Computers & Graphics (2024)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Self-supervised Reconstruction of Re-renderable Facial Textures from Single Image

Mingxin Yang[a,c], Jianwei Guo[b,c,*], Xiaopeng Zhang[b,c], Zhanglin Cheng[a,c,*]

[a] *Shenzhen Key Laboratory of Visual Computing and Analytics, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China*
[b] *MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China.*
[c] *University of Chinese Academy of Sciences, Beijing, China.*

ARTICLE INFO

Article history:
Received October 29, 2024

Keywords: Facial Albedo Reconstruction, 3D Face, Self-supervised, Illumination Reconstruction

ABSTRACT

Reconstructing high-fidelity 3D facial texture from a single image is a quite challenging task due to the lack of complete face information and the domain gap between the 3D face and 2D image. Further, obtaining re-renderable 3D faces has become a strongly desired property in many applications, where the term 're-renderable' demands the facial texture to be spatially complete and disentangled with environmental illumination. In this paper, we propose a new self-supervised deep learning framework for reconstructing high-quality and re-renderable facial albedos from single-view images in the wild. Our main idea is to first utilize a *prior generation module* based on the 3DMM proxy model to produce an unwrapped texture and a globally parameterized prior albedo. Then we apply a *detail refinement module* to synthesize the final texture with both high-frequency details and completeness. To further make facial textures disentangled with illumination, we propose a novel detailed illumination representation that is reconstructed with the detailed albedo together. We also design several novel regularization losses on both the albedo and illumination maps to facilitate the disentanglement of these two factors. Finally, by leveraging a differentiable renderer, each face attribute can be jointly trained in a self-supervised manner without requiring ground-truth facial reflectance. Extensive comparisons and ablation studies on challenging datasets demonstrate that our framework outperforms state-of-the-art approaches.

1. Introduction

Reconstructing high-fidelity 3D human faces is a longstanding problem in computer vision and graphics communities. This task aims to estimate a realistic 3D facial representation, *i.e.,* predicting face geometry, appearance, expression, and scene lighting from the input source. Faithfully reconstructing 3D faces is a crucial prerequisite for many downstream applications including face recognition [1, 2], face editing [3], face alignment [4, 5], and virtual avatar [6, 7].

Recently, single image based 3D face reconstruction has gained much attention. However, it is a highly challenging and ill-posed problem due to the domain gap between the 3D face and 2D image. To learn the mapping from a single image to a 3D face, a parametric model called *3D Morphable Model* (3DMM) [8] is developed as the prior model of a 3D face that transforms the 3D reconstruction to a parameter estimation problem. However, 3DMM largely limits the representation capability of the parametric model because it was constructed by applying linear subspace modelling techniques on a limited number of 3D face scans, thus leading to poor reconstruction fidelity when being applied to in-the-wild images [9, 10, 11]. Recently, many attempts have been conducted to tackle the detail lacking drawback of 3DMM by adding nonlinearity into the parametric model, for example, replacing the

*Corresponding authors: Jianwei Guo (jianwei.guo@nlpr.ia.ac.cn); Zhanglin Cheng (zl.cheng@siat.ac.cn);

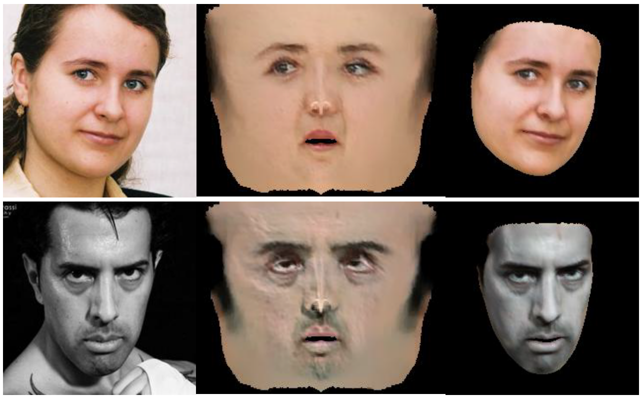

Fig. 1. We learn to reconstruct high-fidelity facial textures from in-the-wild images. Left: input single images; Middle: detailed albedo maps generated by our neural network; Right: re-rendered results using our detailed albedo and fine-grained illumination maps.

linear 3DMM with a completely non-linear one [12, 13, 14] or complementing non-linearity upon the 3DMM coarse reconstruction [15, 16, 17]. In these methods, facial details are either represented in geometry by a displacement map or encoded into appearance by a detailed texture map (or albedo map). In this work, we focus on high-fidelity appearance reconstruction and apply a coarse-to-fine approach to generate textures that capture facial details.

The methods for reconstructing facial textures can be further roughly divided into two categories. The first category extends the basic idea of the parametric model and utilizes a self-collected facial texture dataset to train a generative model [18, 10, 12]. When estimating a new image, these approaches fit the closest texture in the subspace to the input. They could achieve high-quality results even when the inputs are occluded or in extreme light conditions. However, their generation results can not maintain the idiosyncrasy of the human faces well because of the limited representation capacity of the generative model.

The other category typically reconstructs the texture directly from the input image [16]. Although their reconstruction corresponds to the input image better, the reconstruction quality is highly influenced by the input, and noise-like occlusion and extreme environmental illumination will cause artifacts baked in the reconstructed texture. Apart from the requirements of high-fidelity texture reconstruction, many applications (*e.g.,* virtual avatar) demand the texture to be re-renderable. Specifically, the texture should be not only faithful to the input image, but also disentangled with illumination (which is referred to as an albedo). However, above mentioned methods can not solve the disentanglement of face albedo with illumination. The reasons are two-fold: (1) real facial textures are difficult to capture without illumination [12]; (2) the widely used three-band *spherical harmonics* (SH) lighting model has a limited representational capacity [11].

To address these mentioned issues, we propose a new self-supervised learning algorithm that takes both the advantages of above two categories of methods to generate high-fidelity and re-renderable facial albedos. Our method adopts a coarse-to-fine paradigm which first utilizes a *prior albedo generation module* to produce a coarse re-renderable albedo as a prior, then adds facial details on the prior by a *detail refinement module*. Specifically, we adopt a pre-trained inference network based on 3DMM to produce a prior albedo from the input image. Then, we transform the prior albedo into a complete and detailed facial texture by employing an image-to-image translation network to preserve high-frequency details. In addition, we introduce a novel detailed illumination representation and propose a decoder to make the albedo disentangled from environmental illumination. This property is especially useful for rendering from novel viewpoints. Several regularization loss functions are designed on both the illumination side and albedo side for achieving a high-fidelity and re-renderable albedo. Finally, our pipeline can be efficiently trained in a self-supervised manner with the help of differentiable rendering [10]. Fig. 1 gives two examples of our reconstruction results. In summary, our work makes the following contributions:

- We propose a new self-supervised neural network to obtain a high-fidelity and re-renderable facial albedo. We are able to deal with potential occlusions commonly existed in facial images. The self-supervised learning further makes our approach generalize well among other unseen data in-the-wild.

- We devise a novel representation of detailed illumination by a localized spherical harmonics to achieve a more accurate illumination estimation, which alleviates the limited expressiveness of widely used SH-based lighting model.

- We design several novel regularization losses to ensure that the detailed albedo is similar to the prior coarse albedo while keeping high-frequency details. Especially, the cross perceptual loss is effective to disentangle lighting from person-specific details such as beards and wrinkles.

2. Related work

2.1. Parametric Models for the Human Face

The seminal parametric model of 3DMM was introduced by [8], which applies subspace modeling on collected 3D face scans and produces low-dimensional representations for facial identity, expression and albedo. Many variants [19, 20, 21, 22, 23] have extended it to obtain better performance [24]. To improve representation power, parametric models with non-linearity are introduced [25, 26, 27, 13]. Although this model expands the representation capacity of 3DMM, the local modeling scheme leads to stitching artifacts in the generated results. Ganfit [12] utilizes a progressive GAN [28] to construct a generative model, which collects a dataset of high-resolution human facial textures and trains the network on it. However, the model has the drawback that the illumination is baked into the texture. Lattas et al. [6] extend Ganfit by post-processing (super-resolution, de-lighting and BRDF inference) the derived texture. Its limitation is that the captured dataset does not contain sufficient samples of different ethnicities and may produce unfaithful results.

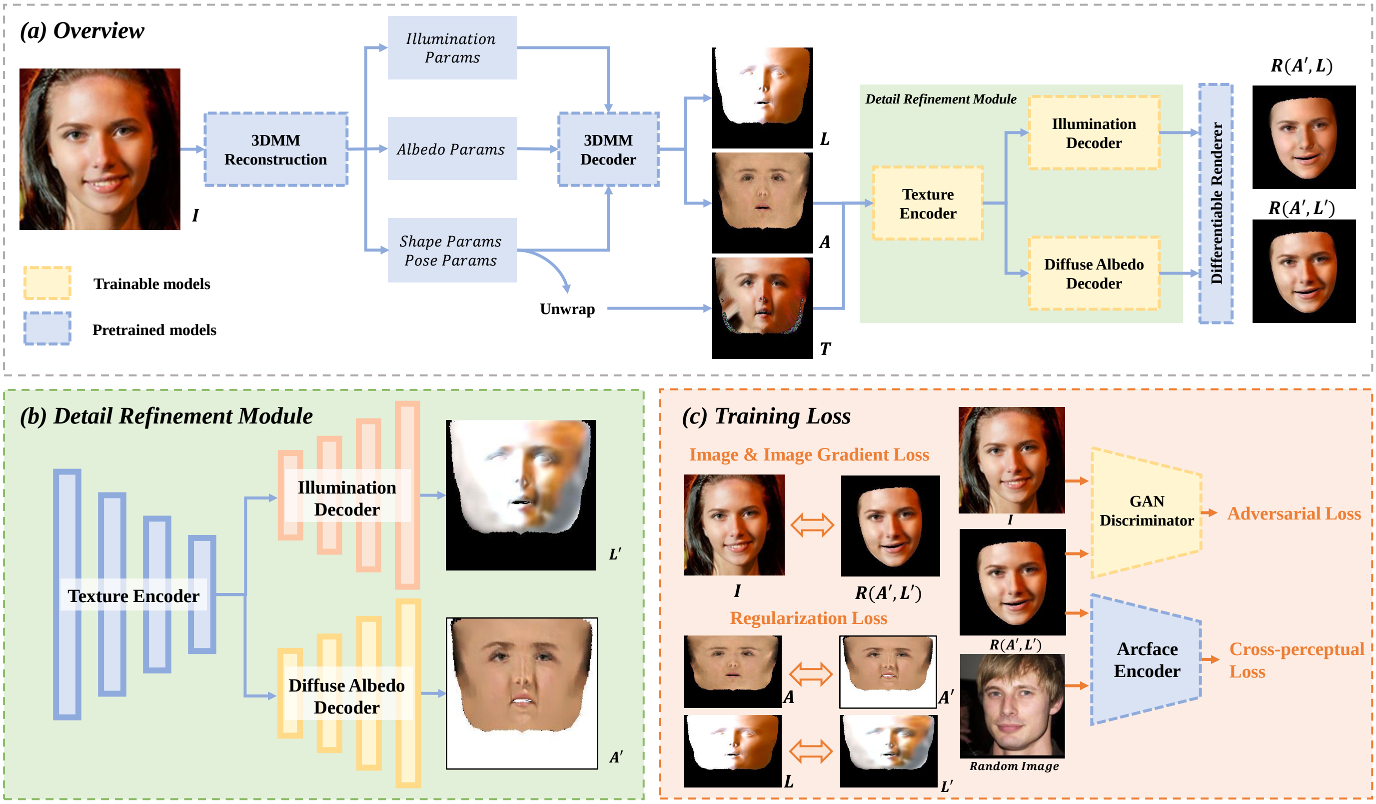

Fig. 2. Network architecture for 3D face reconstruction: The left part of the upper box shows our prior module that takes a 3DMM as proxy model and generates unwrapped facial texture map (T), prior albedo map (A) and prior illumination (L). Our prior module includes a pre-trained 3DMM encoder (which is trained seperately) to regress 3DMM parameters from input image (I) and a fixed 3DMM linear decoder to generate corresponding attributes of the 3D face. The unwrapped facial texture map and prior albedo map are then fed into our detail refinement module (right part of the upper box) to generate the detailed albedo ($A^{'}$) and detailed illumination ($L^{'}$). The detail refinement module (elaborated on the green box) contains one texture encoder with a light decoder and a diffuse albedo decoder. After getting the detailed illumination map and the detailed albedo map, they are sent to a differentiable renderer with other 3D face attributes to obtain re-rendered images ($R(A^{'}, L^{'}), R(A^{'}, L)$) for self-supervision. The main training losses are further illustrated in the orange box.

2.2. 3D Face Reconstruction

Monocular face reconstruction. Zollhöfer et al. [29] give a state-of-the-art report summarizes recent trends in monocular facial reconstruction, tracking, and applications. Given the lack of depth information in RGB images, 3DMM is always included as a proxy model in monocular face reconstruction pipeline. A variety of works [30, 31, 32, 16, 12, 3, 9, 33, 10, 17, 11] utilize this paradigm to transform the reconstruction to a parameter estimation problem. Tian *et al.* [34] and Zollhoefer *et al.* [29] provide comprehensive surveys for face reconstruction approaches. These works can be further divided into two categories by the inferring approaches: fitting-based and learning-based methods. The former provides more accurate reconstruction results but consumes more time, while the latter leverages deep convolutional neural networks to estimate 3DMM parameters leading to a fast inference.

High-fidelity 3D face reconstruction. Although 3DMM can reconstruct 3D faces roughly, it lacks detailed information and certain characteristics such as wrinkles and pores. Recently, many methods were proposed for high-fidelity 3D reconstruction. The direct idea is to capture a dataset with high-quality 3D face ground-truth and train the inference deep network on the dataset [35, 16, 36] to achieve authentic reconstruction results. However, constructing such datasets requires expensive capture equipment (*e.g.,* LED sphere) and leads to laborious work. Meanwhile, the data are mostly captured in a controlled environment and the network trained on it is hard to handle in-the-wild face images. Chen et al. [37] utilize a combination of synthetic and realistic face images, and propose a domain-transfer cGAN to reduce the domain gap between these two kinds of data. Different from this work, we propose a self-supervised framework that depends only on the in-the-wild realistic face dataset.

With the development of CNNs and differentiable rendering, a self-supervised paradigm with re-rendering loss is incorporated into the facial detail reconstruction [38, 39, 17, 15, 14, 13, 40]. These works add a regression network to complement detail information upon the 3DMM coarse reconstruction. However, the coarse model and detail model are often trained separately. These approaches are trained in in-the-wild image datasets and resolve the drawbacks of supervised approaches. [41] leverage the symmetry in the human face and directly regress the depth maps; however, it is unable to reconstruct a complete face model and produce artifacts when encountering extreme inputs (e.g., images with non-frontal faces). Some face texture completion and frontalization approaches [42, 43] may help eliminate the artifacts to some extent. Another work branch focuses on human portrait video (or multi-view images) synthesis [44, 45, 46, 47]. They still utilize the 3DMM as a proxy model and generate high-fidelity dynamic details upon it. However, these works do not produce any detailed 3D mesh model; hence, graphic renderers cannot directly utilize their reconstruction results.

Representation of facial details. 3D facial detail information can be modeled in either geometric space as displacements or normal [43, 48, 49, 13, 16, 14, 50] or in appearance space as texture or albedo [39, 12, 51], or both of them [52]. Considering the representation space, details can be represented as maps in uv-space, or maps in frontal-face space, or vertex attributes on a 3D face mesh. The methods of [16, 13] represent facial detail in uv-space. [16] unwrap partial image texture from image and regress a detailed displacement map from it; [13] directly generate a uv representation of texture and geometry from the input image. [39] utilize graph convolutional network (GCN) and model texture as three-channel vertex attributes on a 3D face mesh to obtain competitive results. [41] take advantage of the symmetry characteristic of human faces and represent the depth, albedo and illumination maps in the frontal-face space. Although applying GCN on mesh could produce convincing results, we argue that uv-representation is still a valid representation for detail reconstruction because the proper face parameterization keeps most of the face topology and can be easily processed by 2D CNN. In this work, we present a monocular high-fidelity 3D face reconstruction approach and represent the detail information by a detailed texture map in the uv-space.

2.3. Image Formation Modeling

Image formation is the process that maps a 3D model with an environmental condition to a 2D image space. The core in the process is the reflectance models that include illumination modeling and interaction pattern between light and the model surface. In 3D face modeling, three-band RGB spherical harmonic lighting representation [53] and Lambertian surface are often considered the default settings [39, 9, 13]. Spherical harmonics is a set of orthonormal basis defined on a sphere that is analogous to Fourier basis in the Euclidean space, and it can be a proper approximation to illumination in the natural lighting. Lambertian surface assumes that the surface irradiance is irrelevant to the observer's position and only depends on the incident light direction. Although these two assumptions provide proper approximation, they neglect other reflection effects (such as specular reflection) in the real scenario and limit the capability to capture complete illumination when encountered with complex environmental light, which is harmful to recover detailed face albedo. This observation motivates us to also refine the reflectance models in our method. Therefore, we propose to retain the Lambertian assumption and attribute all the complex reflectance into our detailed illumination map which is an extension of 3-band spherical harmonics.

3. Overview

Given a single facial image, our goal is to reconstruct a 3D human face, with the emphasis on generating a high-fidelity and re-renderable facial texture that is complete, detailed and disentangled with illumination. To this end, we propose to first generate a prior albedo by a prior generation module and enhance it with the facial texture unwrapped from the image by a detail refinement module, see Fig. 2. We choose a traditional linear 3DMM [54] as our parametric model because a 3DMM albedo excludes most of the illumination. The other parametric models like [12] can also be directly applied in our pipeline.

The input facial image is first fed into the 3DMM encoder to produce 3DMM parameters (including identity, expression, and albedo), pose parameters and illumination parameters. Then, these regressed 3DMM parameters are passed to a fixed 3DMM decoder to acquire the prior albedo, 3DMM shape, camera pose matrix and coarse illumination. Next, we obtain the facial texture by unwrapping the input image according to the projected 3DMM shape. The unwrapped facial texture map and prior albedo map are fed into the detail refinement module, which is composed of a modified version of image-to-image translation network, to generate a detailed albedo map and a detailed illumination map. Finally, the detailed albedo and the detailed illumination combined with the 3DMM shape projected in camera space are rendered to the image space by a differentiable renderer.

4. Prior Albedo Generation Module

Our prior albedo generation module takes a 3DMM as proxy model and uses a convolutional neural network to estimate the parameters of facial geometry, albedo, pose and illumination. We adopt the state-of-the-art 3DMM coefficient regressor [9] for the purpose.

Next, we can derive the prior albedo by 3DMM albedo decoding and generate an image texture by sampling the corresponding projection pixels with 3DMM shape and pose parameters. After these two textures are obtained, they are projected onto 2D uv-space to accommodate with 2D CNN structure of detail refinement module. However, two problems occurred in the image texture sampling procedure. First, the non-frontal face and occlusion problem in the facial images may cause incompleteness in the unwrapped texture. Second, the inaccurate regression of pose and shape parameters in many cases may cause the image texture to generate artifacts, especially in the edge parts. Our detail refinement module introduced in the following can resolve these two problems.

5. Detail Refinement Module

Our detail refinement module adopts an image-to-image translation network in the uv-space where the input includes two maps: a prior albedo map and a partial facial image texture. We first pad the unseen parts with Gaussian noises as being carried out in [51] because filling the 'holes' in the unwrapped image texture is one of the goals of our detail refinement module.

Fig. 3. Motivation of the cross perceptual loss. From left to right are image of identity A, image of identity B, the rendered image constructed by detailed albedo of A in combination with detailed illumination of B without cross perceptual loss, the rendered image constructed by detailed albedo of A in combination with detailed illumination of B with cross perceptual loss.

Then, these two maps are concatenated and fed into the refinement network which also produces two outputs: the detailed albedo map and the detailed illumination map. The detailed albedo map includes information about basic details in human faces, such as facial wrinkles, pores and etc. Meanwhile the detailed illumination map attempts to model spatially-complex environmental illumination that can not be captured in a previous coarse reconstruction. In the following, we elaborate on each part of the module and explain the specifically-devised loss functions.

5.1. Illumination Disentanglement

Illumination regularization. With respect to the illumination modeling, directly utilizing the coarse illumination in the prior generation module does not fully capture the complex illumination for in-the-wild images. This situation will lead to the leakage of light information to the albedo which makes it not rerenderable. Therefore, a detailed representation for illumination is needed.

Given that our framework is trained in a self-supervised way, disentangling the illumination with albedo is not trivial. We take advantage of the coarse illumination spherical harmonics generated by the prior module and develop our illumination representation from it. We introduce a novel representation in the uv-space called spherical harmonics map which models a spherical harmonics illumination for every vertex in the face model. With the illumination map, we could not only model complex illumination in facial images, but also disentangle light by minimizing the distance with coarse spherical harmonics, which is named by illumination regularization loss. In particular, we represent the detailed illumination map $L_{detail} \in \mathbb{R}^{(B,27,H,W)}$ in the uv-space and directly regress it from the detail refinement module. Then, we regularize local illumination by devising a mean square error (MSE) loss to penalize detail and coarse illumination differences. The MSE loss is expressed as follows:

$$L_{reg-illu} = ||M_{uv} * (L_{detail} - L_{coarse})||^2, \qquad (1)$$

where $L_{coarse} \in \mathbb{R}^{(B,27,H,W)}$ is the coarse SH-illumination vector expanding to uv map size, and $M_{uv} \in \mathbb{R}^{(B,1,H,W)}$ stands for the facial regions visible to the camera projected onto the uv-space.

Cross perceptual loss. Despite utilizing the illumination regularization loss mentioned above, we find in our experiments that a small amount of facial internal characteristics (such as wrinkles and beard) are mistakenly included in the illumination map, which means that albedo and illumination are not completely disentangled, leading to the loss of details in our detailed albedo map. Given that the attributes, such as wrinkles and beard are individual specific, we propose to utilize a cross-identity perceptual loss to conduct the further disentanglement. The motivation is stated below. We assume that the re-rendered image that combines person A's detailed illumination map and person B's detailed albedo map should have the same identity with person B in that the correctly-disentangled detailed illumination map would only include environmental illumination information in it. However, if the detailed illumination map is not entirely disentangled, which refers to including the facial attribute specific to person A, then it may change the identity of the rendered image. These two samples can be seen from Fig. 3. Owing to this observation, we utilize an illumination-irrelevant facial recognition network[55] to distinguish whether the two images are of the same identity. The cross perceptual loss is represented as follows:

$$L_{cross-percp} = 1 - < ArcFace(I_r), ArcFace(I_{gt}) >, \qquad (2)$$

where $ArcFace$ stands for the perceptual net, I_r represents the rendered image with detailed illumination map A and detailed albedo map B, and I_{gt} means the ground-truth facial image of B. We adopt the cosine distance as the measurement of the similarity between two normalized facial feature vectors.

We sum these two losses together with corresponding weights as our final illumination disentanglement loss:

$$L_{id} = \lambda_{id1} * L_{reg-illu} + \lambda_{id2} * L_{cross-percp}. \qquad (3)$$

5.2. Albedo Regularization

We utilize prior albedo and the intrinsic characteristics of albedo map to construct several regularization losses for obtaining a complete and re-renderable albedo from the input image. Here, we propose three losses : the symmetry loss, albedo smooth loss, and conditional GAN losses.

(1) Symmetry loss: We propose a symmetry loss on the detailed albedo map. Given that human facial albedos are mostly symmetrical (especially when decoupled with light), we use this loss to regularize an unseen texture problem induced by the non-frontal face in the input image. Another advantage of the albedo symmetry loss is that it ensures robust albedo reconstruction in uneven scene illumination. The symmetry loss is expressed as follows:

$$L_{symm} = ||M_{uv} * (A_{detail} - \hat{A_{detail}})||^2, \qquad (4)$$

where $\hat{A_{detail}}$ is the detailed albedo map flipped along the y-axis.

(2) Albedo smooth loss: We propose a smooth loss to regularize the detailed albedo map. We expect the detailed albedo map inherits this feature because the generated prior albedo is decoupled with illumination. We utilize local weighted smooth loss on the detailed albedo map to achieve this goal. To compute the local weights, we assume the detailed albedo map shares the same smoothness with the prior albedo map. Therefore, we use the local difference between pixels in the prior albedo map to

compute the smoothness weights of the detailed albedo map. The albedo smooth loss is defined as:

$$L_{smooth} = \sum_{i} \sum_{j \in N(i)} \omega_{i,j} ||A_{detail}(i) - A_{detail}(j)||^2, \tag{5}$$

$$\omega_{i,j} = exp(-\alpha * ||A_{prior}(i) - A_{prior}(j)||^2), \tag{6}$$

where A_{detail} and A_{prior} represent the detailed albedo map and the prior albedo map respectively, $N(i)$ indicates the neighbors of texel (pixel in uv-space) $p(i)$, $\omega_{i,j}$ represents the similarity of two texels, which is measured by a decreasing function of corresponding texels' difference in the prior albedo map. α is a super-parameter which we here choose 80 empirically. In the above equation, our albedo smooth loss penalizes more to those texels whose neighborhood difference shares less similarity between the detailed albedo map and the prior albedo map.

(3) GAN loss: Our devised GAN loss includes an L_1 distance loss and an adversarial loss [56] to force the detailed albedo to share the same distribution as prior albedo map. The L_1 distance loss can be written as:

$$L_{L1} = |M_{uv} * (A_{detail} - A_{prior})|. \tag{7}$$

We then define the adversarial loss as:

$$\begin{aligned} L_{GAN_D} =& E_{G(z) \in A_{detail}} log(1 - D(G(z))) + & (8)\\ & E_{x \in A_{prior}} logD(x), & \\ L_{GAN_G} =& E_{G(z) \in A_{detail}} log(D(G(z))). & (9) \end{aligned}$$

where D symbolizes the discriminator to judge whether the generated albedo map falls on the support set of the prior albedo map distribution. $G(z)$ represents the detailed albedo map generator which means the whole framework.

Our albedo regularization loss is then computed by combining above four loss terms with proper weights:

$$\begin{aligned} L_{ar} =& \lambda_{ar1} * L_{symm} + \lambda_{ar2} * L_{smooth} + & (10)\\ & \lambda_{ar3} * L_{L1} + \lambda_{ar4} * L_{GAN_G} & (11) \end{aligned}$$

5.3. Detail Preservation

Besides above regularization losses, we also utilize basic reconstruction losses to facilitate high-fidelity reconstruction. These losses are all applied on the image space; thus a face mask is required for concentrating penalization of the differences on face regions in the images. We adopt the face parsing approach [57] to generate face masks before training. Coarse reconstruction may not be well-suited to image mask because of the inaccurate estimation of 3DMM and camera pose. Hence, we generate our final face mask by multiplying a pre-generated mask with projected face mask. The final face mask can be computed as:

$$M_{face} = M_{parsing} * M_{proj}. \tag{12}$$

After face masks are obtained, we propose two reconstruction losses applied in the mask regions, which contain image gradient loss and image loss.

(1) Image gradient loss: We now propose an image gradient loss to encourage the similarity between the re-rendered facial image gradient and the ground-truth facial image gradient for reconstructing facial details as authentic as possible. This loss is designed according to the assumption that the detail information can be mostly captured by image gradient. We define such gradient loss function as:

$$L_{grad} = \sum ||M_{face} * (Grad(I_r) - Grad(I_{gt}))||^2, \tag{13}$$

where M_{face} is the pre-extracted face mask, I_r and I_{gt} are the rendered image and the input facial image, respectively; and $Grad$ represents the gradient operator. Specifically, we first calculate two directional gradients along the x-axis and y-axis, then compare them with corresponding ground-truth gradient maps. Finally, we obtain the summation.

(2) Image loss: We also adopt image loss to penalize pixel difference between the rendered image and the input facial image, which can be expressed as follows:

$$L_{img} = \sum ||M_{face} * (I_r - I_{gt})||^2. \tag{14}$$

We combine these two losses together to obtain our detail preservation loss:

$$L_{dp} = \lambda_{dp1} * L_{grad} + \lambda_{dp2} * L_{img}. \tag{15}$$

6. Network Architecture and Training Details

We train our neural network on a public dataset CelebA [58] which is a large-scale facial attribute dataset that has more than 200K facial images collected from the internet. We separate the dataset into disjoint training data (85%) and testing data (15%). We pre-process the images by first generating 68-landmarks [59] before feeding them into our detail generation network. Then, we utilize the generated landmarks to crop and scale the images to keep the human faces staying in the center of the images and resize them to 224×224. After pre-processing, these images are fed into our pre-trained prior albedo generation module. In this work, a 3DMM parameter regressing network and a fixed 3DMM decoder [54] are utilized to obtain the prior albedo and other attributes (geometry, camera and illumination). The camera parameters and 3DMM shape are leveraged to unwrap the texture from the input image. Next, the unwrapped texture and prior albedo (in uv-space) are concatenated and fed into the detail reconstruction network to acquire detailed albedo maps and detailed illumination maps.

We adopt the ResNet-50 [60] as the backbone network of our prior albedo generation module and pre-train it on 300W-LP [61] following the state-of-art 3DMM reconstruction work [9]. 300W-LP is a dataset that contains 122,450 facial images with a variety of head poses generated from the original 300W dataset by face profiling techniques. Similar to the training process in [9], we train the network in a self-supervised way with pixel-level, landmark-level, and perceptual-level discrepancy in combination with the parameter regularization loss.

In the detail reconstruction network, we adopt the basic pix2pix network [56] as our backbone and abandon the skip-connection because it may cause the output to inherit the noise from the unwrapped image texture. We also extend pix2pix with two decoders, one for detailed albedo generation and the other for detailed illumination modeling. The light decoder and albedo decoder share the same structure but with different output layers, where the light decoder outputs a 27-channel map while the albedo decoder outputs a 3-channel one. Finally, We use the face mesh renderer [62] for differentiable rendering. We combine the loss functions mentioned above to train our network, which is expressed as follows:

$$L_{total} = L_{id} + L_{ar} + L_{dp}. \tag{16}$$

The coefficients in those loss functions are chosen as λ_{id1}, λ_{id2}, λ_{ar1}, λ_{ar2}, λ_{ar3}, λ_{ar4}, λ_{dp1}, λ_{dp2} : 1.0, 0.5, 5.0, 5.0, 1, 1.0, 0.001, 1.0, 5.0, where λ_{ar3} is set as 0.0 after first epoch training.

We trained our entire network end-to-end for 10 epochs using the Adam optimizer. The initial learning rate was set to 10^{-4} and reduced with attenuation coefficient of 0.98 every 1 epochs until we reached 10^{-5} to avoid overfitting. The batch size was 16 and momentum was 0.9. The training task was completed in 2 days on a workstation with one Nvidia RTX-2080 TI GPU. Once trained, our network can process approximately 30 images per second in the inference stage.

7. Experimental Results

We evaluate our algorithm qualitatively and quantitatively by performing a complete comparison with current state-of-the-art approaches. We further conduct ablation studies to provide a comprehensive evaluation of the individual components of our neural network.

7.1. Self Evaluation

Performance on CelebAHQ dataset. We first analyze the capability of our method by using the CelebAHQ database [63], which includes 30k 1024 × 1024 facial images generated by applying super-resolution algorithm to a subset of CelebA images. Given that the images in CelebAHQ contain more details, we test our neural network on it to demonstrate whether our approach could capture details on high-definition images and achieve high-quality reconstructed albedos. Note that our neural network is only trained on the original CelebA.

Fig. 4 qualitatively shows our reconstruction results on several images randomly selected from CelebAHQ. Our proposed approach successfully keeps facial details in the reconstructed detailed albedo map, and there exists no reflection effects other than diffuse reflection in the albedo map or rendered results with coarse illumination. This phenomenon verifies that most of the environment illumination are explained by the detailed illumination, which leads to a clean diffuse facial albedo. In addition, our albedo regularization loss ensures that the detailed albedo map also exhibits smoothness and completeness which are beneficial to re-render applications. The detailed albedo generated by our network can be directly sent into a renderer with Lambertian reflector to achieve high-fidelity re-rendered results owing to these two characteristics.ior 3DMM albedo by a white map (so it includes no prior information) to evaluate the effect of the inputs.

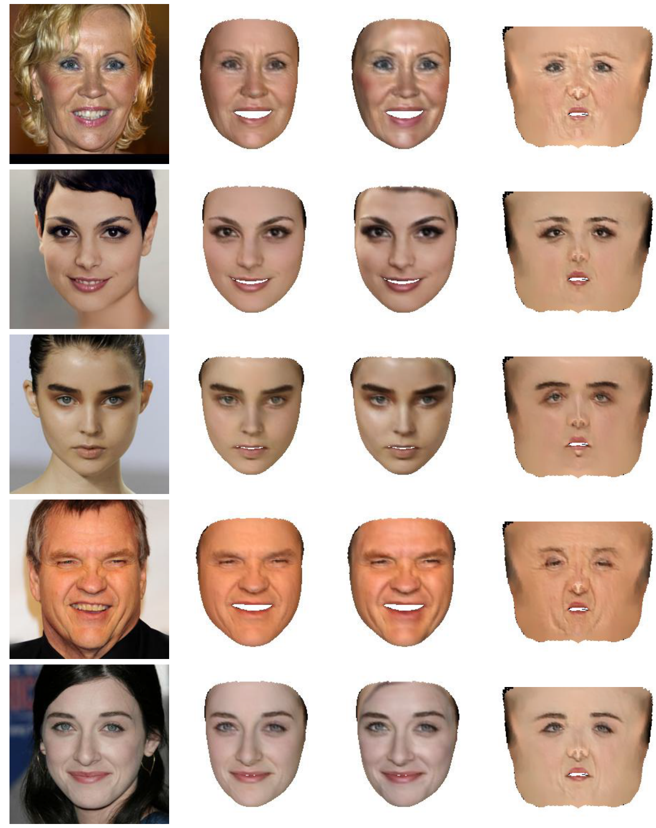

Fig. 4. Our reconstruction results on CelebAHQ. The first column is input, and the second and third columns show the results generated by our detailed albedo combined with coarse illumination and detailed illumination respectively. The last column shows the detailed albedo in the uv-space.

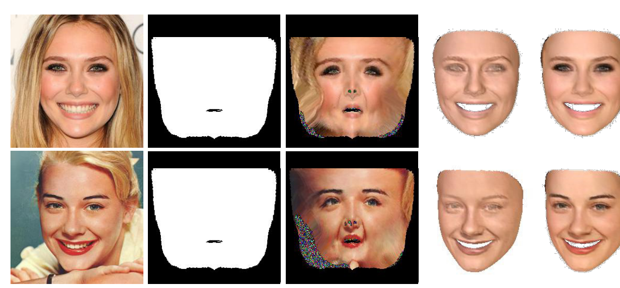

Fig. 5. Evaluation of prior albedo and detail refinement module. The first three columns are input images, unwrapped image textures, and white facial albedo which replaces the coarse albedo. The right two columns show the output rendered images by using white albedo and original coarse albedo. The figure verifies the capability of our model to transfer as much details from the input texture.

Evaluation of prior albedo and detail refinement module. Given that the detail refinement module in our framework takes two inputs, namely the prior albedo generated by 3DMM and the unwrapped texture from the input image, we are interested in exploring what these two inputs are responsible for in the detail refinement module. Accordingly we conduct an experiment by substituting the input prior 3DMM albedo by a white map (so it includes no prior information) to evaluate the effect of the inputs. Fig. 5 shows the experimental results, where the ren-

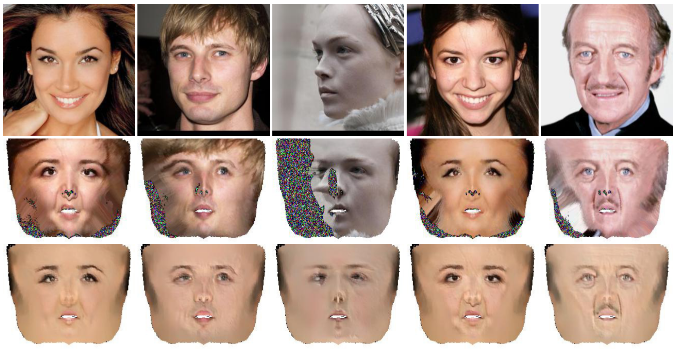

Fig. 6. Effectiveness of eliminating artifacts in unwrapped image texture. From top to bottom are original input facial images, the unwrapped textures from input images and detailed albedos reconstructed by our method.

dered image with white map absorbs most details in the input but loses the appearance consistency with the original image. This phenomenon indicates that our detail refinement module takes these two inputs independently, where the prior albedo guides entire appearance generation and the unwrapped texture is responsible for the detail supplement.

Texture artifact removal. We now evaluate the capability of our detail refinement module in dealing with the textures containing artifacts. The two main artifacts that exist in facial images are show in the second row of Fig. 6. First, the non-frontal face images would lead to incompleteness in the unwrapped image texture. Second, the geometry parameters (including camera pose and shape parameters) regressed from the coarse reconstruction step are not accurate in many cases, which would result in severe stripe-like artifacts in the unwrapped texture, especially in the boundaries of human face. Our detail refinement module can remove these two kinds of artifacts and produce a smooth, complete and high-fidelity albedo (Fig. 6). This phenomenon is mainly due to the introduction of prior albedo and our designed albedo regularization loss, which endow the final reconstructed albedo with completeness and smoothness.

7.2. Qualitative Comparison

For qualitative comparison, we first compare our approach against recent learning-based texture reconstruction and generation methods [19, 12, 10, 39]. Then, we focus on the qualitative evaluation in extreme illumination condition and compare our reconstructed albedo with advanced facial texture generation method [12]. Finally, we compare our albedo reconstruction performance with [36] which shares similar goal with ours whereas utilizing a self-collected high-fidelity dataset.

Comparison on MOFA data. Fig. 7 illustrates the comparison results with state-of-the-art reconstruction and generation works on a subset of MOFA test dataset [11]. Han *et al.* [19] proposes to utilize low-cost publicly-available data to construct a full 3D face texture space containing not only diffuse but also spatially-varying specular materials. As a generation-based model, [12] capture 10,000 high-resolution human facial textures in the uv-space and train a progressive growing GAN to model the distribution human face texture. They leverage this progressive GAN as the generative model and utilize fitting-based paradigm to estimate the parameters in latent space.

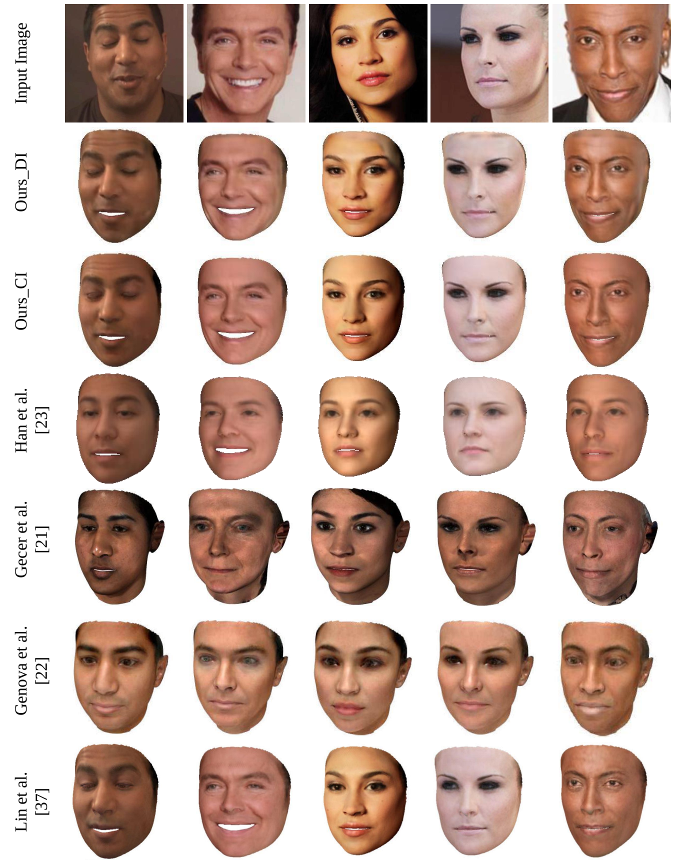

Fig. 7. Qualitative comparison to other competitive methods. The first row is the input images while the remaining rows show the reconstruction results of all methods. Our reconstructions are shown in the second and third rows where Ours_DI stands for reconstruction with detailed illumination and Ours_CI means reconstruction with SH illumination (including only details in diffuse albedo).

Deng *et al.* [9] and Genova *et al.* [10] are the two representative 3DMM-based facial reconstruction methods that are also trained in the self-supervised way on in-the-wild facial image datasets. Lin *et al.* [39] aims to reconstruct high-fidelity facial texture from a single image self-supervised using GCN in a coarse-to-fine manner. As illustrated in the Fig. 7, not only our reconstruction achieves the best detail preservation compared with other competitive approaches (see the second row, Ours_DI), but our diffuse albedo also decouples the environmental illumination and shadows (see the third row, Ours_CI).

Comparison on extreme illumination data. Facial images under extreme illumination condition, including uneven lighting or shadow, are commonly encountered in real-world applications. Due to the loss of information and low quality representation of illumination (three-band Spherical Harmonics), reconstructing high-fidelity 3D face under such circumstances is still challenging thus reconstruction-based methods always fail to reconstruct complete face albedos. Meanwhile, a generation-based approach can deal with extreme lighting because they map the facial texture space to a latent space with a support

Fig. 8. Comparison to Ganfit [12] in extreme lighting. From left to right are input images, reconstructed albedo of [12], our generated albedo, and our re-illuminated results using a different illumination condition, where the used environmental map is illustrated at the left-bottom.

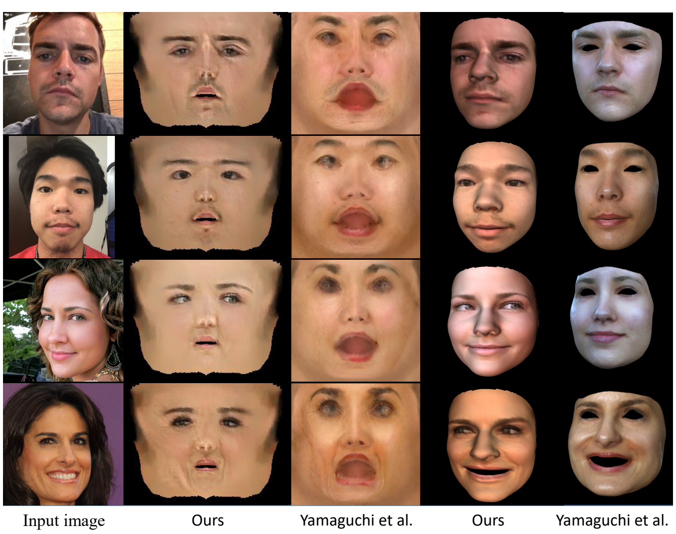

Fig. 9. Comparison to [36] for albedo reconstruction. The first column shows input images; the second and third columns are the reconstructed albedos of our approach and [36]; the last two columns display the rendered images with different viewpoints and illumination conditions.

Table 1. Quantitative comparison on CelebA dataset. Ours_OD means the rendered image constructed by our detailed albedo combined with our detailed illumination, while Ours_OC means the result by using coarse illumination. The best result of each measurement is marked in bold font. Symbol '/' means that we could not test the corresponding method since no open-source implementation.

Methods	$L_1 \downarrow$	$PNSR \uparrow$	$SSIM \uparrow$	$LightCNN \uparrow$	$evoLVe \uparrow$
Deng et al. [9]	0.05	26.58	0.83	0.72	0.64
Gercer et al. [12]	/	26.5	0.898	/	/
Lin et al. [39]	0.034	**29.69**	0.89	0.90	0.85
Dib et al. [38]	0.032	28.72	0.807	/	/
Ours_OC	0.02	24.88	0.89	0.91	0.83
Ours_OD	**0.01**	28.90	**0.93**	**0.93**	**0.86**

set. Our proposed model merging these two methods together should also have the capability to reconstruct convincing face albedos from the facial images under extreme illumination. In Fig. 8, we compare our reconstruction results in extreme lighting to Ganfit [12]. The second and third columns show that our albedo decouples the complex environmental illumination and outperforms [12] by preserving more facial details from the input face. This phenomenon is because of the cooperation between the prior albedo generation module with the detail refinement module of our framework. The former module is responsible for generating the guided albedo and the latter can complement details upon it. As a result, our method not only inherits the advantage of generation-based methods which maintain the diffuse texture smooth in the whole but also has the ability to preserve the details as in a reconstruction-based method. We outperform Ganfit and achieve more convincing results. The fourth column shows re-illuminated results according to our reconstructed albedo, where we apply different illumination conditions to the reconstructed albedo and render it to images where the illumination is randomly selected from a face illumination prior database [64]. The re-illuminated results are rather realistic and keeps the identity information of the original image.

Comparison on albedo reconstruction from a single image. To evaluate the quality of our reconstructed (diffuse) albedo, we compare with the state-of-the-art method [36]. As shown in Fig. 9, both [36] and our approach can decouple environmental illumination well. However, thanks to the novel illumination representation and disentanglement loss, our reconstructed albedos keep more performers' idiosyncrasy than [36], which can be observed from the nasolabial folds from the third and fourth rows.

7.3. *Quantitative Comparison*

For quantitative comparison, we mainly focus on the criteria for measuring the image-level difference. First, L1 distance loss is applied as the basic pixel-level criterion. Then, we utilize two commonly-used image similarity criteria, namely the structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR), to evaluate the similarity between the rendered face image and original input face image. With regard to the human face problem, we also leverage two well-known pretrained face recognition networks as maps from image space to feature space and evaluate the difference between rendered face image and input face image in the facial feature space. The two facial recognition networks we adopted are *LightCNN* [65] and *evoLVe* [66], since their state-of-the-art performance and widely acceptance [39]. In summary, we calculate the difference between two face images in both pixel-level (including L1 distance loss, PSNR and SSIM) and face feature-level (including LightCNN and evoLVe).

Fig. 10. Ablation study of the proposed gradient loss and texture regularization loss: our full model produces the most convincing results than others.

The numerical statistics for each method are reported in Table 1, where the competing methods we choose are state-of-the-art ones trying to reconstruct details in albedo. The table illustrates that our reconstruction results with detailed illumination are better than those of the competing algorithms. Besides, our framework also achieves competitive results by using only detailed albedo combined with coarse illumination, which further demonstrates that our detailed albedo is able to capture most facial details in the input image.

7.4. Ablation Study

Effectiveness of gradient and texture regularization losses. We first demonstrate the functionality of the gradient loss and texture regularization loss in our pipeline using the detailed rendering results with coarse illumination and detailed albedo. As shown in Fig. 10, our proposed L_{grad} helps our model to capture the detailed information from the facial image. Our $L_{tex-reg}$ loss contributes to the disentanglement of illumination and completes the occlusion part according to the prior albedo which renders the detailed albedo map more similar to the prior albedo map. By contrast, our full model produces the most convincing results than others.

Effectiveness of light regularization. To evaluate the effect of our light perceptual regularization loss, we perform an ablation study by showing the rendered detailed illumination images with and without this loss. In Fig. 11, the light perceptual regularization loss helps the disentanglement of illumination with facial characteristics. The illumination map recovered with the help of light perceptual regularization loss includes less facial wrinkles and beard than the one that recovered without this loss. This phenomenon indicates that the facial details are all mostly encoded in the detailed albedo map. Our illumination map has only environmental light information as far as possible thus, it is more suitable for re-renderable 3D facial generation.

7.5. Limitations

Although our model achieves competitive results on most of the in-the-wild facial image datasets, it may still generate unreliable results on huge occlusion cases. The reason is that the prior model we used is only able to produce low-fidelity prior albedo and the lack of information in significantly occluded regions cannot be complemented with only symmetry regularization. Moreover, though our model takes 3DMM reconstructed albedo as prior, our reconstructed albedo is not completely independent of input image quality and generate better result when inputs are of high resolution.

Fig. 11. Ablation study of the proposed perceptual loss for lighting disentanglement: the illumination image generated by our full model separates the facial intrinsic details better.

8. Conclusion and Future Work

We have presented a novel self-supervised neural network for 3D face reconstruction, emphasizing generating re-renderable high-fidelity textures from single images. We utilize the coarse 3DMM model as a prior and fine-tune on it to capture more facial details. We compare our results with state-of-the-art methods in qualitative and quantitative ways. The comparison demonstrates that our method does not require capturing high-resolution face texture datasets and we can generate re-renderable and realistic facial textures.

However, our approach still falls short in fully addressing face reconstruction under extreme conditions, such as top-down viewpoints or exaggerated facial expressions. This limitation is primarily due to the constrained representative capacity of the 3DMM albedo prior or the potential for inaccurate geometry reconstruction by the 3DMM when dealing with exaggerated facial expressions. In the future, we plan to construct a high-fidelity albedo map dataset and train a new generation model, which would significantly improve the reconstruction quality under extreme conditions. Second, we are interested in extending our model to reconstruct geometric details, because high-fidelity geometry and texture would lead to a more competitive and visually appealing result. Finally, we would like to add more dynamics to our model and reconstruct animated facial details from a single image or video.

Acknowledgments

We thank the reviewers for their constructive comments. This work was supported in parts by the National Key R&D Program of China (2022ZD0160801), NSFC (U21A20515, 62262043, 62172416), Guangdong Basic and Applied Basic Research Foundation (2023B1515120026, 2023B0303000016), Shenzhen Science and Technology Program (GJHZ20210705141402008), and Youth Innovation Promotion Association of the Chinese Academy of Sciences (2022131).

References

[1] Zhou, H, Liu, J, Liu, Z, Liu, Y, Wang, X. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2020, p. 5911–5920.

[2] Owusu, E, Appati, JK, Okae, P. Robust facial expression recognition system in higher poses. Visual Computing for Industry, Biomedicine, and Art 2022;5(1):14.

[3] Thies, J, Zollhofer, M, Stamminger, M, Theobalt, C, Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2016, p. 2387–2395.

[4] Wang, Y, Guo, J, Yan, DM, Wang, K, Zhang, X. A robust local spectral descriptor for matching non-rigid shapes with incompatible shape structures. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2019, p. 6231–6240.

[5] Guo, J, Wang, H, Cheng, Z, Zhang, X, Yan, DM. Learning local shape descriptors for computing non-rigid dense correspondence. Computational Visual Media 2020;6:95–112.

[6] Lattas, A, Moschoglou, S, Gecer, B, Ploumpis, S, Triantafyllou, V, Ghosh, A, et al. Avatarme: Realistically renderable 3d facial reconstruction. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2020, p. 760–769.

[7] Hu, L, Saito, S, Wei, L, Nagano, K, Seo, J, Fursund, J, et al. Avatar digitization from a single image for real-time rendering. ACM Trans Graph 2017;:1–14.

[8] Blanz, V, Vetter, T, Rockwood, A. A morphable model for the synthesis of 3d faces. ACM Trans Graph (Proc SIGGRAPH) 2002;:187–194.

[9] Deng, Y, Yang, J, Xu, S, Chen, D, Jia, Y, Tong, X. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019, p. 0–0.

[10] Genova, K, Cole, F, Maschinot, A, Sarna, A, Vlasic, D, Freeman, WT. Unsupervised training for 3d morphable model regression. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2018, p. 8377–8386.

[11] Tewari, A, Zollhofer, M, Kim, H, Garrido, P, Bernard, F, Perez, P, et al. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: IEEE International Conference on Computer Vision Workshops. 2017, p. 1274–1283.

[12] Gecer, B, Ploumpis, S, Kotsia, I, Zafeiriou, S. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2019, p. 1155–1164.

[13] Tran, L, Liu, F, Liu, X. Towards high-fidelity nonlinear 3d face morphable model. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2019, p. 1126–1135.

[14] Tran, L, Liu, X. Nonlinear 3d face morphable model. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2018, p. 7346–7355.

[15] Chen, Y, Wu, F, Wang, Z, Song, Y, Ling, Y, Bao, L. Self-supervised learning of detailed 3d face reconstruction. IEEE Trans Image Process 2020;:8696–8705.

[16] Chen, A, Chen, Z, Zhang, G, Mitchell, K, Yu, J. Photo-realistic facial details synthesis from single image. In: IEEE International Conference on Computer Vision (ICCV). 2019, p. 9429–9439.

[17] Tewari, A, Zollhöfer, M, Garrido, P, Bernard, F, Kim, H, Pérez, P, et al. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2018, p. 2549–2559.

[18] Lattas, A, Moschoglou, S, Ploumpis, S, Gecer, B, Ghosh, A, Zafeiriou, S. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. IEEE Trans Pattern Anal Mach Intell 2021;44(12):9269–9284.

[19] Han, Y, Wang, Z, Xu, F. Learning a 3d morphable face reflectance model from low-cost data. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2023,.

[20] Towards Metrical Reconstruction of Human Faces. 2022.

[21] Booth, J, Roussos, A, Ponniah, A, Dunaway, D, Zafeiriou, S. Large scale 3d morphable models. Int Journal of Computer Vision 2018;:233–254.

[22] Li, T, Bolkart, T, Black, JM, Li, H, Romero, J. Learning a model of facial shape and expression from 4d scans. ACM Trans Graph 2017;:194:1–194:17.

[23] Cao, C, Weng, Y, Zhou, S, Tong, Y, Zhou, K. Facewarehouse: A 3d facial expression database for visual computing. IEEE Trans Vis Comput Graph 2013;:413–425.

[24] Egger, B, Smith, WAP, Tewari, A, Wuhrer, S, Zollhoefer, M, Beeler, T, et al. 3d morphable face models - past, present and future. ACM Trans Graph 2020;.

[25] Hong, Y, Peng, B, Xiao, H, Liu, L, Zhang, J. Headnerf: A real-time nerf-based parametric head model. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2022, p. 20374–20384.

[26] Bao, L, Lin, X, Chen, Y, Zhang, H, Wang, S, Zhe, X, et al. High-fidelity 3d digital human head creation from rgb-d selfies. ACM Trans Graph 2021;.

[27] Li, T, Bolkart, T, Black, MJ, Li, H, Romero, J. Learning a model of facial shape and expression from 4D scans. ACM Trans Graph (Proc SIGGRAPH Asia) 2017;36(6):194:1–194:17.

[28] Karras, T, Aila, T, Laine, S, Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:171010196 2017;.

[29] Zollhöfer, M, Thies, J, Garrido, P, Bradley, D, Beeler, T, Pérez, P, et al. State of the art on monocular 3d face reconstruction, tracking, and applications. Comput Graph Forum 2018;37(2):523–550.

[30] Wood, E, Baltrušaitis, T, Hewitt, C, Johnson, M, Shen, J, Milosavljević, N, et al. 3d face reconstruction with dense landmarks. In: European Conference on Computer Vision (ECCV). Springer; 2022, p. 160–177.

[31] Feng, Y, Feng, H, Black, MJ, Bolkart, T. Learning an animatable detailed 3d face model from in-the-wild images. ACM Trans Graph 2021;40(4):1–13.

[32] Li, C, Morel-Forster, A, Vetter, T, Egger, B, Kortylewski, A. To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. arXiv preprint arXiv:210609614 2021;.

[33] Yi, H, Li, C, Cao, Q, Shen, X, Li, S, Wang, G, et al. Mmface: A multi-metric regression network for unconstrained face reconstruction. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2019, p. 7663–7672.

[34] Tian, Y, Zhang, H, Liu, Y, Wang, L. Recovering 3d human mesh from monocular images: A survey. IEEE Trans Pattern Anal Mach Intell 2023;.

[35] Yang, H, Zhu, H, Wang, Y, Huang, M, Shen, Q, Yang, R, et al. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2020, p. 601–610.

[36] Yamaguchi, S, Saito, S, Nagano, K, Zhao, Y, Chen, W, Olszewski, K, et al. High-fidelity facial reflectance and geometry inference from an unconstrained image. ACM Trans Graph 2018;:162:1–162:14.

[37] Chen, Z, Wang, Y, Guan, T, Xu, L, Liu, W. Transformer-based 3d face reconstruction with end-to-end shape-preserved domain transfer. IEEE Trans Circuit Syst Video Technol 2022;32(12):8383–8393.

[38] Dib, A, Thebault, C, Ahn, J, Gosselin, PH, Theobalt, C, Chevallier, L. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In: IEEE International Conference on Computer Vision (ICCV). 2021, p. 12819–12829.

[39] Lin, J, Yuan, Y, Shao, T, Zhou, K. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2020, p. 5891–5900.

[40] Yang, M, Guo, J, Ye, J, Zhang, X. Detailed 3d face reconstruction from single images via self-supervised attribute learning. In: SIGGRAPH Asia 2020 Posters. 2020, p. 1–2.

[41] Wu, S, Rupprecht, C, Vedaldi, A. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2020, p. 1–10.

[42] Zeng, X, Wu, Z, Peng, X, Qiao, Y. Joint 3d facial shape reconstruction and texture completion from a single image. Computational Visual Media

2022;8:239–256.
[43] Kim, J, Yang, J, Tong, X. Learning high-fidelity face texture completion without complete face texture. In: IEEE International Conference on Computer Vision (ICCV). 2021, p. 13990–13999.
[44] Zheng, M, Haiyu, Z, Yang, H, Huang, D. Neuface: Realistic 3d neural face rendering from multi-view images. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2023,.
[45] Gafni, G, Thies, J, Zollhöfer, M, Niessner, M. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2021,.
[46] Thies, J, Zollhöfer, M, Nießner, M. Deferred neural rendering: image synthesis using neural textures. ACM Trans Graph 2019;:1–12.
[47] Kim, H, Garrido, P, Tewari, A, Xu, W, Thies, J, Niessner, M, et al. Deep video portraits. ACM Trans Graph 2018;:1–14.
[48] Chen, Z, Yin, K, Fidler, S. Auv-net: Learning aligned uv maps for texture transfer and synthesis. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2022, p. 1465–1474.
[49] Feng, Y, Feng, H, Black, MJ, Bolkart, T. Learning an animatable detailed 3D face model from in-the-wild images. ACM Trans Graph 2021;40(8).
[50] Wang, X, Guo, Y, Deng, B, Zhang, J. Lightweight photometric stereo for facial details recovery. IEEE Computer Vision and Pattern Recognition (CVPR) 2020;:737–746.
[51] Deng, J, Cheng, S, Xue, N, Zhou, Y, Zafeiriou, S. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. IEEE Computer Vision and Pattern Recognition (CVPR) 2018;:7093–7102.
[52] Chai, X, Chen, J, Liang, C, Xu, D, Lin, CW. Expression-aware face reconstruction via a dual-stream network. IEEE Trans Multimedia 2021;23:2998–3012. doi:10.1109/TMM.2021.3068567.
[53] Zhang, L, Samaras, D. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. IEEE Trans Pattern Anal Mach Intell 2006;:351–363.
[54] Paysan, P, Knothe, R, Amberg, B, Romdhani, S, Vetter, T. A 3d face model for pose and illumination invariant face recognition. IEEE International Conference on Advanced Video and Signal Based Surveillance 2009;:296–301.
[55] Deng, J, Guo, J, Niannan, X, Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2019, p. 4690–4699.
[56] Isola, P, Zhu, JY, Zhou, T, Efros, AA. Image-to-image translation with conditional adversarial networks. IEEE Computer Vision and Pattern Recognition (CVPR) 2017;:1125–1134.
[57] Yu, C, Wang, J, Peng, C, Gao, C, Yu, G, Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. European Conference on Computer Vision (ECCV) 2018;:325–341.
[58] Liu, Z, Luo, P, Wang, X, Tang, X. Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision (ICCV). 2015, p. 3730–3738.
[59] Feng, Y, Wu, F, Shao, X, Wang, Y, Zhou, X. Joint 3d face reconstruction and dense alignment with position map regression network. In: European Conference on Computer Vision (ECCV). 2018, p. 534–551.
[60] He, K, Zhang, X, Ren, S, Sun, J. Deep residual learning for image recognition. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2016, p. 770–778.
[61] Zhu, X, Lei, Z, Liu, X, Shi, H, Li, SZ. Face alignment across large poses: A 3d solution. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2016, p. 146–155.
[62] Shi, T, Zou, Z, Song, X, Song, Z, Gu, C, Fan, C, et al. Neutral face game character auto-creation via pokerface-gan. In: ACM International Conference on Multimedia. 2020, p. 3201–3209.
[63] Lee, CH, Liu, Z, Wu, L, Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2020, p. 5549–5558.
[64] Schneider, A, Schonborn, S, Frobeen, L, Egger, B, Vetter, T. Efficient global illumination for morphable models. In: IEEE International Conference on Computer Vision (ICCV). 2017, p. 3865–3873.
[65] Wu, X, He, R, Sun, Z, Tan, T. A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security 2018;13(11):2884–2896.
[66] Zhao, J, Li, J, Tu, X, Zhao, F, Xin, Y, Xing, J, et al. Multi-prototype networks for unconstrained set-based face recognition. In: IJCAI. 2019,.