder and a context encoder. The face texture encoder takes the source image, the corresponding light image and the normal map as input, aiming to remove the illumination effect from the source image and learn a pure texture representation. We design a view synthesis module to warp the face region of the texture feature to the driving frame I_t using the dense correspondence obtained between the source and driving reconstruction results. The primary step is to build the pixel correspondences between source-driving pairs (I_s, I_t). Given a pixel coordinate p_t in I_t, we compute the pixel value p_s by bilinear-sampling (Jaderberg et al. 2015):

$$p_s \sim \mathbf{K}_s[\mathbf{P}_{t\to s}^{rel}]\mathbf{D}_t(p_t)\mathbf{K}_t^{-1}p_t, \tag{2}$$

where $\sim$ represents the equality in the homogeneous coordinates; $\mathbf{K}_s$ and $\mathbf{K}_t$ are the intrinsic matrices of the source and driving poses, respectively; $\mathbf{P}_{t\to s}^{rel}$ is the projection matrix from the target view to the source view; $\mathbf{D}_t$ is the rendered depth map of the target view; $\mathbf{D}_t(p_t)$ is the depth for this particular pixel p_t in $\mathbf{D}_t$. The non-face region of the feature maps is set to zero. In this way, the warped texture features are encoded with the driving geometry, and thus it is pixel-aligned with driving frames, thus greatly benefiting the subsequent synthesis task.

The context encoder takes a source image as input and aims to encode the global context of the input image. We use dilated convolution layers and more down sampling operations to enlarge the receptive field. We upsample the context features using interpolation and concatenate them with the warped texture features. The output of the context encoder is used as the input of the generation network as shown in Figure 3.

Conditional Generator. Although our encoders provide features containing geometry and texture information, there are still two issues: (1) by compressing the source image into more compact texture features, some detailed information is lost; (2) a source image often does not contain the full texture of the face, *i.e.*, there are some invisible areas under the source pose and expression.

Hence, on the one hand, to compensate for the missing detail information, we carefully design our conditional input for the generator: we warp the source image to the driving view by the view synthesis module as Equation 2. On the other hand, to help the generator better hallucinate the missing information caused by view differences, we introduce a co-visible mask C, which encodes the validity of elements in the warped images. It is a map of the same size as the input image, with each element $C_{ij} = 1$ indicating the corresponding pixel in the driving view is visible in the source view and $C_{ij} = 0$, otherwise. In order to explicitly highlight the illumination of the driving frame, we concatenate the co-visible mask and the warped source image with the light and normal maps whose 3D geometry is composed by the 3D reconstruction results of the source image and driving frame. This concatenation is the conditional input of the generator. Inspired by the recent advance of deep generative models (Park et al. 2019; Karras et al. 2020; Karras, Laine, and Aila 2019), we transform the conditional input into modulation parameters that are multiplied and added to the normalized activation element-wise at each scale of the generator.

Losses. In order to generate reasonable face images, we apply the $\mathcal{L}_1$ loss between the generated image and the ground-truth image (driving image).

$$\mathcal{L}_1 = |I_{gen} - I_t|. \tag{3}$$

We also use the perceptual loss, which calculates the $\mathcal{L}_1$ distance between activation maps of the pre-trained VGG-19 network Θ, which can be written as:

$$\mathcal{L}_{vgg} = |\Theta(I_{gen}) - \Theta(I_t)|. \tag{4}$$

We use the adversarial loss to encourage the generator to synthesize visually realistic results:

$$\mathcal{L}_{adv} = \text{ReLU}[1 - D(I_{gen})], \tag{5}$$

where D represents the PatchGAN (Isola et al. 2017) discriminator. The loss for the discriminator is as follows,

$$\mathcal{L}_D = \text{ReLU}(1 - D(I_t)) + \text{ReLU}(1 + D(I_{gen})). \tag{6}$$

Expression Tracking

In this section, we aim to improve the expression module of RecNet, since RecNet fails to reconstruct large expressions. Our key idea is to generate image-3DMM expression coefficient pairs by our DGNet as training data for expression prediction. Then we propose the training procedure of our ETNet and a baseline, which represents a training strategy used in our initial experiments.

Expression-specific Image Generation. We generate image-3DMM expression coefficient pairs by our DCNet. In particular, the input to DCNet is different from face reenactment, since the source image and its reconstruction results are as the same as ones in face reenactment. For the driving frames, we replace the shape, pose, and illumination in the driving frame reconstruction results by the source image ones. In this way, we can generate large expression images without changing the shape, pose, and illumination of the source image, thus decreasing the artifacts of cGAN.

The 3DMM expression coefficients of the driving frames are captured by our 3D animators, *i.e.*, Facial Motion Capture , as shown in Figure 2-II. In this way, we synthesize the images from the source image with only different expressions.

Tracking Network. Given image-3DMM expression coefficient pairs, we train an expression tracking network (ETNet). In detail, firstly, our ETNet predicts expression coefficients from a ResNet34 (He et al. 2016), and the coefficients are supervised by an activation loss as $\mathcal{L}_{exp} = max(\beta, \beta_{gt}) * (\beta - \beta_{gt})$, where β_{gt} is the ground-truth expression coefficients. Since the $\mathcal{L}_1$ loss is difficult to balance the variance of each dimension of expression coefficients, we apply the activated loss. Then, we use the ETNet to replace the expression predictor in RecNet, and call the upgraded face reconstruction method as RecAugNet.

Baseline. Given image-3DMM expression coefficient pairs, there is another direction to improve the face reconstruction, i.e., upgrading the self-supervised training strategy of RecNet to semi-supervised training. In detail, we add the above pairs to the training data of RecNet, and add the expression loss on such pairs with the original losses; hence, such pairs can provide expression coefficients supervision, which may mitigate the entanglement of face shape and expression. However, through our initial experiments, we found that the ground-truth expression coefficients would conflict with the inaccurate facial landmark ground-truth, leading to even worse results than RecNet (please see the ablation study).

Figure 4: Qualitative comparisons of different methods on the Voxceleb2 dataset (Chung, Nagrani, and Zisserman 2018) under different person reenactment.

Experiment

Implementation Details

Training Data. To train our RecNet, we combine multiple datasets, including 300W-LP (Zhu et al. 2016), CelebA (Liu et al. 2015), LS3D (Bulat and Tzimiropoulos 2017), and Voxelceleb2 (Chung, Nagrani, and Zisserman 2018), which provide diversified illumination and background for training. We use Voxelceleb2 (Chung, Nagrani, and Zisserman 2018) to train DCNet, and pick source images from this dataset to generate the 3DMM-images pairs. Detailed network construction and data processing can be found in the *suppl.* material.

Method	EFL ↓	FID ↓
Bilayer	2.88	55.10
FOM	17.29	41.48
GANimation	10.89	47.36
Ours	**2.50**	**37.45**

Table 1: Quantitative results on the Voxceleb2 dataset with cross-identity reenactment.

Method	L1 ↓	PSNR ↑	SSIM ↑	FID ↓
Bilayer	0.1486	13.32	0.543	67.20
FOM	0.0604	21.28	0.713	19.57
GANimation	0.0999	17.02	0.512	24.89
Dual-G	0.1123	18.82	0.573	43.38
Ours	**0.0418**	**23.01**	**0.7798**	**19.25**

Table 2: Quantitative results on the Voxceleb2 dataset with same-identity reenactment.

Evaluation Datasets and Metrics

FaceWarehouse. To better evaluate the improvement brought by the expression enhancement, we use FaceWarehouse as the evaluation dataset. FaceWarehouse (Cao et al. 2013) is a dataset containing 20 different expressions for 150 persons. To fairly compare with the existing methods, we crop ground-truth meshes to 85 mm around the nose tip. For the alignment, we first apply similarity transformation from the ground-truth mesh to the predicted mesh through 3D landmarks, and then use the iterative closest point (ICP) algorithm (Besl and McKay 1992) to achieve fine rigid alignment. We adopt the following evaluation metrics with the unit being mm: point-to-plane root-mean-square error (RMSE), the standard deviation (STD) of RMSE, the median of RMSE, and the largest 80% of RMSE.

FRGC v2.0. We crop each ground-truth mesh to 95*mm* around the nose tip for proper alignment with all the compared methods. The alignment method used here is the same as that for processing the FaceWarehouse dataset. Furthermore, we adopt mean average error (MAE) following (Chen et al. 2020) and the STD of MAE as the evaluation metrics.

VoxCeleb2. VoxCeleb2 (Chung, Nagrani, and Zisserman 2018) is a face dataset that contains 1M videos of different celebrities. We follow the training and test split strategies proposed in their paper. We evaluate the face reenactment under two conditions: same-identity reenactment, where the

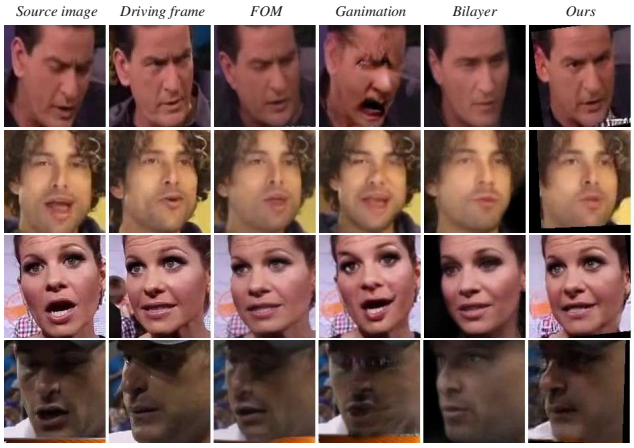

Figure 5: Qualitative comparisons on the Voxceleb2 dataset (Chung, Nagrani, and Zisserman 2018) under same-identity reenactment.

driving and source images are of the same persons, and cross-identity reenactment, where the driving and source images are of different persons.

Cross-identity Evaluation. For same-identity reenactment, the driving frames can be seen as the ground-truth, and thus the performance can be evaluated by directly measuring errors like L1 error, PSNR, and SSIM. However, since there is no ground-truth available for cross-identity reenactment, we resort to indirect evaluation. We measure the RMSE of **F**acial **L**andmarks (EFL) of generated and driving images to reflect the difference between the transferred pose/expression and ones from driving frames. As face reenactment should retain the identity of the source image, we measure the cosine similarity between the face embedding of the reenactment results and the source images by FID scores (Obukhov and Krasnyanskiy 2020)

Comparison with State-of-the-art Methods

We compare the face reenactment results of our approach with those by three state-of-the-art methods, including FOM (Siarohin et al. 2019), Bilayer (Zakharov et al. 2020), GANimation (Pumarola et al. 2018), and Dual-G (Hsu, Tsai, and Wu 2022).

Cross-identity Reenactment. Figure 4 shows the results for cross-identity reenactment. Benefiting from using 3D reconstruction results as the input and conditional input for the generator, our DCNet can transfer the poses and expressions of the driving frames while keeping the subject's identity in each source image. As seen in the first and second rows of Figure 4, our method can transfer the target motion accurately. In comparison, limited by using the facial landmarks as the representation of facial motion, FOM and Ganimation fail to reenact pose and expression when there is a large difference between the source and target poses/expressions and they end up copying the source images, as shown in the first and second rows of Figure 4. The third row of Figure 4 shows that our method preserves the white beard and the identity of the source face, and extracts the evil smile from the driving image, but Bilayer's result cannot retain the characteristics and identity of the source image.

Table 1 reports the quantitative results with cross-identity reenactment. Our results gain the lowest FID and EFL, since our DCNet benefits significantly from face reconstruction results and the dense pixel-level conditions, and avoids the entanglement of face shape and expression.

Same-identity Reenactment. Figure 5 shows the results for same-identity reenactment. As seen in the first and the second rows, our method generates more realistic images while still preserving the original identity, face detail, and vivid expression thanks to our disentangled guidance. In the third row, our method preserves the face details (forehead wrinkles) as part of the face identity, while other methods generate blurry or distorted results. As shown in the fourth row, previous approaches rarely consider the illumination change caused by the head pose variation, and thus fail to generate visually realistic results. In contrast, based on the dense pixel-level conditions, our method can extract the right light from the driving images, generate more realistic images with better face texture and more reasonable light (*e.g.*, shadow under the hat and highlight on the nose).

Table 2 reports the quantitative results with the same-identity reenactment. It can be seen that our method outperforms the competing methods on all metrics. In contrast, the artifacts of Ganimation and the missing texture details of Bilayer lead to large errors.

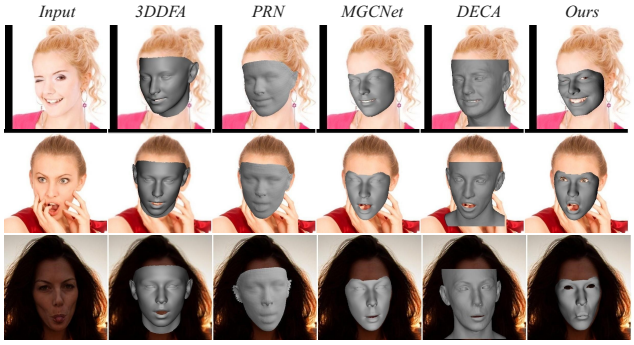

Figure 6: Qualitative comparisons of face reconstruction on the in-the-wild dataset. We overlay each result of the compared methods on the input image.

Method	MAE ↓	STD ↓
3DDFA	2.32	0.75
PRNet	2.35	0.67
MGCNet	1.94	0.67
DECA	1.93	0.59
MICA	2.12	0.61
Ours	**1.85**	**0.56**

Table 3: Quantitative reconstruction results on the FRGC v2.0 dataset.

Face Reconstruction. We compare our method with the state-of-the-art methods, including 3DDFA (Zhu et al. 2016), PRNet (Feng et al. 2018), MGCNet (Shang et al. 2020), DECA (Feng et al. 2021). We evaluate the qualita-

tive results of the compared methods on the AFLW20003D (Zhu et al. 2016) dataset, as shown in Figure 6. Our predicted 3DMM coefficients produce more vivid expressions than the other methods, *e.g.*, the closed eyes in the first row and the open mouth in the second row. Besides, since our RecAugNet/ETNet benefits from the ground-truth expression coefficients, our method does only rely on the accuracy of facial landmarks, and thus it achieves more expressive potentials than the other methods.

On the FRGC v2.0 dataset, Table 3 shows that our method outperforms the state-of-the-art methods. Our RecAugNet achieves higher fidelity and more robust variance, due to the combined advantages of unsupervised learning and expression coefficients supervision. On the FaceWarehouse dataset, as shown in Table 4, our RecAugNet also gains the best performance. It indicates that compared with the other methods or even RecNet, the paired training dataset generated from the face reenactment module can mitigate depth ambiguity and achieve better results.

Method	RMSE ↓	STD ↓	Med ↓	80% ↓	Time
3DDFA	2.24	0.44	2.22	2.60	4ms
PRNet	2.48	0.44	2.46	2.85	4ms
MGCNet	1.97	0.39	1.96	2.27	20ms
DECA	1.90	0.36	1.86	2.18	20ms
RecNet	1.90	0.40	1.85	2.24	20ms
Baseline	1.98	0.42	1.97	2.35	20ms
RecCopyNet	**1.89**	**0.34**	**1.73**	**2.04**	20ms
RecDataNet	**1.92**	**0.35**	**1.74**	**2.14**	20ms
RecAugNet	**1.78**	**0.35**	**1.74**	**2.08**	20ms

Table 4: Quantitative reconstruction results on the FaceWarehouse dataset.

Method	L1 ↓	PSNR ↑	SSIM ↑
Ours-R	0.044	22.31	0.772
Ours-P	0.045	22.38	0.773
Ours-L	0.048	22.11	0.761
Ours	**0.042**	**23.01**	**0.779**

Table 5: Quantitative results on the Voxceleb2 dataset with same-identity.

Ablation Study

Reenactment. (1) Dense pixel-level conditions: Previous studies (Yao et al. 2020; Kim et al. 2018) have explored 3DMM-based face reenactment by using rendered face images as conditional input (Kim et al. 2018). We argue that our dense pixel-level conditions are more effective since they explain the image generation process and provide more flexibility to the generator network. To demonstrate the advantage of the dense pixel-level conditions, we compare the CGANs conditioned by rendered images (Ours-R) and our dense pixel-level conditions (Ours-P) in Table 5. It can be seen that the model with dense pixel-level conditions yields better quantitative results. As discussed in the previous sections, the change of illumination conditions in which face pose and expression variation affect is also an important factor, which is rarely considered by previous approaches. We present the quantitative evaluation results of the model with and without light control (Ours-L) in Table 5. We can see that removing light control leads to significantly worse results. (2) Image warping: the last row of Table 5 reports our final results with both dense pixel-level conditions and the warped image as the conditional input. By comparing the second and the last rows of Table 5, we can see that including warped images can further improve the results.

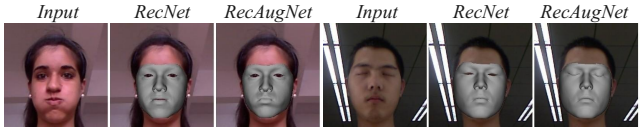

Figure 7: Qualitative comparisons for the ablation study of face reconstruction on the FaceWarehouse dataset.

Face Reconstruction. (1) RecAugNet: As shown in Figure 7, we show the improvement gains by the RecAugNet qualitatively over RecNet. Furthermore, from Figure 8, the expression (left eye) is not accurate on the expression part but on the identity part, while our RecAugNet outputs the correct expression part and achieves better final results. (2) Baseline: We add the image-3DMM expression coefficient pairs and activation expression loss to the RecNet directly and train it under a semi-supervised strategy, but the best result of this baseline is even worse than the original RecNet, shown as "Baseline" in Table 4. (3) RecCopyNet: To eliminate the improvement from the ETNet itself, we evaluate the performance of "RecCopyNet" in Table 4, which is trained as the same as RecAugNet but uses an empty 'ETNet' to predict expressions. (4) RecDataNet: To eliminate the improvement of the ETNet training data itself, we add this data to the training data of RecNet to train "RecDataNet". As shown in Table 4, such data does not directly improve the RecNet, since the landmark detector almost fails at reenacted large expression images, thus affecting the training procedure.

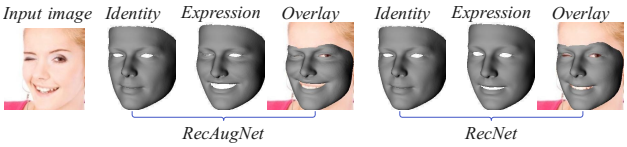

Figure 8: Visualization of the face shape and expression disentangling.

Conclusion

We have presented a novel framework for joint face reconstruction and face reenactment. Our key insight is to explore the inner connection between the two tasks, and we find that these two tasks can benefit each other. We show that using our RecNet to provide dense pixel-level conditions improves the face reenactment performance. Then, the DCNet pays back by providing paired 3DMMs-images training data to train the ETNet, which further enhances the expression module of RecNet to get RecAugNet. Hence, our JR2Net can handle large expressions and disentangle face shape and expression for face reconstruction.

References

Besl, P. J.; and McKay, N. D. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, 586–606. International Society for Optics and Photonics.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, 1021–1030.

Burkov, E.; Pasechnik, I.; Grigorev, A.; and Lempitsky, V. 2020. Neural Head Reenactment with Latent Pose Descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13786–13795.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3): 413–425.

Chen, A.; Chen, Z.; Zhang, G.; Mitchell, K.; and Yu, J. 2019. Photo-Realistic Facial Details Synthesis from Single Image. In *Proceedings of the IEEE International Conference on Computer Vision*, 9429–9439.

Chen, Y.; Wu, F.; Wang, Z.; Song, Y.; Ling, Y.; and Bao, L. 2020. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29: 8696–8705.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.

Daněček, R.; Black, M. J.; and Bolkart, T. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20311–20322.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 534–551.

Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2020. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10893–10900.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hsu, G.-S.; Tsai, C.-H.; and Wu, H.-Y. 2022. Dual-Generator Face Reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 642–650.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

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

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 3730–3738.

Masi, I.; Trn, A. T.; Hassner, T.; Sahin, G.; and Medioni, G. 2019. Face-specific data augmentation for unconstrained face recognition. *International Journal of Computer Vision*, 127(6): 642–667.

Mo, L.; Li, H.; Zou, C.; Zhang, Y.; Yang, M.; Yang, Y.; and Tan, M. 2022. Towards Accurate Facial Motion Retargeting with Identity-Consistent and Expression-Exclusive Constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2): 1981–1989.

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

Obukhov, A.; and Krasnyanskiy, M. 2020. Quality assessment method for GAN based on modified metrics inception score and Fréchet inception distance. In *Proceedings of the Computational Methods in Systems and Software*, 102–114. Springer.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2337–2346.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 296–301. Ieee.

Peng, B.; Fan, H.; Wang, W.; Dong, J.; and Lyu, S. 2021. A unified framework for high fidelity face swap and expression reenactment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6): 3673–3684.

Phillips, P. J.; Flynn, P. J.; Scruggs, T.; Bowyer, K. W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; and Worek, W. 2005. Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 947–954. IEEE.

Ploumpis, S.; Ververas, E.; O'Sullivan, E.; Moschoglou, S.; Wang, H.; Pears, N.; Smith, W.; Gecer, B.; and Zafeiriou, S. P. 2020. Towards a complete 3D morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence*.

Pumarola, A.; Agudo, A.; Martinez, A. M.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, 818–833.

Ramamoorthi, R.; and Hanrahan, P. 2001a. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 497–500. ACM.

Ramamoorthi, R.; and Hanrahan, P. 2001b. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 117–128. ACM.

Shang, J.; Shen, T.; Li, S.; Zhou, L.; Zhen, M.; Fang, T.; and Quan, L. 2020. Self-Supervised Monocular 3D Face Reconstruction by Occlusion-Aware Multi-view Geometry Consistency. *arXiv preprint arXiv:2007.12494*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32: 7137–7147.

Tewari, A.; Bernard, F.; Garrido, P.; Bharaj, G.; Elgharib, M.; Seidel, H.-P.; Pérez, P.; Zollhofer, M.; and Theobalt, C. 2019. FML: face model learning from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10812–10822.

Tewari, A.; Seidel, H.-P.; Elgharib, M.; Theobalt, C.; et al. 2021. Learning Complete 3D Morphable Face Models from Images and Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3361–3371.

Tewari, A.; Zollhöfer, M.; Garrido, P.; Bernard, F.; Kim, H.; Pérez, P.; and Theobalt, C. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2549–2559.

Tewari, A.; Zollhofer, M.; Kim, H.; Garrido, P.; Bernard, F.; Perez, P.; and Theobalt, C. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 1274–1283.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387–2395.

Tran, L.; and Liu, X. 2018. Nonlinear 3D face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7346–7355.

Wang, L.; Chen, Z.; Yu, T.; Ma, C.; Li, L.; and Liu, Y. 2022. FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20333–20342.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.

Wu, F.; Bao, L.; Chen, Y.; Ling, Y.; Song, Y.; Li, S.; Ngan, K. N.; and Liu, W. 2019. MVF-Net: Multi-View 3D Face Morphable Model Regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 959–968.

Wu, W.; Zhang, Y.; Li, C.; Qian, C.; and Loy, C. C. 2018. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European conference on computer vision (ECCV)*, 603–619.

Yao, G.; Yuan, Y.; Shao, T.; and Zhou, K. 2020. Mesh Guided One-shot Face Reenactment Using Graph Convolutional Networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1773–1781.

Zakharov, E.; Ivakhnenko, A.; Shysheya, A.; and Lempitsky, V. 2020. Fast Bi-layer Neural Synthesis of One-Shot Realistic Head Avatars. In *European Conference on Computer Vision*, 524–540. Springer.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9459–9468.

Zeng, X.; Pan, Y.; Wang, M.; Zhang, J.; and Liu, Y. 2020. Realistic face reenactment via self-supervised disentangling of identity and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12757–12764.

Zhang, J.; Zeng, X.; Wang, M.; Pan, Y.; Liu, L.; Liu, Y.; Ding, Y.; and Fan, C. 2020. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5326–5335.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1851–1858.

Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 146–155.