Expression-Aware Face Reconstruction via a Dual-Stream Network

Xiaoyu Chai, Jun Chen, *Member, IEEE*, Chao Liang, Dongshu Xu, and Chia-Wen Lin, *Fellow, IEEE*

***Abstract*—Recently, 3D face reconstruction from a single image has achieved promising progress by adopting the 3D Morphable Model (3DMM). However, face images taken in-the-wild usually involve expressions with a large range of variety. This poses difficulty to use 3DMM to represent such various facial expressions owing to the limited expressive ability of its linear model, thereby resulting in distortion and ambiguity in local facial regions. To tackle this problem, we present a novel dual-stream network composed of a geometry stream and a texture stream to deal with expression variations. Specifically, in the geometry stream, we propose novel Attribute Spatial Maps (ASMs) to decompose a face into the identity and expression attributes and then separately record the essential spatial information of the two facial attributes in the 2D image space. This avoids the interaction between the two attributes, thus preserving the identity information and further improving the ability of coping with expression variations. In the texture stream, we propose to generate facial appearance with realistic texture and canonical layout by our Semantic Region Stylization Mechanism (SRSM), that transfers the style from an input face to a 3DMM albedo map in a region-adaptive manner. Moreover, we also propose a Shared Semantic Region Prediction Module (SSRPM) to explore the common correspondence of semantic regions between the above two face texture representations. Both quantitative and qualitative evaluations on public datasets demonstrate the effectiveness of our approach in face reconstruction under expression variations.**

***Index Terms*—Attribute spatial map, dual-stream network, expression-aware, face reconstruction, facial texture synthesis.**

I. Introduction

MONOCULAR face reconstruction aims to recover the corresponding 3D face model from a single image. In recent years, intensive research efforts have been devoted to this field owing to its wide range of applications, such as face alignment [1], face editing [2], expression recognition [3], face augmentation [4], and virtual reality (VR) [5].

However, it remains challenging to reconstruct an accurate face geometry and recover photo-realistic textures from a single image. One intractable difficulty is expression variability, an attribute that humans are born with, which becomes a key issue to tackle during 3D face reconstruction owing to its variety and ambiguity [6].

To improve the performance of face reconstruction, many studies have reported the remarkable progress by adopting deep learning-based approaches. For example, the methods in [7], [8] directly regresses the shape and texture parameters of 3DMM from an input face image by using Convolutional Neural Networks (CNNs) for 3D face reconstruction. Chang *et al.* [9] employed an additional CNN to estimate robust expression parameters of faces to handle expression variations. Ferrari *et al.* [10] proposed to learn a dictionary of deformations from the deviations of a shape model between each 3D scan and the average mode. However, the 3DMM-based methods suffer from an inherent drawback that the performance of 3DMM is restricted by its linear assumption of PCA-based models: representing the face geometry by the expression and identity components within two linear sub-spaces. Human faces are, however, nonlinear in nature, making them beyond the representation ability of linear 3DMM. As a result, the 3DMM face geometry is incapable of accurately describing the full variety of human faces. Moreover, the face texture can only represent low-frequency components without sufficient details. Fig. 1 illustrates two examples that 3DMM-based 3DDFA [1] cannot well capture face geometry variations and texture details.

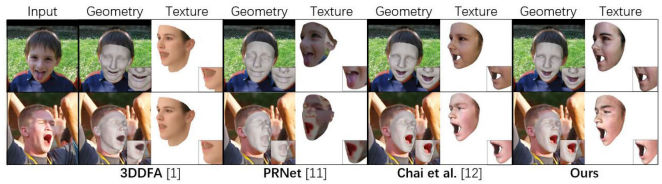


Fig. 1. Illustrations of reconstructed face geometries and textures under large expressions. 3DDFA [1] fails to accurately reconstruct the variations in geometry and details on texture. In the results of PRNet [11], the local regions around mouth and nose are ambiguous and the textures are distorted. The method in [12] produces geometry with slight misalignment around the jaw and face silhouette. In contrast, our method produces more accurate geometries and faithful textures.

Manuscript received August 17, 2020; revised November 28, 2020; accepted December 28, 2020. Date of publication March 31, 2021; date of current version September 24, 2021. This work was supported in part by National Nature Science Foundation of China 62071338, U1611461, U1736206, U1903214, 61876135, 61872362, 61671336, 61801335, and 61862015, in part by Hubei Province Technological Innovation Major Project under Grants 2017AAA123, 2018AAA062, 2018CFA024, and 2019CFB472, in part by Nature Science Foundation of Hubei Province 2018CFA024, 2019CFB472, and in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2634-F-007-013. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Jian Zhang. *(Corresponding author: Jun Chen.)*

Xiaoyu Chai, Jun Chen, Chao Liang, and Dongshu Xu are with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: stevenchai@whu.edu.cn; chenj.whu@gmail.com; cliang@whu.edu.cn; xudongshu@whu.edu.cn).

Chia-Wen Lin is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3068567.

Digital Object Identifier 10.1109/TMM.2021.3068567

Recently, nonlinear models were proposed to capture detailed face structures beyond the space of 3DMM. For instance, PRNet proposed in [11] represents face geometry with a 2D UV position map for joint 3D face reconstruction and face alignment. The methods proposed in [13], [14] learn nonlinear spaces of shapes and textures from face images via an encoder-decoder structure. Gecer *et al.* [15] adopted GAN to generate high-fidelity face textures in an unwrapped UV space. Nevertheless, since these methods resort to directly predicting the whole geometry of a face without decoupling a face's identity and expression attributes, the interactions between the two attributes make these methods vulnerable to large variations in facial expression [16]. On the other hand, synthesizing facial appearance with realistic texture directly using a generative model is challenging since the input face images are often affected by certain interference (*e.g.*, expression, pose or illumination), thus bring about distortion and blur effect into the results. Fig. 1 shows the 3D face reconstruction results with PRNet [11], where the face geometries and textures are distorted by expression variations, especially in the regions around the mouth and nose.

To decouple the identity and expression attributes of a face, the method proposed in [17] represents face geometry as a linear combination of the identity bases and the expression bases, which are independent of each other. The expression basis can be calculated from the discrepancy between expression faces and their corresponding neutral faces. The nonlinear methods in [11], [14] have demonstrated their greater representation ability by embedding face geometry into the image space than the linear 3DMM. It is thus feasible to replace the linear bases of identity and expression attributes with the 2D image space to enhance the model ability of representing the two attributes while avoiding the interaction between them, making it competent for reconstructing challenging expressions. Besides, it is convenient to deploy CNNs for generating 2D images with facial attributes, yielding a more faithful face geometry.

As for the face texture, our observation is that although the linear 3DMM albedo map has limited expressive power, it can still maintain a complete face structure and is insensitive to expression variations [8], [18]. Meanwhile, the original face image may involve distorted or incomplete regions due to expression and pose changes [11] but the real skin color is preserved. Besides, these two types of face texture representations belong to different domains, but there exist some semantic correspondences in terms of local face regions. It is, therefore, reasonable to combine the advantages of 3DMM albedo map and face image for producing facial appearance with canonical structure and faithful texture by utilizing the correspondences of semantic regions between them.

In this paper, we propose a novel dual-stream network, which involves a geometry stream to recover an accurate face shape and a texture stream to produce faithful facial appearance under expression changes. Specifically, we propose a novel Attribute Spatial Map (ASM) to embed the spatial information of decoupled identity and expression attributes into the 2D image space, rather than using the entire face geometry mixing up the two attributes. Utilizing ASMs to represent the face attributes in an unconstrained manner, the identity information can be preserved well. More importantly, it can improve the ability to cope with expression variations. Further, the reconstruction of face textures can be cast as an image style transfer problem. We regard the 3DMM albedo map as the content input and face image as the style input, then we conduct a flexible stylization process on the content input with our well-designed Semantic Region Stylization Mechanism (SRSM), which takes an adaptive semantic-region action to transfer style information, instead of conventional global stylization as in [19], [20], for synthesizing more vivid face appearance with a precise structure and photo-realistic textures. The SRSM can align the mean and variance of the content features with those of style features in a region-adaptively manner for generating credible face appearance depending on the shared semantic regions between 3DMM albedo map and face image inferred from the Shared Semantic Region Prediction Module (SSRPM), which is a key component to explore the correspondence of semantic regions between the two inputs. We further devise a fusion module to combine the output face geometry and texture into a complete 3D face model to conduct the self-supervised leaning scheme.

Our contributions are summarized as follows:

- We propose a novel dual-stream network model for expression-aware 3D face reconstruction with accurate face geometry and realistic facial textures.
- For the face geometry, we introduce ASMs to represent identity and expression attributes in an unconstrained 2D image space, which effectively enhances the expressive power for representing face geometry with expression variations.
- For the facial texture, we propose a Semantic Region Stylization Mechanism (SRSM) to generate photo-realistic face appearance in a manner of semantic region-based adaptive stylization.
- We further devise a Shared Semantic Region Prediction Module (SSRPM), which explores the correspondence of semantic regions between the 3DMM albedo map and the original face, to be incorporated in SRSM.

Compared with its preliminary conference version [12], this paper has been significantly expended in the following aspects. First, instead of synthesizing face style globally, we propose the SRSM with high flexibility to generate realistic face textures by transferring the style from face image to 3DMM albedo map in a semantic region-based adaptive manner. Second, we introduce the SSPRM, which aims to find the shared correspondence of facial semantic regions between 3DMM albedo map and face image, to be able to combine with SRSM. Third, we enhance our geometry stream by improving network architecture and applying another loss term to produce more accurate face geometry. With these newly added components, the superiority of the proposed method is thoroughly validated through extensive experiments.

The remainder of this paper is organized as follows. Section II reviews related research. In Section III, we detail the proposed dual-stream framework for face reconstruction. Section IV reports the experimental results. Finally, the conclusion is drawn in Section V.

II. Related Works

In this section, we briefly survey the related works in the literature including linear and non-linear 3D Morphable Models, style transfer in image generation, and conditional image synthesis with normalization.

A. Linear 3D Morphable Model

In the past two decades, the most widely used method for representing 3D faces is 3DMM [21], that adopts Principal Component Analysis (PCA) to build a statistical 3D face model. This seminal work facilitates representing or generating 3D facial shapes and textures by regressing the parameters of the linear 3DMM. Previously, the methods often utilize landmarks [22] and local features [23] to align an input face with the 3D template, then solve the nonlinear optimization regression function for estimating the 3DMM coefficients. With the development of deep learning, CNNs have been adopted to predict 3DMM parameters. For instance, Jourabloo *et al.* [24] and Zhu *et al.* [1] proposed using cascaded CNNs to predict accurate 3DMM coefficients, which, however, consumes much more computational complexity caused by iterations. Recently, the methods proposed by Dou *et al.* [7] and Geneova *et al.* [8] directly regress the shape and texture parameters of 3DMM for an input face image by training a CNN. Nevertheless, the linear 3DMM has rather limited power in describing the facial variability for images captured in the wild conditions.

B. Non-Linear 3D Morphable Model

More recently, there were several works that aimed to learn non-linear 3DMM to enhance model representation power. These methods use deep neural networks to represent the face model, which consist of a CNN-based decoder to estimate non-linear 3DMM, coupled with an image encoder. For instance, Tewari *et al.* [25] proposed to learn a decoder composed of multilayer perceptrons to represent the shape and albedo bases which can reconstruct arbitrary facial images. Feng [11] employed a 2D UV position map with an encoder-decoder structure to represent face shapes for joint 3D face reconstruction and alignment. The approaches proposed in [13], [14] employ either fully-connected layers or 2D convolutions to represent a facial geometry or skin albedo map. The above-mentioned methods utilize deep neural networks to define the non-linear 3DMM with great representation power. However, they still have difficulty in recovering accurate local shapes and faithful textures owing to the large variations of expression attributes and limited texture generation ability.

C. Image Transformation With Stylization

The task of image translation aims to learn the mapping between different image domains, which generally leverages conditional Generative Adversarial Networks (GANs) [26] trained on paired or unpaired data to solve this problem. Most of these approaches adopt stochastic sampling from a latent space to synthesize output images based on category labels. However, none of these methods performs local control of synthesized images since the absence of an explicit correspondence to the image local style within the latent representations.

Image stylization makes use of the image translation framework to transfer the style of a reference image onto a content image. Gatys *et al.* [27] for the first time proposed to synthesize style transferred images by matching global feature statistics in CNN layers. Since then, numerous methods have been proposed to improve the performance of synthesized images. [19] and [28] introduced feed-forward style transfer approaches which are much faster than the optimization-based alternatives. Then, [29] and [30] demonstrated other feed-forward methods that can transfer arbitrary styles thanks to a style swap layer or through a meta-network. [31] utilized a color transfer approach for the color design process to the fabric images which can generate more vivid color transfer results. Nevertheless, all the mentioned methods can only perform global style transfer, thus causing deformation in local structures, or have limited local region stylization on synthesized images. Therefore, it is desirable to adopt other strategies, for example, combining the global and local information as in [32]–[36], to improve the final results.

D. Normalization in Deep Learning

Normalization is crucial in learning CNNs such as the batch normalization proposed in [37], which accelerates training convergence of CNNs and makes training deep networks feasible. The normalization aims to make the input features approximately independently and identically distributed by using a shared mean and variance. With this property, some normalization variants find applications in conditional image synthesis including style transfer which requires additional training data. For instance, adaptive instance normalization (AdaIN) [20] and spatially-adaptive (de)normalization (SPADE) [38] normalize given feature maps, which are further affine transformed with parameters learned from features or conditions from external data. The normalization not only helps realize a flexible transfer with arbitrary new styles controlled by affine parameters, but also facilitates synthesizing images in a spatially-varying manner, *e.g.*, semantic mask based local style transfer. In general, these conditional normalization methods, such as [20], [38], [39] and [40], can promote the performance of an image transformation network to generate more flexible and plausible synthetic images.

III. Proposed Method

In this section, we introduce our dual-stream network for expression-aware monocular facial reconstruction. As shown in Fig. 2, given an input face image $\mathbf{I}_o$, the geometry stream extracts face shape $\mathbf{I}_g$ via generating identity ASM $\mathbf{A}_{id}$ and expression ASM $\mathbf{A}_{exp}$, along with parameter set $\mathcal{P}$. The texture stream aims to synthesize facial texture $\mathbf{I}_a$ with image $\mathbf{I}_o$ and texture parameter p_t. Then the fusion module combines them into final 3D face model $\mathbf{F}_m$.

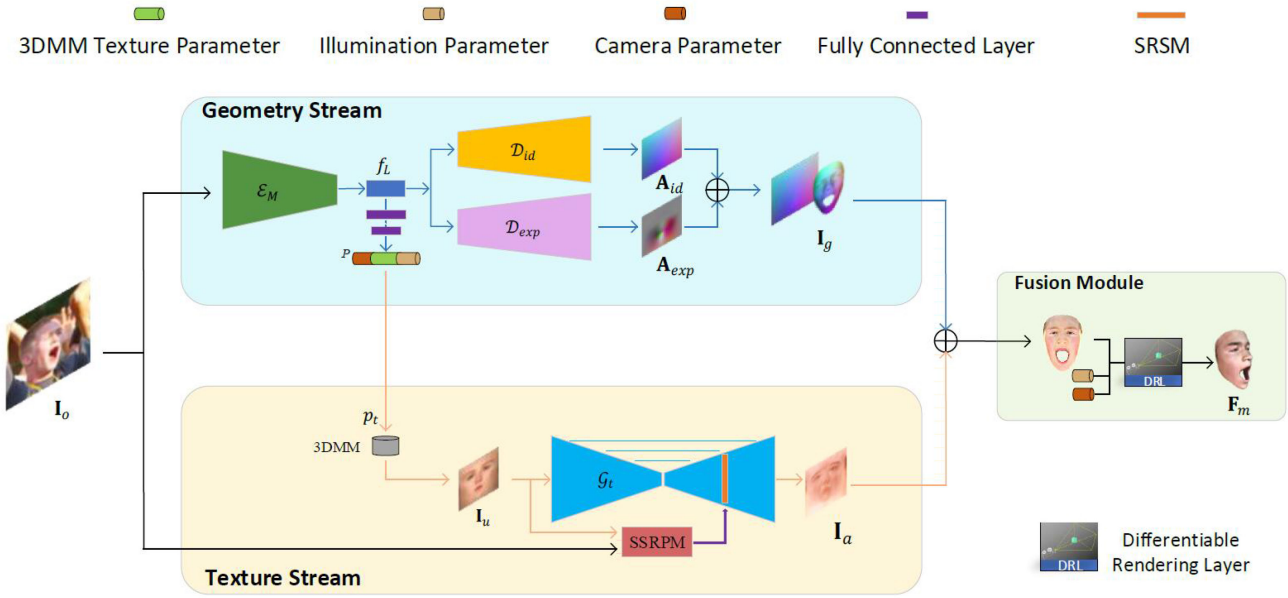


Fig. 2. Proposed dual-stream framework for expression-aware face reconstruction. The $\mathbf{I}_o$ represents the input face image in-the-wild. The following blue block indicates the Geometry Stream to produce the face spatial map $\mathbf{I}_g$ with our proposed ASMs. The parallel yellow block indicates the Texture Stream to generate unwrapped face texture map $\mathbf{I}_a$ in a semantic region-based stylization manner. Finally, the fusion module, shown in the green block, concatenates the outputs of the dual-stream together as the final 3D face model $\mathbf{F}_m$ to implement end-to-end training.

A. Geometry Stream

For an input face image, we propose a geometry stream based on a modified encoder-decoder structure to generate the corresponding attribute spatial maps including an identity spatial map and an expression spatial map, which carry the face spatial information for reconstructing the 3D facial geometry.

1) Attribute Spatial Maps: Feng *et al.* [11] first proposed to employ a 2D position map, which records the 3D shape of a complete face in the UV space, to represent the face geometry, thereby enhancing the representation power of a face model and can be directly regressed with a simple CNN. Nevertheless, these methods tend to be vulnerable to extreme expressions or large poses. Chu *et al.* [17] revealed that a face shape is a linear combination of identity and expression attributes, which are two isolated bases without affecting each other. Moreover, the expression basis can represent the offset between a face with expression and its neutral state.

Inspired by [11], [14], we introduce *Attribute Spatial Maps (ASMs)*, which comprises an identity ASM and an expression ASM, as two basic attributes to represent 3D face geometry. We first utilize the facial structure with neutral expression as a 2D image to obtain the identity ASM, where the spatial locations of all vertices for representing an identity are recorded in the R, G and B channels. Then, the offsets of spatial locations between a face expression and its neutral version are estimated as the expression ASM, which can be obtained by subtracting the identity ASM from the full 3D facial spatial map. An example of the decomposed expression and identity ASMs is illustrated in Fig. 3.

2) Face Geometry Reconstruction: From the image generation point of view, the prediction of ASMs can be regarded as the process of image syntheses from an input face image. Naturally, a deep neural network is suitable to conduct the prediction task owing to its powerful representation ability. To this end, we devise a novel encoder-decoder structure to generate ASMs, in which the geometry stream is shown as the blue block of Fig. 2.

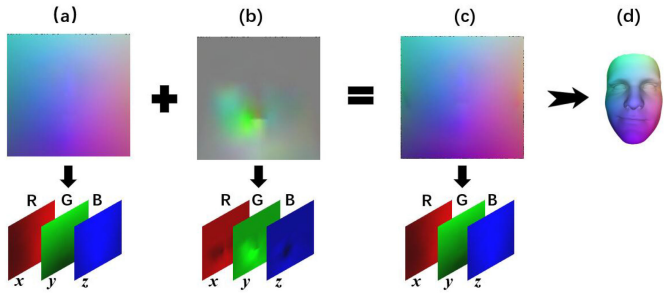


Fig. 3. Illustration of Attribute Spatial Maps (ASMs). (a) and (b) the identity and expression ASMs, respectively. (c) the full 3D face spatial map combining (a) and (b). The second row shows the R, G and B channels of ASMs correspond to the x, y and z spatial dimensions. (d) the recovered face geometry corresponding to (c) textured by normalized spatial value.

We first design a multi-task encoder (MTE) based on VGG-Face [41] to extract the latent representation as well as predict the parameters for Texture Stream (see Sec. III-B). Given image $\mathbf{I}_o$, the latent representation is the feature maps extracted from the last convolutional layers of MTE with a pooling operation. Following MTE, we design two attribute decoders (ADs) to decompose the identity ASM and expression ASM with three channels corresponding to the spatial dimension x, y and z of the attribute map. To obtain faithful ASMs, we adopt the generator architecture of Pix2pix [42] with eight layers in the decoder but change the transposed convolutions to bilinear upsampling.

Each upsampling layer followed by convolutions, denoted as Convolution-BatchNorm-ReLU layer, is with filter size 3, stride 1 and padding 1. The output layer employs a *tanh* activation to produce the output ranging in $[-1, 1]$. This procedure can be formulated as

$$\mathcal{E}_M(\mathbf{I}_o) = f_L, \tag{1}$$

$$\mathbf{I}_g = \mathbf{A}_{id} + \mathbf{A}_{exp} = \mathcal{D}_{id}(f_L) + \mathcal{D}_{exp}(f_L), \tag{2}$$

where $\mathbf{I}_o$ is the input face, $\mathcal{E}_M(\cdot)$ denotes MTE, $f_L \in \mathbb{R}^{1\times1\times512}$ is the extracted latent representation, $\mathbf{I}_g$ is the full 3D face spatial map, $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ represent the identity ASM and expression ASM, respectively. $\mathcal{D}_{id}(\cdot)$ and $\mathcal{D}_{exp}(\cdot)$ denote the ADs that generate $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ with f_L. The generated identity ASM and expression ASM are respectively visualized in Fig. 3(a) and Fig. 3(b).

Besides the two ADs following MTE, there exists a branch network, denoted as parametric sub-network (PSN), to predict the parameters of Texture Stream. PSN consisting of two fully connected layers utilizes the latent representations to predict the parameters. This procedure can be formulated as

$$\mathcal{P} = \mathcal{F}_p(f_L), \tag{3}$$

where $\mathcal{F}_p(\cdot)$ denotes the PSN, $\mathcal{P} = \{p_t, p_l, p_h\}$ is the parameter set predicted from f_L, which contains albedo parameters $p_t \in \mathbb{R}^{199}$, illumination model parameters $p_l \in \mathbb{R}^9$, and head pose parameters $p_h \in \mathbb{R}^3$.

3) Loss Functions: We train the geometry stream to predict the ASMs in a supervised learning manner with the following four loss functions.

The parametric loss ($\mathcal{L}_p$) is measured by the Euclidean distance between the predicted parameters and their ground-truths, with a regularization term on the 3DMM albedo coefficients to prevent degeneration. $\mathcal{L}_p$ is defined as

$$\mathcal{L}_p = \|\mathcal{P} - \hat{\mathcal{P}}\|_2^2 + \tau\|p_t\|_2^2, \tag{4}$$

where $\hat{\mathcal{P}}$ denotes the ground-truth parameters containing $\hat{p}_t$, $\hat{p}_l$, and $\hat{p}_h$. τ is the weight factor for the regularization term and empirically set to 5×10^{-2}.

The ASM loss ($\mathcal{L}_{asm}$) is defined as the Mean Squared Error (MSE) between the predicted ASM and their ground-truths. Besides, we apply a facial weight mask $\mathbf{W}$ on identity ASM, as suggested in [11], to emphasize the discriminative semantic locations. The ASM loss functions is defined as

$$\mathcal{L}_{asm} = \sum_{m,n} \|\mathbf{A}_{id}(m,n) - \hat{\mathbf{A}}_{id}(m,n)\|_2^2 \cdot \mathbf{W}(m,n) + \sum_{m,n} \|\mathbf{A}_{exp}(m,n) - \hat{\mathbf{A}}_{exp}(m,n)\|_2^2, \tag{5}$$

where $\mathbf{A}(m,n)$ and $\hat{\mathbf{A}}(m,n)$ represent the generated ASM and ground-truth ASM, respectively. $\mathbf{W}(m,n)$ denotes the facial weight mask, and (m,n) represents the pixel coordinates.

The symmetry loss ($\mathcal{L}_{sym}$) is used to ensure a plausible identity since the facial identity ASM without expression is symmetry. We utilize a horizontal image flip operation $\mathcal{F}_h(\cdot)$ to implement this constraint as

$$\mathcal{L}_{sym} = \|\mathbf{A}_{id} - \mathcal{F}_h(\mathbf{A}_{id})\|_1. \tag{6}$$

The regularization term ($\mathcal{L}_{reg}$) is used to encourage local smoothness, as in [25] and [14], by adding Laplacian regularization on the vertex locations for the set of all vertices:

$$L_{\text{reg}} = \frac{1}{N} \sum_{\mathbf{v}_i^{\text{mn}} \in \hat{\mathbf{A}}} \left\| \hat{\mathbf{A}}(\mathbf{v}_i^{\text{mn}}) - \frac{1}{|\mathcal{N}_i|} \sum_{\mathbf{v}_j^{\text{mn}} \in \mathcal{N}_i} \hat{\mathbf{A}}(\mathbf{v}_j^{\text{mn}}) \right\|_2^2 \tag{7}$$

where $\hat{\mathbf{A}} = \hat{\mathbf{A}}_{id} + \hat{\mathbf{A}}_{exp}$ represents the full 3D face spatial map as shown in Fig. 3(c), N is the number of vertices, $\mathbf{v}^{\text{mn}}$ is the projection of $\mathbf{v}$ onto the ASM space with location (m,n), and $\mathcal{N}_i$ is the 1-ring neighborhood of the *i*th vertex.

The total loss of geometry stream becomes

$$\mathcal{L}_{gs} = \eta_p\mathcal{L}_p + \eta_a\mathcal{L}_{asm} + \eta_s\mathcal{L}_{sym} + \eta_r\mathcal{L}_{reg}, \tag{8}$$

where η_p, η_a, η_s and η_r are the weights for different terms to scale their values to similar magnitudes. We empirically set them to 5×10^{-1}, 1, 5×10^{-1}, and 1×10^{-3}, respectively.

B. Texture Stream

The texture stream aims to generate an expression and pose invariant facial appearance with photo-realistic style learned from arbitrary face photos in-the-wild by our well-designed *Semantic Region Stylization Mechanism (SRSM)*, the pipeline of texture stream is illustrated in the yellow block of Fig. 2.

A high-fidelity synthetic facial appearance should maintain the consistency of the spatial structure and the texture with the original input face. However, the face images in-the-wild are often disturbed by some factors (*e.g.*, expression, pose or illumination), making the generated face albedo maps distorted and blurred by a conventional encoder-decoder network trained on such face images. To tackle these problems, we propose to generate facial appearance with a more flexible style transfer manner by our SRSM, which combines the advantages of both the input face with realistic texture and the unwrapped albedo map with a canonical content structure.

The key towards fulfilling this goal is the proposed Shared Semantic Region Prediction Module (SSRPM), that takes the features of the original input image and unwrapped texture map derived from the albedo coefficients p_t to predict face semantic region maps sharing the same spatial locations between them.

1) Shared Semantic Region Prediction Module: Generally, a face is composed of different regions such as eyes, nose or mouth. However, instead of directly using these physiologically defined *hard* regions, we use *n soft* semantic regions within the feature maps to establish the correspondences between $\mathbf{I}_o$ and $\mathbf{I}_u$. We assume each feature map represents a certain facial region corresponding to one of *n* semantic locations in both $\mathbf{I}_o$ and $\mathbf{I}_u$. As a result, we can synthesize the final face texture from $\mathbf{I}_u$ under the guidance of the spatial information derived from $\mathbf{I}_o$.

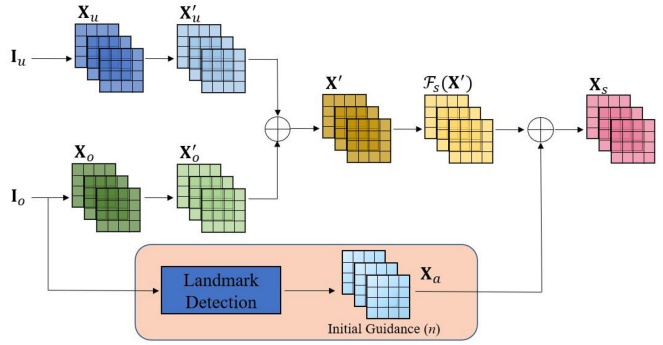


Fig. 4. Illustration of Shared Semantic Region Prediction Module (SSRPM). $\mathbf{X}_u$ and $\mathbf{X}_o$ are mid-level features extracted from $\mathbf{I}_u$ and $\mathbf{I}_o$, respectively. $\mathcal{F}_s(\mathbf{X}')$ means the n semantic regions, which are transformed from the fused feature maps $\mathbf{X}'$. The orange block represents the landmark-branch to regularize the learning process by offering a reasonable initial region guidance $\mathbf{X}_a$ with landmark locations. $\mathbf{X}_s$ is the final shared semantic region maps.

Given *n* initial desired soft semantic regions, we define their correspondence to the feature maps of input images as their inner product. The semantic representations for these soft regions are learned through back-propagation during training. In order to establish the correspondence between $\mathbf{I}_u$ and $\mathbf{I}_o$ sharing the same locations, we first align their mid-level features, denoted as $\mathbf{X}_u$ and $\mathbf{X}_o$, using two groups of filters, to produce $\mathbf{X}'_u$ and $\mathbf{X}'_o$, and then add them to obtain $\mathbf{X}'$. Then, we employ a convolutional layer with *n* filters to extract semantic regions from $\mathbf{X}'$ as soft regions. This operation transforms the joint feature maps $\mathbf{X}'$ into a new feature space $\mathcal{F}_s(\mathbf{X}') \in \mathbb{R}^{h\times w\times n}$. Generally, more accurate region scope and sufficient semantic information can be learned with a larger *n* value, which, however, increase the computational cost as well. To achieve a good trade-off between accuracy and complexity, we adopt $n = 16$ in our implementation. The structure of SSRPM is illustrated in Fig. 4.

Nevertheless, it may not be reliable to adopt this structure to generate reasonable results since it tends to assign all feature maps to a single certain region. This is due to the lack of effective guidance which leads to model performance degradation and insufficient learning of the distinctions between different semantic regions.

To address this problem, we introduce a landmark-guided semantic regions branch to regularize the learning process of SSRPM. As shown in the orange block in Fig. 4, the landmark-guided branch aims to provide reasonable initial semantic regions. It first utilizes a 3D-based face landmark detection method [43], with a stacked hour-glass architecture [44], to produce 68 channels, one for each landmark location. Then, these landmark channels are concatenated with the input face image $\mathbf{I}_o$ to aggregate to the *n*-channel feature maps, acting as the alternatives for the semantic facial regions. This procedure can be formulated as

$$\mathbf{X}_a = \mathcal{F}_a[\mathcal{K}(\mathbf{I}_o), \mathbf{I}_o], \tag{9}$$

where $\mathcal{K}(\cdot)$ is a pre-trained landmark detection model, which produces 68 channels with locations in $\mathcal{K}(\mathbf{I}_o)$, $\mathcal{F}_a(\cdot)$ aims to generate $\mathbf{X}_a \in \mathbb{R}^{h\times w\times n}$ that represents initial semantic regions.

The final shared semantic regions maps are obtained by adding $\mathcal{F}_s(\mathbf{X}')$ to the scaled $\mathbf{X}_a$, followed by a softmax operation as:

$$\mathbf{X}'_s = \mathcal{F}_s(\mathbf{X}') + k\mathbf{X}_a, \tag{10}$$

$$\mathbf{X}_{s_j} = \frac{\exp\left(\mathbf{X}'_{s_j}\right)}{\sum_{i=1}^{n}\exp\left(\mathbf{X}'_{s_i}\right)}, \tag{11}$$

where k is a scaling factor initialized as 1×10^{-2}, i and j are the indices of feature channels. Each channel in $\mathbf{X}_s$ reflects the probability of every feature pixel belonging to one of the n semantic regions.

2) Face Texture Generation: The Texture Stream takes albedo parameter p_t and original face image $\mathbf{I}_o$ as inputs to synthesize facial albedo map $\mathbf{I}_a$ with realistic texture, this process can be formulated as

$$\mathbf{I}_a = \mathcal{G}_t(\mathbf{I}_u, \mathbf{I}_o), \tag{12}$$

where $\mathcal{G}_t(\cdot)$ is the texture generation network. $\mathbf{I}_u$ is the unwrapped face albedo map derived from p_t as in [14]. To improve the faithfulness of the synthesized facial texture, we incorporate the Semantic Region Stylization Mechanism (SRSM) into the generator network for semantic region-based adaptive stylization.

Our texture stream is mainly an image translation network that receives $\mathbf{I}_u$ with unreal texture and synthesizes $\mathbf{I}_a$ with faithful appearance. This divergence focuses on the style change between the input and output, but both of them share the same content structure. To this end, existing methods [45]–[47] employ an encoder-decoder network [48], in which the features first pass through a series of progressively downsampling layers until reaching a bottleneck layer, then the process is reversed. However, these methods tend to produce blurry images.

In order to overcome the difficulty, our $\mathcal{G}_t(\cdot)$ adopts a modified version of U-Net structure to synthesize the facial texture, where skip-connections are added to better preserve low-level information shared between the input and output. Specifically, the whole network has totally h layers where skip-connections are added in between layer l and layer $h-l$. Each skip connection simply adds all channels at layer l with those at layer $h-l$. We then integrate the SRSM into $\mathcal{G}_t(\cdot)$ in the mid-level layers to import the style from $\mathbf{I}_o$ with a faithful appearance to $\mathbf{I}_u$ possessing a canonical layout.

Specifically, SRSM receives the extracted mid-level feature maps $\mathbf{X}_u^m$ and $\mathbf{X}_o^m$ of $\mathbf{I}_u$ and $\mathbf{I}_o$, respectively, and the semantic region maps $\mathbf{X}_s$ from SSRPM. It first calculates the corresponding activation maps $\mathbf{X}_u^s$ and $\mathbf{X}_o^s$ with respect to $\mathbf{X}_u^m$ and $\mathbf{X}_o^m$. Then, each semantic region in $\mathbf{X}_u^m$ with corresponding mean and variance are normalized as an instance. The style of the shared semantic regions in $\mathbf{X}_{u_i}^s$ are controlled by α_i and β_i extracted from $\mathbf{X}_{o_i}^s$ by two convolutional layers. The process can be formulated as

$$\mathbf{X}_{u_i}^s = \mathbf{X}_u^m \odot \mathbf{X}_{s_i}, \mathbf{X}_{o_i}^s = \mathbf{X}_o^m \odot \mathbf{X}_{s_i}, \tag{13}$$

$$\mathbf{X}_r^s = \sum_{i=1}^{n}\left(\frac{\mathbf{X}_u^m - \mu\left(\mathbf{X}_{u_i}^s\right)}{\sigma\left(\mathbf{X}_{u_i}^s\right)+\epsilon}\times\beta_i + \alpha_i\right)\odot\mathbf{X}_{s_i}, \tag{14}$$

where $\odot$ denotes Hadamard product, $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation of $\mathbf{X}^s_{u_i}$ computed in the way of instance normalization (IN) [28]. α_i and β_i are the learnable modulation parameters of the transformation, which depend on $\mathbf{X}^s_{o_i}$. ϵ is a small constant added to the variance for numerical stability. The final $\mathbf{X}^s_r$ guided by the semantic-region-based style information from $\mathbf{X}^m_o$ maintains the spatial structure of the original face.

3) Loss Functions: The learning objective of our texture stream consists of the following three loss terms.

The pixel reconstruction loss ($\mathcal{L}_{rec}$) enforces the generated face albedo map to be consistent with the real unwrapped face by measuring the discrepancy of the dense pixel values between the two texture representations as defined below:

$$\mathcal{L}_{rec} = \frac{1}{WHC} \sum \|\mathbf{I}_a - \hat{\mathbf{I}}_a\|_2^2, \tag{15}$$

where $\mathbf{I}_a$ and $\hat{\mathbf{I}}_a$ represent the generated albedo map and the ground-truth, respectively. W, H and C refer to the size of images in texture stream.

However, solely relying on $\mathcal{L}_{rec}$ with pixel-wise L_2-norm tends to preserve low-frequency information, but miss high-frequency details. To alleviate this, we use adversarial loss ($\mathcal{L}_{adv}$) based on a discriminator, which aims to distinguish between the generated $\mathbf{I}_a$ and real $\hat{\mathbf{I}}_a$. Specially, we employ patchGAN [49] in our discriminator to produce high-quality textures with local high-frequency details. $\mathcal{L}_{adv}$ is formulated as

$$\mathcal{L}_{adv} = \log \mathcal{D}(\hat{\mathbf{I}}_a) + \log\left(1 - \mathcal{D}\left(\mathbf{I}_a\right)\right), \tag{16}$$

where $\mathcal{D}(\cdot)$ denotes the discriminator of patchGAN.

To further encourage the generated $\mathbf{I}_a$ to fit the appearances of the semantically corresponding regions in $\hat{\mathbf{I}}_a$, we employ the contextual loss $\mathcal{L}_{con}$ proposed in [50] to match the statistics between the two faces as defined by

$$\mathcal{L}_{\text{con}} = -\log\left[\text{CX}\left(\phi^l_p(\mathbf{I}_a), \phi^l_q(\hat{\mathbf{I}}_a)\right)\right], \tag{17}$$

where p and q index the feature map of layer ϕ^l that contains n_l features, which relies on pre-trained VGG features. $\text{CX}(X, Y)$ denotes the similarity between features X and Y as in [50]. $\mathcal{L}_{\text{con}}$ uses $relu2\_2$ up to $relu5\_2$ layers since low-level features capture richer style information (*e.g.*, color or texture) useful for transferring the real appearance.

To sum up, the overall loss term in texture stream is

$$\mathcal{L}_{ts} = \omega_r \mathcal{L}_{rec} + \omega_c \mathcal{L}_{con} + \omega_a \mathcal{L}_{adv}, \tag{18}$$

where ω_r, ω_c and ω_a are the trade-off weights for different terms to have similar magnitudes. We set them as follows: 1×10^{-1}, 3 and 2×10^{-1}, respectively.

C. Fusion Module

From the geometry stream and texture stream, we receive expression-aware ASMs and faithful albedo maps, respectively. But it is difficult to achieve high quality reconstruction by independently training these two streams in a supervised manner since the available high-precision face datasets used for 3D tasks are far from enough. Thus, it is vital to find a way of making use of large amounts of off-the-shelf face datasets.

To this end, we design a Fusion Module, which incorporates $\mathbf{I}_g$ and $\mathbf{I}_a$ into the complete 3D face $\mathbf{F}_m$ according to pre-defined topology and pixel coordinates, to conduct unsupervised training, as shown in the green block of Fig. 2. Our implementation first recovers the coordinates and albedo values of each vertex on the surface, which has the same mesh topology as defined in 3DMM [21], in the way of [14]. Then, we use a differentiable renderer $\mathcal{R}$ [8] to rebuild the input face $\mathbf{I}_R$. The renderer is essentially a differentiable rasterizer based on a deferred shading model, which generates triangle IDs and barycentric coordinates for each pixel on the image plane. The rendering procedure adopts full perspective, the illumination and position parameters are also computed in the pipeline. Finally, the rendered face image $\mathbf{I}_R$ is formulated as:

$$\mathbf{I}_R = \mathcal{R}(\mathbf{I}_g, \mathbf{I}_a, p_l, p_h), \tag{19}$$

In this stage, we employ three loss terms in a way that emphasizes the collaboration of the whole framework instead of each independent sub-network. Firstly, the occlusion-aware pixel loss ($\mathcal{L}_{pix}$) aims to keep the photometric consistency between $\mathbf{I}_o$ and $\mathbf{I}_R$ with a visibility mask $\mathcal{M}$, which is estimated from a face segmentation method [52]. Secondly, the identity-preserving loss ($\mathcal{L}_{id}$) is to facilitate the similarity of faces in $\mathbf{I}_o$ and $\mathbf{I}_R$ by calculating the cosine distance between the identity features. At last, the alignment loss ($\mathcal{L}_{ali}$) calculates the Euclidean distances between detected face landmark locations of $\mathbf{I}_o$ and $\mathbf{I}_R$ extracted by a face alignment network [53]. Since the face pose is more vulnerable to degrade result quality before an accurate geometry estimation [51]. The above loss functions are formulated as follows

$$\mathcal{L}_{pix} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \|\mathbf{I}_o - \mathbf{I}_R\|_1 \odot \mathcal{M}, \tag{20}$$

$$\mathcal{L}_{id} = 1 - \frac{\mathcal{F}_{id}\left(\mathbf{I}_o\right) \cdot \mathcal{F}_{id}\left(\mathbf{I}_R\right)}{\|\mathcal{F}_{id}\left(\mathbf{I}_o\right)\|_2 \|\mathcal{F}_{id}\left(\mathbf{I}_R\right)\|_2}, \tag{21}$$

$$\mathcal{L}_{ali} = \frac{1}{K} \sum_{k=1}^{K} \|\mathcal{H}(\mathbf{I}_o) - \mathcal{H}(\mathbf{I}_R)\|_2 \tag{22}$$

where $\mathcal{M}$ is the visibility mask, $\odot$ is the Hadamard product operation, W and H are the width and height of an image, and $\mathcal{F}_{id}(\cdot)$ is a face recognition network [54]. $\mathcal{H}(\cdot)$ is the face alignment network [53], and k represents the number of landmarks.

All the loss functions in the fusion module for end-to-end training can be expressed as:

$$\mathcal{L}_{fm} = \lambda_1 \mathcal{L}_{pix} + \lambda_2 \mathcal{L}_{id} + \lambda_3 \mathcal{L}_{ali}, \tag{23}$$

where λ_1, λ_2 and λ_3 are the weights to balance the loss terms, which set as 1×10^{-1}, 2 and 1×10^{-2}.

IV. Experimental Results

In this section, we first introduce the datasets and the training strategies of our experiments. Then we demonstrate and discuss the improvements of our dual-stream network via extensive experiments and comparison with other approaches. We further

analyze the contribution of key components in our framework through ablation studies.

A. Datasets.

The training datasets mainly include 300W-LP [1], CelebA [55], and two additional synthetic datasets generated based on 300W-LP. To train the geometry stream, we first utilize the annotated coefficients in 300W-LP to generate face spatial maps with expressions and their corresponding identity ASMs. Then, the difference maps between the face spatial maps and their corresponding identity ASMs are taken as their expression ASMs. Totally, 60 000 ASM training face images including $\hat{\mathbf{A}}_{id}$ and $\hat{\mathbf{A}}_{exp}$ are used to form the ASM Dataset (ASMD). To train the texture stream, we construct a Facial Texture Dataset (FTD) by first selecting face images from 300W-LP with rare occlusion and normal lighting conditions denoted $\hat{\mathbf{I}}_o$. Then, each corresponding annotated coefficient vector $\hat{p}_t$ is transformed to the unwrapped face texture map, denoted as $\hat{\mathbf{I}}_u$. The texture of the input face image is extracted as the ground-truth albedo map $\hat{\mathbf{I}}_a$ with realistic texture. We totally collect 40 000 texture images consist of $\hat{\mathbf{I}}_o$, $\hat{\mathbf{I}}_u$, and $\hat{\mathbf{I}}_a$ in the FTD. All the images are cropped and resized to 256×256.

For face geometry evaluation, we adopt two 3D face datasets as follows. The MICC Florence 3D Faces dataset [58] provides ground-truth scans of 53 subjects with the neutral expression along with their short video footage under three settings: 'cooperative', 'indoor' and 'outdoor'. The BU-3DFE (Binghamton University 3D Facial Expression) dataset [59] contains a large range of facial expressions, in which all images are rendered with frontal views to facilitate observing the influence of expression components.

B. Training Strategies

We adopt the facial mesh topology defined in the Basel Face Model (BFM) [60] and remove the regions of ears and neck, then record the 2D locations of individual vertices on the 3D mesh. In the training process, we set all the batch size to 8 and employ the Adam Optimizer [61]. To stabilize the training process of the network, and make it easier to converge, we gradually train each part in the whole network with the following three steps. First, we utilize the annotated parameters in 300W-LP [1] to train the multi-task encoder with its parametric sub-network, then train the whole geometry stream on the ASMD. The learning rate is set as 2×10^{-4}. Second, we train the texture stream on FTD in the manner suggested in [42] with a learning rate of 1×10^{-4} for both the generator and discriminator. Finally, after the above two sub-networks have converged, we combine the dual-stream outputs with the fusion module to conduct self-supervised training and fine-tune the entire network on CelebA. We set a lower learning rate as 5×10^{-5} in this stage.

C. Face Geometry Evaluation

In this part, we evaluate the performance of our geometry stream on 3D shape reconstruction against the linear and nonlinear methods. We also conduct quantitative comparisons of reconstruction error to verify the effectiveness of our method.

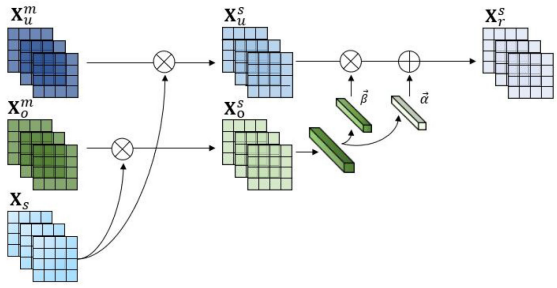


Fig. 5. Illustration of Semantic Region Stylization Mechanism (SRSM). $\mathbf{X}_u^m$ and $\mathbf{X}_o^m$ are the features extracted from $\mathbf{I}_u$ and $\mathbf{I}_o$, respectively. $\mathbf{X}_u^s$ and $\mathbf{X}_o^s$ are the features with shared semantic regions calculated by $\mathbf{X}_s$. α_i and β_i represent the affine parameters of style in semantic regions from $\mathbf{X}_o^s$. And $\mathbf{X}_r^s$ is the final region-based stylization features.

We first visually compare our geometry results with two linear 3DMM-based approaches, 3DDFA+ [51] and ExpNet [9], that reconstruct the face geometry using 228-d and 128-d shape coefficients, respectively. To evaluate the accuracy of predictions, we visualize the reconstruction error using heatmaps. Since the geometry of different methods and ground-truth may have distinct mesh topologies, we first use the ICP algorithm [62] to align the predicted results with their ground-truth, then calculate the point-to-plane distance, rather than the point-to-point distances, as in [8], to measure the face reconstruction error. As demonstrated in Fig. 6, 3DDFA+ fails to capture some extreme expressions, *e.g.*, opening or pouting mouth, whereas ExpNet leads to ambiguous in local regions description around the mouth and face contour shape, especially for those faces with head pose involving yaw and roll angles. Fig. 6 shows that our proposed method and its preliminary version [12] using ASMs can faithfully recover the face geometry with a large range of expressions. In particular, compared with its preliminary version in [12], our method can capture more accurate local details of the geometry around mouth and eyes, thanks to the newly introduced symmetry loss and alignment loss.

We also present the visual comparisons with several nonlinear model-based methods: VRN [56], PRNet [11], N-3DMM [13], and CMD [57], as shown in Fig. 7. VRN using a volumetric representation for a face, but the surface is non-smooth and does not preserve local details. PRNet and N-3DMM both use a position map to represent the whole face shape without considering the decomposition of attributes. CMD is the most recent method that adopts mesh convolutions to model face shapes. By decomposing the facial attributes into identity and expression parts with the proposed ASMs, our method can better reconstruct face geometry under a large range of expression changes. Especially from the close-up views, we can observe that, by utilizing the symmetry loss and alignment loss as supervision for the formation of face geometry, our method can produce finer local details to cover the facial silhouette of the input face, compared with the preliminary version [12].

We further conduct a user study to verify the effectiveness of our proposed method. In the experiment, We first randomly pick

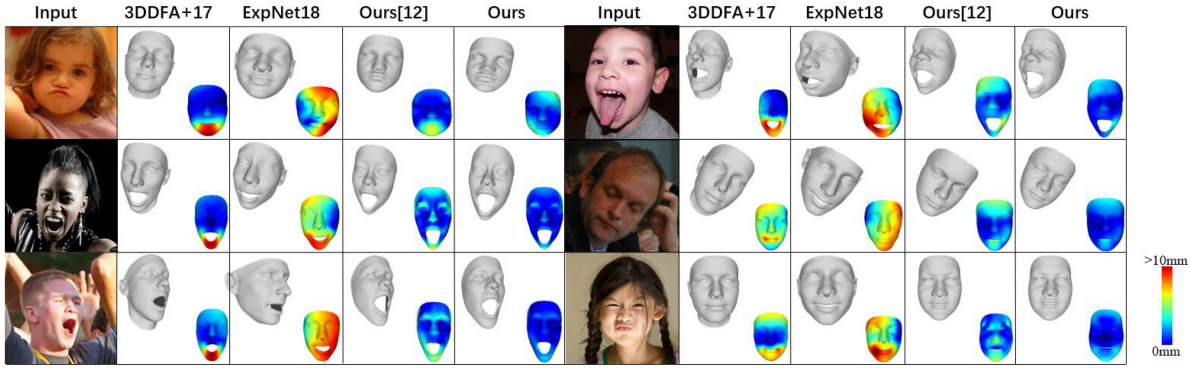


Fig. 6. Visualization of reconstruction and error comparison of face geometry with linear 3DMM-based approaches for face images with challenging expressions and various poses. The performance of 3DDFA+ [51] and ExpNet [9] are confined by the linear 3DMM basis, which can hardly reconstruct precise face shape owing to its limited representation ability. In contrast, our preliminary version [12] and this method using ASMs can faithfully recover the face geometry in terms of expression and local region reconstruction. We remove the regions about neck and ears in the heatmap for better visualization.

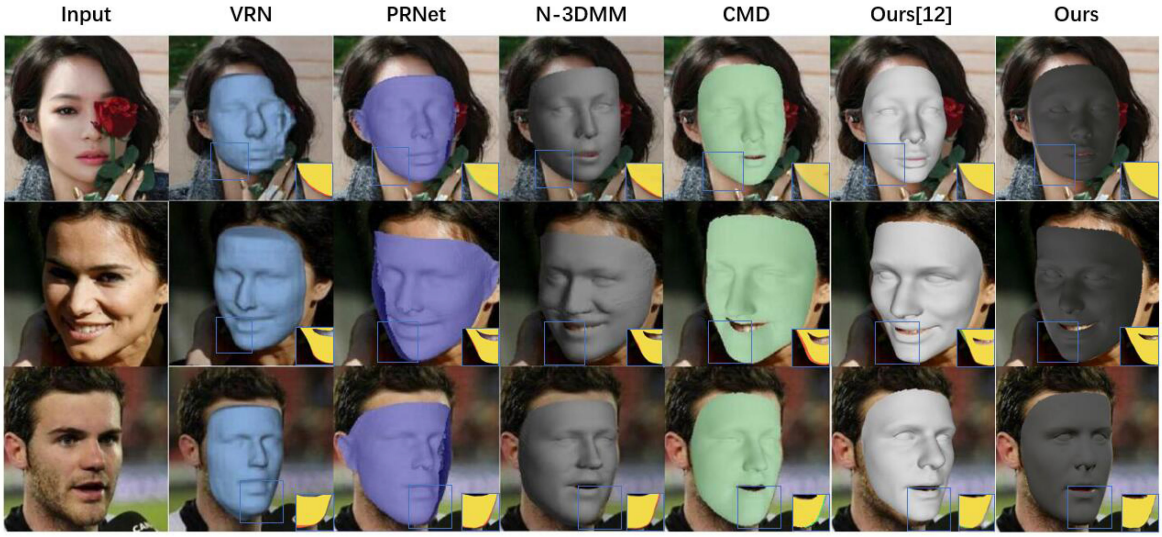


Fig. 7. Examples of face geometry reconstructions on CelebA [55]. VRN [56] is prone to lost facial structure shape with volumetric representations. PRNet [11] and N-3DMM [13] do not consider attribute decomposition, making the accurate expression variations hard to recover. CMD [57] is the recent work that using mesh convolutions to model face geometry. Compared with our preliminary work in [12] that only decouples face attributes, this method further utilizes the symmetry loss and alignment loss so that the predicted geometry can more accurately match the contours of face image. To better visualize the differences of these methods, we present the close-up views in different colors to indicate different meanings (yellow: the overlap between the geometry and face region, red: the part of the geometry stretched outside the face silhouette, green: the part of the geometry shrunk inside the face region).

a set of face geometry results (e.g., see the three examples in the first column of Fig. 7) produced by the aforementioned nonlinear approaches. We then ask 50 subjects to rate from "1" (the *worst*) to "6" (the *best*) the results reconstructed by six methods in terms of how well the geometry reflects the original facial shape subjectively. The rating distributions of the six methods illustrated in Fig. 8 show that our results receive a higher percentage of preference scores compared to the others.

We then conduct quantitative comparisons to evaluate the accuracy of the reconstructed face geometry on two datasets, MICC [58] and BU-3DFE [59], containing 3D face scans. Table I shows the results on MICC [58] dataset for identity preservation evaluation since it excludes the influence of expression

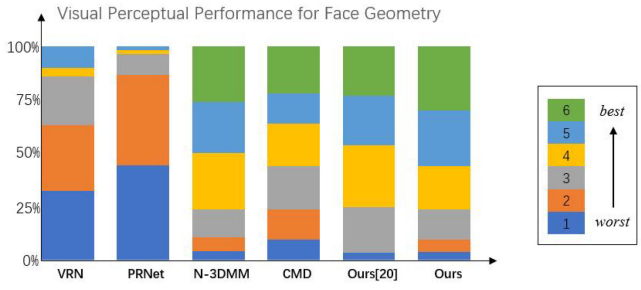


Fig. 8. Rating distributions of the perceptual user study on the quality of face geometries reconstructed by the six methods (see the horizontal axis) compared in Fig. 7. The vertical axis indicates the percentages of rating scores from 6 (the best) to 1 (the worst) received from 50 subjects.

TABLE I
FACIAL GEOMETRY ACCURACY COMPARISON ON MICC [58] WITH NEUTRAL EXPRESSION SCANS. OUR METHOD ACHIEVES LOWER AVERAGE ERROR AND VARIANCE MEASURED BY THE POINT-TO-PLANE DISTANCE UNDER DIFFERENT CONDITIONS

Method	Cooperative Mean	Cooperative Std.	Indoor Mean	Indoor Std.	Outdoor Mean	Outdoor Std.
Tran *et al.* [18]	1.93	0.27	2.02	0.25	1.86	0.23
Genova *et al.* [8]	1.50	0.13	1.50	0.11	1.48	0.11
Ours	**1.34**	**0.11**	**1.35**	**0.09**	**1.29**	**0.09**

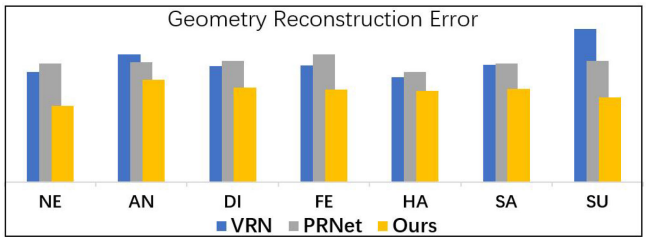


Fig. 9. Normalized geometry reconstruction error for faces in BU-3DFE involving seven expressions. Our method leads to lower error than VRN [56] and PRNet [11]. From left to right: Neutral, Anger, Disgust, Fear, Happy, Sad, Surprise.

components. We adopt the ICP algorithm [62] to align the reconstructed geometry with the ground-truths to calculate the point-to-plane distance as the reconstruction error similar to [8]. Our method outperforms the two compared methods proposed in [18] and [8] under different conditions in terms of the mean and standard deviation of reconstruction error. We also evaluate our method on BU-3DFE [59], Fig. 9 compares the normalized geometry reconstruction errors of our method, VRN [56], and PRNet [11] for faces with seven different expressions.

D. Face Texture Evaluation

In this part, we compare the performances of various methods on texture generation to validate the effectiveness of the proposed texture stream in synthesizing faithful facial appearances.

Fig. 10 compares the face textures generated by the methods in [14] and ours. Given an original image $\mathbf{I}_o$, as shown in Fig. 10(a), [14] directly adopts an encoder-decoder structure to generate the face albedo, and predicts the lighting condition, then combines them as the final face texture, as shown from Fig. 10(b) to Fig. 10(d), respectively. In contrast, our texture stream first generates the 3DMM albedo map with texture coefficient p_t. Then, the real texture in $\mathbf{I}_o$ is transferred to the albedo map using our SRSM. Subsequently, the environmental illumination of the input image is simulated and merged with the face texture to synthesize the final faithful facial appearance. From Fig. 10(e) to Fig. 10(h), we show the results of our preliminary version [12], while the last four columns demonstrate the results of our present method, where the colors of final face textures are more consistent with the input faces than the other two methods.

Fig. 11 compares some facial textures synthesized using our methods and three state-of-the-art methods. Shu *et al.* [63] proposed an unsupervised intrinsic decomposition scheme to synthesize faces. The method proposed in [64] uses high-resolution data to train networks for synthesizing textures. Gecer *et al.* [15] proposed a progressively growing GAN to generate high-fidelity face textures. To evaluate the performances of these approaches, we compute the photometric error measured by the difference between an input face and the associated rendered texture as $E_{pho} = \|R_s \odot (I_o - I_r)\|_2^2 / \mathcal{N}_p$, where R_s represents the same common face regions across all the rendered images for a fair comparison, $\mathcal{N}_p$ is the total number of pixels in R_s, and $\odot$ means the Hadamard product operation. The results in Fig. 11 demonstrate that, compared with the three stat-of-the-arts, the facial textures generated by our method are visually clearer and their colors are more consistent with those of the input faces. Moreover, the generated face textures of the present method, thanks to the proposed SRSM, have lower photometric error value than that of our previous version [12].

TABLE II
COMPARISON OF FACE TEXTURE RECONSTRUCTION ERRORS

Metrics	Linear (p_t)	Non-Linear [13]	Ours [12]	Ours
MAE	0.101	0.062	0.056	0.049

We further quantify the performance gain of our texture stream by evaluating the Mean Absolute Error (MAE) between the input face and the rendered facial appearance using our method, a linear method in p_t, and a non-linear method in [13], as shown in Table II.

E. Ablation Study

In this part, we conduct ablation studies to evaluate the effectiveness of the key components in our framework.

1) Attribute Spatial Maps (ASMs) in Geometry Stream: Our proposed ASM facilitates the geometry stream to handle extreme expression variations by separating an identity from his/her expression attributes by using the associated ASMs in an unconstrained manner. Here, we study the effects of expression ASMs on face images with various expressions. For a fair comparison, we design a non-expression ASM structure, which eliminates the expression ASM and only uses one spatial map to represent the whole face shape, while keeping the remaining configurations to train the network. Fig. 12 compares the generated face geometry with and without the expression ASM. For the various expressions in face images, the non-expression ASM structure fails in some extreme conditions, such as a wide open-mouth, which causes a distinct mismatch in the view of the overlay. In contrast, owing to its strong expressive power, the complete structure with ASMs demonstrates its superiority to faithfully reconstruct face geometries with various expressions.

2) Loss Functions in Geometry Stream: In our geometry stream, it mainly contains three loss terms to promote the generation of expression-aware face spatial maps. The ASM loss contains the MSE term weighted by the facial weight mask $\mathbf{W}$, which imposes constraints on the global and local discriminative regions, respectively. The symmetry loss imposes a constraint on the identity ASM since the non-expression spatial map is horizontally symmetric. The regularization term used to suppress

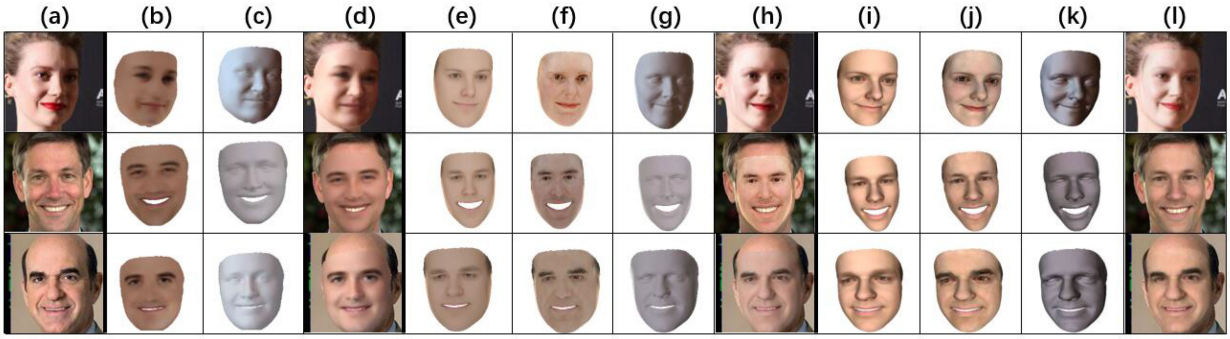


Fig. 10. Comparison of facial texture of our methods with a state-of-the-art [14]. The final texture is composed of pure albedo and scene illumination, which constitutes a high-fidelity appearance. (a) the input images. (b) to (d) the results in [14]. (e) to (h) the results of our preliminary method in [12]. (i) to (l) the results of our present method.

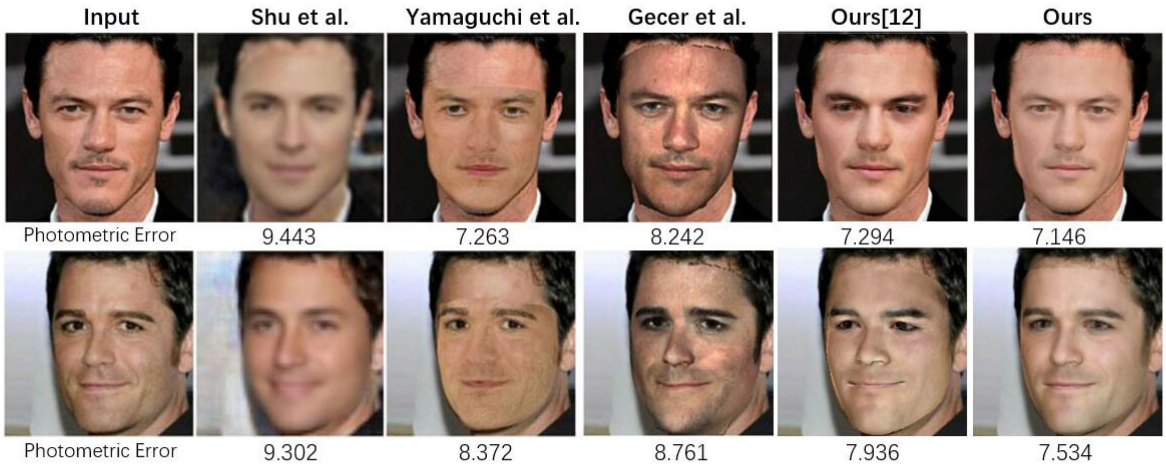


Fig. 11. Comparison with three state-of-the-art methods by overlaying the texture on the input images. The method in [63] produces blurry face image, whereas the GAN-based methods in [64] and [15] recover more faithful face textures. Our method outperforms [64] and [15] in both skin color and illumination. We calculate the photometric error for the face regions to better demonstrate the difference between the inputs and the generated textures.

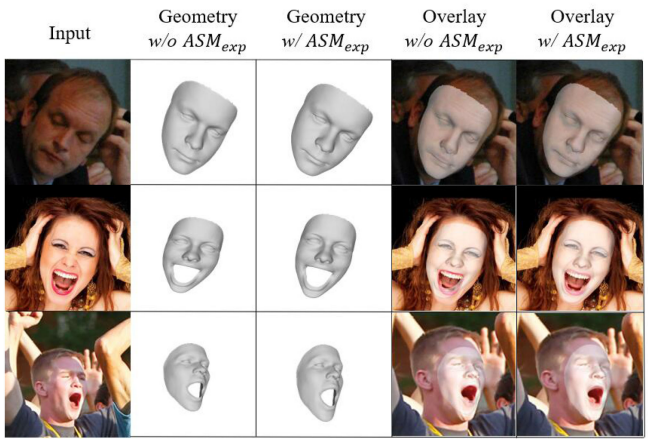


Fig. 12. Comparison of face geometry generation with and without the expression ASM. The non-expression ASM structure fails in some extreme expressions, leading to the distortions on the mouths in the view of the overlay. The complete structure with ASMs demonstrates its superiority in reconstructing the geometry under various expressions. Please zoom in for better observation.

TABLE III
ABLATION STUDY OF OUR METHOD ON AFLW2000

$\mathcal{L}_{asm}(MSE)$	$\mathcal{L}_{asm}(MSE+\boldsymbol{W})$	$\mathcal{L}_{sym}$	$\mathcal{L}_{reg}$	NME
✓				4.651
✓	✓			4.203
✓	✓	✓		4.124
✓	✓	✓	✓	3.912

noise by imposing a local smoothness constraint. To analyze the effects of these terms, we compare our complete model with two partial variants: one only adopts the MSE term without the facial weight mask and the regularization and the other utilizes the standard ASM loss which excludes the regularization term. Table III compares the NME performances of different variants on AFLW2000 [1], where the NME values are calculated between the generated and ground-truth shapes as the per-vertex mean error. It shows that the three terms in loss functions significantly reduce the reconstruction error of the generated face geometry.

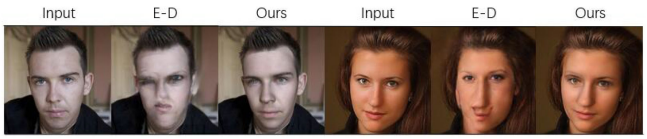


Fig. 13. Comparison of our model with a conventional encoder-decoder model for generating facial textures. The encoder-decoder approach tends to produce blurry face appearances with local spatial deformation, especially around the mouth and nose. Our approach generates more faithful face texture map. Please zoom in for better observation.

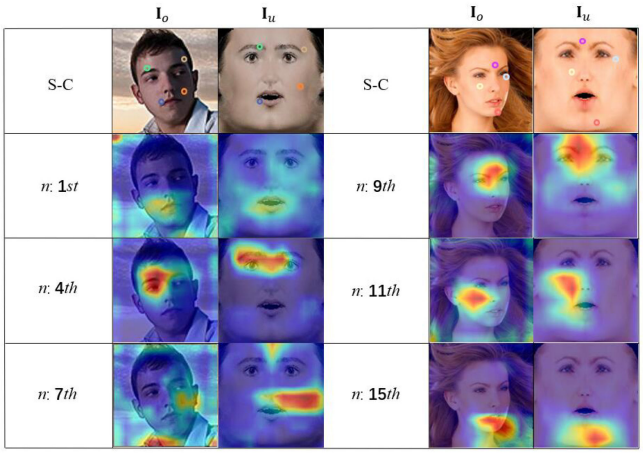


Fig. 14. Visualization the correspondence relations and shared semantic regions between input face image and unwrapped texture map. The first row demonstrates the Sparse Correspondence (S-C) using circles with the same color. The second to fourth rows further visualize the shared semantic regions by choosing some activation maps to demonstrate the function of SSRPM.

3) Generator Architecture in Texture Stream: The generator architecture of our texture stream is a U-Net based network, which utilizes skip-connections between the mirror layers to preserve the local spatial information in the original face image thus producing high-fidelity texture map. To verify the ability of the generator structure, we compare our model against an encoder-decoder version without skip-connections on face texture map generation in Fig. 13.

4) Semantic Region Stylization in Texture Stream: The core function of texture stream is to stylize semantic regions using our SSRPM, that establishes a soft correspondence of shared semantic regions between the input face and its unwrapped texture map. These relations can be utilized to synthesize realistic facial appearance with the semantic region-based stylization rather than the global manner. Fig. 14 shows the correspondence relations in SSRPM, where circles with the same color are used to illustrate some correspondences between the two images, as shown in the first row. We then visualize the shared semantic regions, as in [65] to illustrate the locations of a case in an image, with activation maps from $\mathbf{X}_r^s$ each representing a semantic region shared between them. The 1st, 4th, 7th, 9th, 11th, 15th activation maps are randomly chosen to show some regions, such as mouth, eyes, cheek, and jaw.

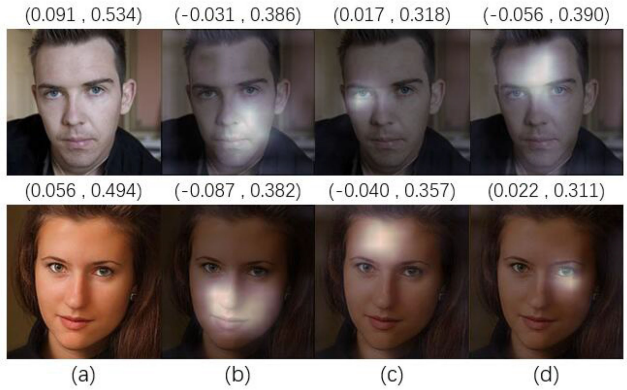


Fig. 15. Visualization of semantic regions and their corresponding feature statistics. (a) The original face images without indicating semantic regions. (b) to (d) randomly chosen activation maps to respectively illustrate the semantic regions they focus on. Above the images are the corresponding mean and standard values for the feature maps, calculated from the whole images in (a), and only from the highlighted regions in (b), (c), and (d).

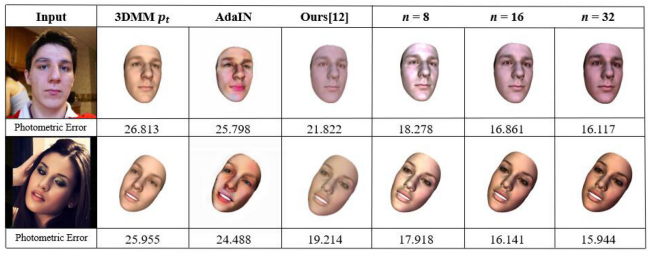


Fig. 16. Stylization ability comparison. The first column shows the input face images. The first group presents face textures, generated by 3DMM, AdaIN [20] and our previous method [12], which are global-based methods and cause color inconsistency with the inputs. The second group shows our semantic-region-based stylization facial textures with different n values in SSRPM. When the value of n increases, more regions can establish their correspondence, hence the generated face texture gets closer to the original face texture. The photometric error indicates below each generated face texture further demonstrates the performance of synthetic face textures and the superiority of our semantic-region-based method.

The learned semantic regions by SSRPM in $\mathbf{X}_r^s$ indicate those regions with high correlation in semantics. As shown in Fig. 15, we randomly select several activation maps to highlight the semantic areas (from Fig. 15(b) to Fig. 15(d)) of the corresponding face images (Fig. 15(a)). Then the statistics, including the mean and standard deviation, computed from diverse regions are different from others, which represent specific style feature information within one semantic region.

Fig. 16 verifies the effectiveness of SSRPM with different stylization manners for face texture generation. We divide all the methods into two groups: one only using global-based methods and the other adopting semantic-region-based manner with different settings. In group one, we first show the initial face appearance by directly using the 3DMM texture parameter p_t. Then we adopt AdaIN [20] to perform style transfer from the input face to the 3DMM texture map. Since the affine parameters α and β of AdaIN represent the global style information in the source style image, AdaIN cannot well stylize local regions, thereby

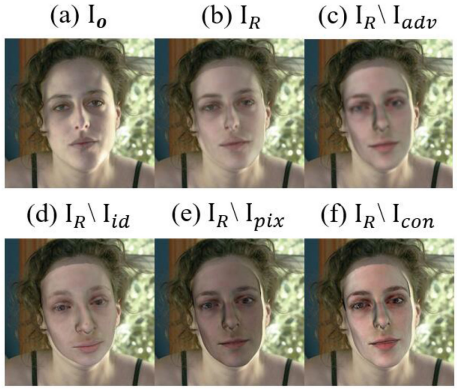


Fig. 17. Illustrations of the contributions of individual loss terms of our method with the leave-one-out ablation study. The results demonstrate the necessity of each component, which makes a significant contribution towards a good reconstruction. Please zoom in for better observation.

introducing local color inconsistency artifacts. We also present the results of our preliminary version [12] that uses the transformation network (TN) to directly synthesize the face texture map without considering the semantic regions on faces. In group two, we demonstrate the results of our proposed SSRPM with different n values. A small value of n (e.g., $n = 8$) causes insufficient stylization since some regions lack adequate correspondences. In contrast, $n = 16$, or $n = 32$ achieves higher-quality reconstruction. The photometric error are further computed to evaluate the performance of different settings. Finally, we choose $n = 16$ as it reaches a good trade-off between performance and computational cost.

5) Other Loss Terms: Here, we conduct an ablation study on our method to verify the full model can reconstruct the input face better than its variants. Specifically, we adopt the leave-one-out strategy for loss terms related to the final facial appearance to investigate the individual contributions of them. As shown in Fig. 17, each of our components significantly contributes towards a good face reconstruction. Fig. 17(a) and Fig. 17(b) show the input face and the final rendered version. Fig. 17(c) shows that the adversarial loss effectively avoids hazy effects and faithfully recovers the details of texture. Fig. 17(d) shows the identity terms contributes to identity preservation. Moreover, the pixel intensity constraint can promote color consistency as shown in Fig. 17(e). Finally, Fig. 17(f) demonstrates the content loss can better capture the albedo and illumination.

V. Conclusion

In this paper, we proposed a dual-stream network, involving a geometry stream and a texture stream, to achieve expression-ware monocular face reconstruction. In the geometry stream, we recover accurate geometries under various expressions by learning decoupled attribute spatial maps for both identity and expression. In the texture stream, we synthesize realistic facial appearances by using a semantic region-based stylization method by combining the advantages of 3DMM albedo map and original face image. Quantitative and qualitative results demonstrate the effectiveness of our method in handling challenging expressions.

References

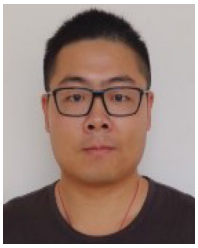
[1] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 146–155.
[2] Z. Geng, C. Cao, and S. Tulyakov, "3D guided fine-grained face manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9821–9830.
[3] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2D+ 3D facial expression recognition with deep fusion convolutional neural network," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2816–2831, Dec. 2017.
[4] H.-C. Shao, K.-Y. Liu, W.-T. Su, C.-W. Lin, and J. Lu, "Domain-transferred face augmentation network," in *Proc. Asian Conf. Comput. Vis.*, Dec. 2020, pp. 309–325.
[5] J. Lou *et al.*, "Realistic facial expression reconstruction for VR HMD users," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 730–743, Mar. 2020.
[6] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, no. 6-7, pp. 884–906, 2019.
[7] P. Dou, S. K. Shah, and I. A. Kakadiaris, "End-to-end 3D face reconstruction with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5908–5917.
[8] K. Genova *et al.*, "Unsupervised training for 3D morphable model regression," in *Proc. IEEE Conf. Comput. Vis*. Pattern Recognit., 2018, pp. 8377–8386.
[9] F.-J. Chang *et al.*, "ExpNet: Landmark-free, deep, 3D facial expressions," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 122–129.
[10] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "A dictionary learning-based 3D morphable shape model," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2666–2679, Dec. 2017.
[11] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Prof, Eur. Conf. Comput. Vis.*, 2018, pp. 534–551.
[12] X. Chai, J. Chen, C. Liang, D. Xu, and C.-W. Lin, "Expression-aware face reconstruction via a dual-stream network," in *Proc. IEEE Conf. Multimedia Expo*, 2020, pp. 1–6.
[13] L. Tran and X. Liu, "Nonlinear 3D face morphable model," in *Proc. IEEE Conf. Comput. Vis*. Pattern Recognit., 2018, pp. 7346–7355.
[14] L. Tran and X. Liu, "On learning 3D face morphable model from in-the-wild images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 157–171, Jan. 2021.
[15] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1155–1164.
[16] K. Fisher, J. Towler, and M. Eimer, "Facial identity and facial expression are initially integrated at visual perceptual stages of face processing," *Neuropsychologia*, vol. 80, pp. 115–125, 2016.
[17] B. Chu, S. Romdhani, and L. Chen, "3D-aided face recognition robust to expression and pose variations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1899–1906.
[18] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3D morphable models with a very deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5163–5172.
[19] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[20] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.
[21] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. Conf. Comput. Graph. Interact. Tech.*, 1999, pp. 187–194.
[22] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2387–2395.
[23] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Rätsch, "Fitting 3D morphable face models using local features," in *Proc. IEEE Int. Conf. image Process.*, 2015, pp. 1195–1199.
[24] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4188–4196.
[25] A. Tewari *et al.*, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz," in *Proc. IEEE Conf. Comput. Vis*. Pattern Recognit., 2018, pp. 2549–2559.

[26] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[27] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016, pp. 2414–2423.
[28] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2017, pp. 6924–6932.
[29] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," 2016, *arXiv:1612.04337*.
[30] F. Shen, S. Yan, and G. Zeng, "Neural style transfer via meta networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2018, pp. 8061–8069.
[31] Y. Han, C. Xu, G. Baciu, M. Li, and M. R. Islam, "Cartoon and texture decomposition-based color transfer for fabric images," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 80–92, Jan. 2017.
[32] H.-M. Hu, H. Zhang, Z. Zhao, B. Li, and J. Zheng, "Adaptive single image dehazing using joint local-global illumination adjustment," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1485–1495, Jun. 2020.
[33] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.
[34] W. Ruan *et al.*, "Multi-correlation filters with triangle-structure constraints for object tracking," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1122–1134, May 2019.
[35] W. Ruan, C. Liang, Y. Yu, J. Chen, and R. Hu, "SIST: Online scale-adaptive object tracking with stepwise insight," *Neurocomputing*, vol. 384, pp. 200–212, Mar. 2020.
[36] W. Ruan *et al.*, "Poinet: Pose-guided ovonic insight network for multi-person pose tracking," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 284–292.
[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.
[38] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2019, pp. 2337–2346.
[39] Y. Wang, Y.-C. Chen, X. Zhang, J. Sun, and J. Jia, "Attentive normalization for conditional image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 5094–5103.
[40] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2020, pp. 5143–5153.
[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[42] 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*., 2017, pp. 1125–1134.
[43] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem?(and a dataset of 230000 3D facial landmarks)," in *Proc. Int. Conf. Comput. Vis*., 2017, pp. 1021–1030.
[44] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis*., 2016, pp. 483–499.
[45] Y. Liu, W. Chen, L. Liu, and M. S. Lew, "SwapGAN: A multistage generative approach for person-to-person fashion style transfer," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2209–2222, Sep. 2019.
[46] Y. Guo *et al.*, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2726–2737, Nov. 2019.
[47] X. Han, H. Yang, G. Xing, and Y. Liu, "Asymmetric joint GANs for normalizing face illumination from a single image," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1619–1633, Jun. 2020.
[48] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[49] A. Shrivastava *et al.*, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2017, pp. 2107–2116.
[50] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis*., 2018, pp. 768–783.
[51] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, Jan. 2019.
[52] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit*., 2018, pp. 98–105.
[53] J. Deng, Y. Zhou, S. Cheng, and S. Zaferiou, "Cascade multi-view hourglass model for robust 3D face alignment," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit*., 2018, pp. 399–403.
[54] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2019, pp. 4690–4699.
[55] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. Int. Conf. Comput. Vis*., 2015, pp. 3730–3738.
[56] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric cnn regression," in *Proc. Int. Conf. Comput. Vis*., 2017, pp. 1031–1039.
[57] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou, "Dense 3D face decoding over 2500fps: Joint texture & shape convolutional mesh decoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2019, pp. 1097–1106.
[58] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2D/3D hybrid face dataset," in *Proc. Joint ACM Workshop Hum. Gesture Behav. Understanding*, 2011, pp. 79–80.
[59] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. Automa. face Gesture Recognit*., 2006, pp. 211–216.
[60] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. AVSS*, 2009, pp. 296–301.
[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[62] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor fusion IV: control paradigms data structures*, vol. 1611, 1992, pp. 586–606.
[63] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2017, pp. 5541–5550.
[64] S. Yamaguchi *et al.*, "High-fidelity facial reflectance and geometry inference from an unconstrained image," *ACM Trans. Graph*, vol. 37, no. 4, 2018, Art. no. 162.
[65] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam : Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

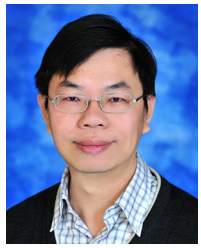
Xiaoyu Chai received the M.S. degree from the China University of Geosciences, Wuhan, China, in 2016. He is currently working toward the Ph.D. degree in communication and information system with the National Engineering Research Center for Multimedia Software, Computer School of Wuhan University, Wuhan, China. His research interests include image and video processing, computer vision, and facial analysis.

Jun Chen (Member, IEEE) received the M.S. degree in instrumentation from the Huazhong University of Science and Technology, Wuhan, China, in 1997 and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008. He is currently the Deputy Director of the National Engineering Research Center for Multimedia Software and a Professor with the School of Computer Science, Wuhan University. He has authored or coauthored more than 50 papers in his research fields, which include multimedia analysis, computer vision, and security emergency information processing.

Chao Liang is currently an Associate Professor with the National Engineering Research Center for Multimedia Software, Computer School of Wuhan University, Wuhan, China. He has authored or coauthored more than 60 papers, including premier conferences, which include CVPR, ECCV, ACM MM, AAAI, IJCAI and honorable journals like TNNLS, TMM, and TCSVT. His research interests include multimedia content analysis and retrieval, computer vision, and pattern recognition. He was the recipient of the Best Paper Award of PCM 2014.



Dongshu Xu received the M.E. degree from Yunnan University, Kunming, China, in 2016. He is currently working toward the Ph.D. degree in communication and information system with the School of Computer Science, Wuhan University, Wuhan, China. His research interests include person re-identification and computer vision.



Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000. He is currently a Professor with the Department of Electrical Engineering, Institute of Communications Engineering, NTHU. He is also the Deputy Director of the AI Research Center, NTHU. During 2000–2007, he was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Minxiong, Taiwan. Prior to joining academia, he was with Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, during 1992–2000. His research interests include image and video processing, computer vision, and video networking. Dr. Lin was a Distinguished Lecturer of IEEE Circuits and Systems Society from 2018 to 2019, a Steering Committee Member of IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2015, and the Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2015. He was the recipient of the Best Paper Award of IEEE VCIP 2015, the Top 10% Paper Awards of IEEE MMSP 2013, and the Young Investigator Award of VCIP 2005 for his articles. He was also the recipient of the Outstanding Electrical Professor Award presented by Chinese Institute of Electrical Engineering in 2019 and Young Investigator Award presented by Ministry of Science and Technology, Taiwan, in 2006. He is currently the Chair of the Steering Committee of IEEE ICME. From 2019 to 2020, he was the President of the Chinese Image Processing and Pattern Recognition Association, Taiwan. He was the Technical Program Co-Chair of the IEEE ICME 2010, the General Co-Chair of the IEEE VCIP 2018, and the Technical Program Co-Chair of the IEEE ICIP 2019. He was an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE MULTIMEDIA, and *Journal of Visual Communication and Image Representation*.