Hybrid Face Reflectance, Illumination, and Shape from a Single Image

Yongjie Zhu, Chen Li, Si Li, Boxin Shi, Senior Member, IEEE, and Yu-Wing Tai, Senior Member, IEEE

Abstract—We propose HyFRIS-Net to jointly estimate the hybrid reflectance and illumination models, as well as the refined face shape from a single unconstrained face image in a pre-defined texture space. The proposed hybrid reflectance and illumination representation ensure photometric face appearance modeling in both parametric and non-parametric spaces for efficient learning. While forcing the reflectance consistency constraint for the same person and face identity constraint for different persons, our approach recovers an occlusion-free face albedo with disambiguated color from the illumination color. Our network is trained in a self-evolving manner to achieve general applicability on real-world data. We conduct comprehensive qualitative and quantitative evaluations with state-of-the-art methods to demonstrate the advantages of HyFRIS-Net in modeling photo-realistic face albedo, illumination, and shape.

Index Terms—Face modeling, intrinsic image decomposition, shading, reflectance, illumination

1 INTRODUCTION

Modeling face reflectance, shape, and the environment illumination is crucial to photo-realistic face image manipulations, such as virtual makeup [1, 2], face inpainting [3], facial attribution editing [4], and so on. The realistic manipulation requires the image to be decomposed into several independent components before editing. However, different from multi-view stereo [5] or photometric stereo [5] that could use multiple images to capture face appearance, single face image decomposition under an unconstrained environment is a highly ill-posed problem because the number of unknowns is several folds to the number of known variables. Since the ground truth is hard to access, conventional solutions adopt optimization with low-level priors, including chromaticity variation [7–9], geometry guided shading consistency [10, 11], reflectance consistency over varying lightings [12, 13], and statistical/physical-based reflectance model [14, 15]. In addition to the significant runtime for solving the complicated non-linear optimizations, the over-simplified assumptions on face reflectance/illumination and the request of frontal face images [14, 16] make the optimization-based solutions impractical especially under arbitrary head poses or complex illumination. By incorporating parametric representations of face reflectance, illumination, and shape, deep learning is applied to formulate the process as a regression problem [17–19] or a generation problem [20, 21] that evaluates how well the network outputs fit the input images photometrically.

Although previous approaches offer impressive results, they still have four major drawbacks: 1) most of them [17, 19, 21] assume the face reflectance is Lambertian and cannot correctly handle the common specular reflection caused by oily skin; 2) the ambiguity between the color of face albedo and illumination cannot be well separated, resulting in different face albedo of the same person under different illuminations [17, 19, 21]; 3) existing
methods can handle different face poses, but the estimated albedo either loses skin details under parametric representation [17, 19] or is rather incomplete in occluded areas under non-parametric representation [4, 21]; 4) the networks are usually trained with synthetic data rendered by over-simplified model (e.g., Lambertian), as supervision [18, 21], which results in unrealistic face appearance modeling and gap between synthetic and real data during training.

We propose the Hybrid Face Reflectance, Illumination, and Shape modeling Network, HyFRIS-Net, to solve the above four problems, with major contributions summarized as follows:

- We propose hybrid reflectance (Lambertian + Blinn-Phong) and illumination (spherical harmonics + distant lights) models for both physically and visually plausible face appearance modeling.
- We propose novel reflectance consistency and identity constraints to guide our network training by fully exploring the relationship of human face images with the same identity.
- We estimate complete face albedo from arbitrary poses by solving the alignment problem in the face texture space and filling the occluded region using a generative-adversarial network.
- We propose a self-evolving training scheme that takes the advantage of earlier output to iteratively improve the later results so that the network can be trained directly on real images, which naturally bridges the gap between synthetic and real data.

We compare our results with state-of-the-art face modeling methods [4, 17, 21], facial specular removal methods [20, 22], and facial geometry reconstruction method [23], which demonstrate our advantages including estimating complete face albedo map with correct skin color, hybrid lighting model with rich frequency information, and refined face normal map with enhanced details.

2 RELATED WORK

Estimating reflectance, illumination, and shape from a general scene have been studied for decades and we only discuss prior single image works.

Face reflectance modeling. The complete face reflectance model is rather complex when considering the specular highlights [20, 22] or sub-surface scattering [14, 24], so the majority of approaches still assume the face reflectance is Lambertian due to its simplicity [4, 17–19, 21, 25]. But such simplification usually causes the specular color to incorrectly remains in the estimated face albedo. Besides, some methods represent the face albedo in a low-dimensional parametric space and their estimated albedo generally loses skin details [17, 18]. All the methods, including ours, require a photometric loss to constrain the estimation under certain face appearance model fits with the input images. The image formation can be implemented as a forward rendering process [14, 21] or modeled by a differentiable image formation layer using CNNs [17, 18]. Because the face appearance modeling problem is highly unconstrained, most methods apply synthetic data to pre-train their model or mix the synthetic data with real images [20, 21]. These methods usually over-fit the synthetic data and cannot properly handle the data distribution in real images. What’s more, some methods [5, 26] use the light stage to access the ground truth face appearance data and use it to train a deep network to estimate the face appearance [26]. However, such data are expensive to access and the captured face images in the lab have less diverse appearances from unconstrained images shared on the Internet by casual users.

Face illumination estimation. A real-world environment illumination can be formulated using different models, including a set of parametric lights [15, 16, 22, 27, 28], and non-parametric panorama maps [29]. Most illumination estimation works on face images assume a Lambertian face appearance and use second-order spherical harmonics to only approximate the diffuse components of reflectance and illumination [4, 17–19, 21, 25]. Recently, researchers start to use the specular highlights caused by oily skin to reconstruct the high-frequency part of illumination which is expressed as high-order spherical harmonics [14], a mixture of light chromaticity and amplitude [22], or a non-parametric environment map [20]. These methods all restrict the illumination chromaticity via statistical priors about skin color. Different from them, we separate the whole environment map into two independent parts: spherical harmonics and sparse distant light which are used to calculate the diffuse shading and specular reflection, respectively. In this way, our method can estimate the complete intrinsic components and then easily use them for inversely rendering a face image.

Face shape estimation. Recent 3D face reconstruction works represent facial shapes by the PCA-based 3D morphable model (3DMM) [30] and have shown impressive results by aligning the 3DMM to known 2D facial landmarks [31–33]. The 3DMM parameters can be regressed from the input image using CNNs directly [17–19]. Although face shapes can also be represented by other formulations such as volumetric space [34], point clouds [35], or normal map [4, 21], these methods still rely on 3DMM for generating the training data. To produce the missing facial structure details due to the low-dimensionality of 3DMM, there are some methods applying shape-from-shading [16, 36], generated bump map [23, 37] and displacement map [38] to refine the low-dimensional shape by adding high-frequency details. In this paper, we introduce a shape refine strategy inspired by shape-from-shading methods. Moreover, despite we do not use the captured ground truth data to train deep models like [38], we still achieve high accuracy compared with other methods [21, 23] which also have no-ground truth labels on real data.

3 PROPOSED METHOD

Given a face image, we first crop the face region and fit the 3DMM [30] with its 3D facial landmarks [33, 40] to get a rough initialization of the face normal map. The face image and its
3.1 Hybrid image formation model

Face appearance is determined by the complicated interaction between face reflectance, environment illumination, and face shape. Simultaneously decomposing all these factors from a single image is prohibitively difficult if both face reflectance and environment illumination are considered as general as possible (e.g., using BSSRDF for face reflectance [14] and non-parametric environment map for illumination). To make a good balance between realistic modeling and efficient learning, we adopt a hybrid representation that consists of a diffuse component and a specular component as [1, 20]:

\[ I = A \odot D + S, \]

where A is the face albedo which scales the diffuse shading D through a pixel-wise multiplication \( \odot \) and S is the specular reflection.

Since D and S encode the low- and high-frequency illumination respectively, we model the illumination as a combination of low-frequency part \( L^d \) and high-frequency details \( L^s \). Both \( L^d \) and \( L^s \) are represented in a parametric manner rather than a panorama map to significantly reduce the degree of freedom - [39, 41].

**Diffuse component.** We adopt the Lambertian model for the diffuse component [1, 4, 16, 17] and use the 2nd-order spherical harmonics (SH) to approximate the diffuse shading \( D_p \) for pixel \( p \) in a quadratic polynomial of the face surface normal direction \( \mathbf{n}_p = (x, y, z, 1) \) as:

\[ D_p = \mathbf{n}_p^T \mathbf{H}(L^d) \mathbf{n}_p, \]

where \( L^d \) is the SH coefficients and \( \mathbf{H}(L^d) \) is a symmetric \( 4 \times 4 \) matrix defined in [39].

**Specular component.** We use a set of distant lights, \( L^s = \{ \mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_{K_s} \} \), to model the remaining high-frequency component of illumination. Therefore, the specular highlights \( S_p \) can be represented using the Blinn-Phong model [14, 42] as:

\[ S_p = \sum_{i=1}^{K_s} \int_{\omega} (\mathbf{n}_p \cdot \mathbf{h})^m \mathbf{l}_i(\omega) d\omega. \]

Here \( m \) is the surface smoothness of human skin, and \( \mathbf{h} \) is the half vector defined as the bisector of incident angle \( \omega \) and the viewing direction \( \mathbf{v} \).

In our implementation, we model the RGB channels of illumination separately. We use the 2nd order SH to model diffuse illumination so the number of \( L^d \) is 27. As a trade-off between accuracy and computational complexity for \( L^s \), we use \( K_s = 22 \) distant lights with pre-defined lobe axes. Since the amplitude is also defined in RGB channels, there are a total of \( 3 \times 22 \) parameters in \( L^s \). Fig. 2 illustrates an example of our hybrid illumination model and the 22 pre-defined directional light positions are marked in red dots. Although our model cannot measure the full BRDF at each surface point, the normal map refined by our HyFRIS-Net is of a sufficient resolution to yield some of the appearance of spatially-varying specular smoothness. And we set the surface smoothness \( m \) of human skin as 22 to reduce computational complexity, according to the statistical MERL/ETH Skin Reflectance Database [42].

3.2 Network architecture

The proposed HyFRIS-Net includes three sub-networks, namely albedo generation network, illumination estimation network, and shape refinement network, to reconstruct the facial albedo, environment illumination, and face shape from a single image. Among these three sub-networks, the albedo generation network uses only the facial image as input, and the rest two concatenate the initialized normal map and facial image together as their input. All the input is transferred into the pre-defined texture space [18]. The projection of a facial vertex \( \mathbf{v} = (x, y, z)^T \) onto the texture space \( t = (u, v) \) is computed as:

\[ u = \alpha_1 \arctan\left(\frac{x}{z}\right) + \beta_1, \]

\[ v = \alpha_2 y + \beta_2, \]

where \( \alpha_1, \alpha_2, \beta_1, \) and \( \beta_2 \) are constant scale and translation to place the unwrapped face into the texture space.

**Albedo generation network.** One challenge of estimating the facial albedo map is that the pixel needs to be transferred to the albedo domain by decoupling illumination effects. It is not straightforward because of the ambiguity existing between colors of skin and illumination. Moreover, the occluded regions caused by different face orientations should be in-painted properly to obtain a complete estimation of the face albedo.

To solve these problems, we adopt a generative adversarial network architecture. We use a standard U-Net [43] as our generator in the albedo generation network, and the sizes of input and output are \( 448 \times 448 \) with 3 channels. This generator consists of a contracting path and an expansive path. The contracting path is a typical convolutional network and includes repeated blocks of two \( 3 \times 3 \) convolutions followed by a rectified linear unit (ReLU) and a \( 2 \times 2 \) max pooling with stride 2. The expansive path upsamples the extracted feature map to meet the output resolution and concatenates the features in the contracting path at different resolution levels to preserving local details.

And then the discriminator takes the output of our generator as input and outputs a \( 28 \times 28 \) patch-based descriptor for better preserving the personal characteristic of generated facial albedo. It includes five \( 4 \times 4 \) convolution layers in total where the strides of the first four layers are set to 2 and the stride of the last layer is set to 1.

**Illumination estimation network.** As described in Sec. 3.1, we model the illumination as a hybrid combination of 2nd-order SH \( L^d \) and 22 distant lights \( L^s \). The total number of parameters is 93 since we consider RGB channels separately.

According to the results of our ablation study (details in Sec. 5.2), we choose VGG11 [44] as the backbone of our illumination estimation network. In order to estimate the low-frequency and high-frequency illumination, we use two independent branches to obtain SH coefficients \( L^d \) and distant lights \( L^s \).
Face image database

Image data. To solve this problem, we propose a self-evolving training strategy to iteratively fine-tune the pseudo-labels obtained from real data and eliminate the gap between synthetic data and real data.

Our self-evolving training consists of a multi-stage process as follows: (a) We train our albedo generation network on labeled synthetic data. (b) We apply this network on real data to obtain the initial albedo proxy, and we also initialize the normal proxy using its 3DMM fitting results (see “self-evolving training” labeled in red and green boxes in Fig. 3). (c) We train the entire HyFRIS-Net from scratch only using real images with albedo and normal proxies and keep updating these two proxy buffers. In each generation, the albedo generation network and illumination estimation network are trained jointly while the shape refinement network is trained separately, by minimizing:

\[ L_{\text{AL}} = \lambda_r L_{\text{photo}} + \lambda_c L_{\text{cons}} + \lambda_a L_{\text{adv}} + \lambda_l L_{\text{triplet}} + \lambda_s L_{\text{smooth}} + \lambda_p L_{\text{specular}}, \]

\[ L_N = \lambda_r L_{\text{photo}} + \lambda_n L_{\text{normal}}, \]

where \( L_{\text{photo}}, L_{\text{cons}}, L_{\text{triplet}}, L_{\text{smooth}}, \) and \( L_{\text{normal}} \) identify the photometric loss, the reflectance consistent loss, the face identity loss, the albedo smoothness loss, the specular highlight loss, and the shape refinement loss respectively, and \( \lambda \) balances each individual loss.

After we finish the training of one generation, both albedo proxy and normal proxy are updated with the output of the current generation before starting the training of the next generation. According to our ablation study, although the results of our network could be in low accuracy at the earlier generations, it keeps evolving and finally achieves highly precise results for albedo, illumination, and shape. The red and green dashed arrows in Fig. 3 identify the self-evolving path during the training stage.

4 SELF-EVOLVING TRAINING

In this section, we explain our self-evolving training strategy for dealing with the absence of ground truth supervision followed by introducing the definitions of important training losses.

4.1 Training strategy

The self-supervised learning in previous method [21] uses model trained with synthetic data to directly test on real data and obtain pseudo-labels for training. In our training process, the photometric loss in self-supervised learning plays a key role in detailed information recovery. But during the training, the pseudo-labels are not fine-tuned which limits the model’s performance on real data. To solve this problem, we propose a self-evolving training strategy to iteratively fine-tune the pseudo-labels obtained from real data and eliminate the gap between synthetic data and real data.

4.2 Training losses

Photometric loss. As the proposed hybrid model introduced in Sec. 3, the face image can be decomposed into three components: albedo \( A \), diffuse \( D \) and specular \( S \); all of them are RGB images with the same resolution as the input image; \( D \) and \( S \) can be computed given the normal map and diffuse/specular lighting.
where \( A, L^d, L^s, N \) are the output of the proposed HyFRIS-Net. \( M \) is a mask in facial texture space to identify the non-occluded pixels and is determined when warping the input image to facial texture space.

**Reflectance consistent loss.** Previous methods alleviate the ambiguity between skin and illumination colors by incorporating statistical priors on either human skin color \([14]\) or illumination color \([15]\), but such priors cannot guarantee face skin color to be correctly estimated under different illuminations. We thus propose the reflectance consistent constraint based on the assumption that the facial reflectance of the same person is consistent in varying face poses or illuminations. Such a constraint can hardly be employed in previous optimization-based approaches \([11, 14, 15]\) because the number of the input image is limited to one. By adding it in the training stage, we achieve more robust skin and illumination color disambiguation. In the testing stage, we still use only one image as input. In each training batch, we randomly select \( p = 3 \) persons and \( k = 3 \) images for each person, so the averaged albedo map of one person is:

\[
\hat{A}_i = \frac{1}{\sum_{j=1}^{k} M_{i,j}} \sum_{j=1}^{k} M_{i,j} \odot \hat{A}_{i,j}. \tag{9}
\]

We then compute the reflectance consistent loss as:

\[
\mathcal{L}_{\text{cons}} = \sum_{i=1}^{p} \sum_{j=1}^{k} \|\hat{A}_{i,j} - \hat{A}_i\|_2^2, \tag{10}
\]

to minimize the distance between estimated albedo map \( \hat{A}_{i,j} \) and masked average albedo map \( \hat{A}_i \). The calculation of \( \hat{A} \) merges the albedo map of one person from different face orientations, so a byproduct of \( \mathcal{L}_{\text{cons}} \) is to generate a complete albedo map without any missing estimations in occluded regions.

**Face identity loss.** To preserve the face identity information, we incorporate a face identity loss by considering a training triplet \( \{A^n, A^p, A^\mu\} \) as:

\[
\mathcal{L}_{\text{triplet}} = \sum_{i} \left( \|\hat{A}^n_i - A^p_i\|_2^2 - \|\hat{A}^n_i - A^n_i\|^2_2 + \delta \right)_+, \tag{11}
\]

where \( A^p, A^\mu, \) and \( A^n \) is the anchor albedo, a positive albedo sample, and a negative sample of the face identity \( i \), respectively. \( A^p_i \) and \( A^n_i \) are from the same face identity, and \( A^\mu_i \) is randomly selected from different identities. We set \( \delta = 0.19 \) in our implementation as the minimal margin between positive and negative samples. In addition, the face identity loss can help avoid a local minimal solution with a uniform albedo to all different images.

**Specular highlight loss.** Different from the diffuse reflection that arises from subsurface scattering and is mostly unpolarized, specular reflectance from skin preserves the plane of polarization information, so the specular highlights can be separated by controlling the polarization state of the light source. But such an operation cannot be applied to Internet images. According to previous research on specular removal \([22, 46]\), specular highlights can be determined once the specular illumination color \( \mu \) and the chromaticity of face albedo are both known:

\[
\tilde{S} = \max_{u \in \{r, g, b\}} \frac{I_{\mu u} - \Lambda_{\max}}{1 - 3 \Lambda_{\max}} \sum_{u \in \{r, g, b\}} I_{\mu u}. \tag{12}
\]

We can easily calculate the diffuse chromaticity \( \Lambda^2 \) from the estimated albedo \( A \). We then approximate the illumination color \( \mu \) from the estimated distant lights \( L^s \). Thus a specular highlight loss is defined as:

\[
\mathcal{L}_{\text{specular}} = \|S(L^s) - \tilde{S}\|_2^2, \tag{13}
\]

which complements the photometric loss \( \mathcal{L}_{\text{photo}} \) and the diffuse shading model in Eq. (2) to separate specular highlights from input images.

2. The chromaticity of a pixel with intensity \( \{r, g, b\} \) is \( \{\frac{r}{r+g+b}, \frac{g}{r+g+b}, \frac{b}{r+g+b}\} \).
**Albedo adversarial loss.** In addition to supervising the network output pixel-wise through photometric loss \( L_{\text{photo}} \) and reflectance consistency loss \( L_{\text{cons}} \), we further incorporate a discriminator to train our albedo generation network through a generative-adversarial process for generating the albedo map as realistic as possible. The adversarial loss is expressed as:

\[
L_{\text{adv}} = \mathbb{E}[\log D_R(\hat{A})] + \mathbb{E}[\log(1 - D_R(A))],
\]

where \( D_R \) denotes the discriminator in the albedo generation network, \( \hat{A} \) denotes the estimated albedo from the generator in the albedo generation network, and \( A \) denotes the positive albedo sample in the albedo proxy buffer.

**Albedo smoothness loss.** In order to classify an image gradient to either albedo or shading changes, we follow a common approach based on the Retinex theory [47] as:

\[
L_{\text{smooth}} = \| (\nabla \Lambda_{\leq 0.1} \circ \nabla A) \|_2^2,
\]

where \( \nabla \) denotes the gradient operator and \( \Lambda_{\leq 0.1} \) denotes the chromaticity of estimated facial reflection in one training generation before.

**Shape refinement loss.** The shape refinement network aims to refine the normal \( \text{N} \) in the normal proxy buffer by introducing more geometric details. We employ a pixel-wised data term and a smoothness term to restrict its output as:

\[
L_{\text{normal}} = \| \text{N} - \tilde{\text{N}} \|_2^2 + \lambda_{ns} \sum_p \sum_{q \in \mathcal{N}_p} \|1 - n_p^T n_q\|_2^2,
\]

where \( \text{N} \) is the output of the shape refinement network, \( \mathcal{N}_p \) is the 4-connected neighborhood of facial pixel \( p \), and \( \lambda_{ns} \) is a weighting term fixed as 0.2.

### 4.3 Implementation details.

To generate synthetic data, we use 3DMM with different face properties, including shape and texture parameters. Specially, we use face parameters in the 300W-LP dataset [48] which is composed of paired large-pose face 3D models and images to generate different samples with large pose and texture variety. For real data, we use the VGGFace2 [49], a human face dataset containing 9131 subjects. For each subject, 10 images are manually selected for covering different face orientations and illuminations. Then we use the images and corresponding shape initializations in the face texture space with 448 \( \times \) 448 resolution as the input to our network. Our framework is implemented in PyTorch and Adam optimizer is used with default parameters. We set \( \{\lambda_p, \lambda_c, \lambda_t, \lambda_a, \lambda_n, \lambda_{sp}\} = \{0.0, 0.1, 1, 0.1, 0.01, 1.0\} \) as constant during the training stage. We train our network for 3 generations. Within one generation, each sub-network is trained for 20 epochs until convergence. And after each generation, the albedo and normal in proxy buffers will be updated automatically.

### 5 Experimental Results

We compare the results of HyFRIS-Net with six state-of-the-art methods, namely SfSNet [21], MLFace [17], NeuralFace [4], FaceHR [22], FaceProbe [20] and Pix2Vertex [23]. The results are either provided by the authors [17, 20, 22] or generated using the released model [4, 21, 23]. We inversely warp our results (HyFRIS-Net operates in the facial texture space) back to the original image space for easy reference.

**Comparison on face albedo estimation.** Besides the results shown in Fig. 4, we further compare our face albedo results with NeuralFace [4] and SfSNet [21] with images from the same subjects. In addition to our improvement achieved by incorporating the hybrid illumination and face reflectance model, our method demonstrates more consistent face albedo for the same subject in Fig. 5 than others due to the utilization of reflectance consistent loss. To quantify the consistency of albedo, we compare MSE and SSIM errors between the estimated albedo map and mean albedo map with the same identity on the MultiPie dataset [50]. Since the estimated albedos of NeuralFace [4] and SfSNet [21] are represented in the projected image space, we compute the

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Methods} & \text{NeuralFace [4]} & \text{SfSNet [21]} & \text{Ours} \\
\hline
\text{MSE} & 0.122 & 0.017 & 0.001 \\
\text{SSIM} & 0.803 & 0.884 & 0.990 \\
\hline
\end{array}
\]

Fig. 5. Comparison with NeuralFace [4] and SfSNet [21] on face albedo estimation using images from the same subjects.
3D geometry and its UV texture coordinate using its 3DMM fitting, and then we warp the estimated albedo to the aligned UV space so that we can calculate the errors and compare them with our estimated albedo maps. During the warping process, the misaligned and occluded parts are automatically removed to ensure the reliability of the results. Our method shows much smaller errors.

Comparison on specular highlight separation. We compare our highlight separation results with FaceProbe [20] and FaceSHR [22] on the FaceProbe dataset which contains 30 frontal face images and the corresponding ground truth of diffuse component in Table 1 and Fig. 6. The ground truth of diffuse reflection is captured by cross-polarization under an indoor directional light. The results of FaceProbe and FaceSHR are provided by their authors. We employ two metrics, RMSE and SSIM, to measure the diffuse reflection estimation errors and similarity with ground truth, respectively. Our method quantitatively outperforms FaceSHR [22] and FaceProbe [20], especially in RMSE.

FaceProbe [20] first separates specular highlights from an input image, then projects the highlight to an illumination probe coordinate, and finally corrects the illumination color. In contrast to such a single-forward “image → illumination” solution, our solution is bounded by the physically based photometric loss to perform a cycle constraint as “image → illumination → image”. In the qualitative examples shown in Fig. 7, our method shows a more reasonable estimation of specular magnitude where FaceProbe [20] and FaceSHR [22] tend to produce a too strong specular reflection.

Comparison on illumination estimation. We follow the quantitative evaluation protocol in SfSNet [21] for evaluating our parametric illumination estimation on the MultiPIE dataset [50]. We sample 5 camera directions around the frontal face direction, namely 13_0, 14_0, 05_1, 05_0, 04_1, and 05_1 is the frontal camera direction. The result is significantly improved by our method, especially on top-1 accuracy.
estimation for each lighting condition. We then evaluate our estimated illuminations of test images under each camera as a classification problem with 19 classes of lighting conditions. For fully using this dataset, we carry out 10 cross-validations for this classification. The detailed top-1, top-2 and top-3 classification accuracy are shown in Fig. 8. Our method outperforms SfSNet in 12 of the total 15 comparisons, especially achieving a significant improvement in the accuracy of frontal camera 05_1.

Comparison on shape estimation. We use the Florence dataset [51] which contains 53 scanned face meshes (14 females and 39 males) to evaluate our face shape estimation and measure the accuracy as the angular error between predicted normal and ground truth normal. For each model, we render a facial image under frontal camera direction with weak perspective projection, and then compute the ground truth of its normal map accordingly. We use these rendered facial images as the test images and compare the results with SfSNet [21] and Pix2Vertex [23] in Table 2 and Fig. 9. Because the output of Pix2Vertex is a 3D point cloud, we additionally employ a rigid registration with ICP to align its results with the ground truth model and compute the corresponding normal map estimation. As aforementioned, our method achieves better accuracy in the estimation of face albedo and illumination due to the hybrid modeling with corresponding training losses and the self-evolving training strategy. The more reliably estimated albedo and illumination directly contribute to the improvement in the face shape estimation over previous methods, especially in high-quality regions (< 20°).

Additional comparisons. Besides the qualitative comparison presented in our paper, we further include more results by comparing with MLFace [17] (Fig. 10), NeuralFace [4] (Fig. 11), and SfSNet [21] (Fig. 12).

The 3DMM based method MLFace [17] uses the low dimensional representations for both face albedo and shape, so their results lack skin details. In contrast, our estimated albedo and shape preserve the small facial structure better because of our jointly parametric and non-parametric representations.

Although NeuralFace [4] estimates reasonable illumination in Fig. 11, the estimated albedo is over-smooth and results in a large photometric error as demonstrated in their reconstruction results. SfSNet [21] shows good performance on the estimation of facial albedo, illumination, and shape in Fig. 12. However, it does not model the specular reflection and wrongly leaves it in the albedo component. Moreover, the synthetic data used to pre-train SfSNet [21] do not include human hairs, so they cannot correctly handle heavy bangs or beards in the face region. Compared

<table>
<thead>
<tr>
<th>Methods</th>
<th>Mean Err.</th>
<th>&lt; 20°</th>
<th>&lt; 30°</th>
<th>&lt; 45°</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pix2Vertex</td>
<td>27.4°</td>
<td>41.4%</td>
<td>64.5%</td>
<td>84.7%</td>
</tr>
<tr>
<td>SfSNet [21]</td>
<td>25.6°</td>
<td>71.1%</td>
<td>85.9%</td>
<td>91.7%</td>
</tr>
<tr>
<td>Ours</td>
<td>18.8°</td>
<td>81.5%</td>
<td>90.1%</td>
<td>94.1%</td>
</tr>
</tbody>
</table>

Fig. 10. Comparison with MLFace [17] on albedo, shading, shape, and reconstruction. MLFace [17] uses 3DMM to fit the face image which loses face details. In our results, such details are faithfully recovered.
5.2 Ablation study

We analyze the performance of different backbone architectures for illumination estimation as well as the importance of proposed loss functions by training the proposed HyFRIS-Net using synthetic data with ground truth labels. We also validate the necessity of applying our self-evolving training strategy on specular separation task using the FaceProbe dataset [20].

Lighting estimation network architectures. In order to determine an appropriate structure for the illumination estimation network, we evaluate five types of structures on RMSEs of spherical harmonics (SH Err.) and Distant lights (DL Err.), including AlexNet [52], ResNet18 [53], ResNet50 [53], VGG11 [44], and VGG16 [44]. According to the evaluation results shown in Table 3, we found a shallow network, e.g., VGG11 [44], always outperforms its deep version for modeling both the low-frequency and high-frequency components of the illumination. We believe a good estimation of such illumination only requires local image features without the need of handling large and deep receptive fields, so a shallow network performs surprisingly better. This conclusion is consistent with the study presented in [19]. Therefore, we choose VGG11 [44] as our feature backbone for the two sub-networks in the illumination estimation network. To validate the effectiveness of our hybrid illumination model, we compare our method with alternative light source representations with fewer parameters, including three distant lights and eleven distant lights with unknown directions. As shown in Fig. 13, using three distant lights has the highest error. By increasing the numbers of light sources to eleven, the network starts to be able to lower the error, but also exhibits unstable fluctuation in the prediction. By using our hybrid lighting model, the network is able to converge faster and smoother with smaller fitting errors.

Training loss justification. Our method estimates face albedo, parametric low- and high-frequency illumination and face shape from a single input image in a self-evolving training manner,

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet [52]</td>
<td>3.56</td>
<td>6.01</td>
</tr>
<tr>
<td>ResNet18 [53]</td>
<td>4.52</td>
<td>5.86</td>
</tr>
<tr>
<td>ResNet50 [53]</td>
<td>5.13</td>
<td>6.01</td>
</tr>
<tr>
<td>VGG11 [44]</td>
<td>3.22</td>
<td>2.85</td>
</tr>
<tr>
<td>VGG16 [44]</td>
<td>3.88</td>
<td>4.15</td>
</tr>
</tbody>
</table>
so reasonable training losses are necessary to achieve accurate results. Table 4 shows an ablation study on the proposed reflectance consistent loss $L_{\text{cons}}$ and face identity loss $L_{\text{triplet}}$ using synthetic training data. Not surprisingly, use both $L_{\text{triplet}}$ and $L_{\text{cons}}$ together achieve the highest accuracy and $L_{\text{triplet}}$ shows more importance than $L_{\text{cons}}$. Without using $L_{\text{triplet}}$, only $L_{\text{photo}}$ and $L_{\text{cons}}$ cannot prevent the albedo generation network from generating a uniform albedo while the illumination network and shape network could be wrongly trained when given this uniform albedo. We also evaluate the effect of albedo adversarial loss in Fig. 14. Since the predicted albedo maps are represented in UV texture space, we compute the proxy mesh by employing 3DMM fitting on face key points and then render the images using this mesh with corresponding albedo maps. We rotate the reconstructed face mesh to three different viewpoints for better visualization. From the results, we can observe that without the adversarial loss, the network tends to produce blurry estimations on the occluded regions. By adding the adversarial loss, the network is able to infer more realistic high-frequency albedo textures.

**Self-evolving justification.** We validate the self-evolving training strategy by boosting our network with four different albedo proxies, including only using the synthetic albedo and using the albedo proxies in three training generations. Because we do not have the ground truth of face albedo, we instead show the evaluations of these four networks on specular separation task using the FaceProbe dataset [20], as shown in Table 5. We evaluate the RMSE and SSIM errors between the reconstructed diffuse image and captured ground truth diffuse face. Since the synthetic face albedo deviates far from the real albedo, it performs badly at the beginning. By keeping updating the albedo proxy with the output of our network, the estimated face albedo is self-evolved and becomes closer to the real albedo, which in turn results in more accurate specular separation. By using the proposed training scheme, our network eliminates the ambiguity between albedo and illumination in both low- and high-frequency domains.

### 6 Applications

Photo-realistic face manipulations depend on face reflectance, shape and illumination heavily. For better validating our estimation.
of each component, we apply our results on various applications, including diffuse/specular editing, relighting and light transfer.

**Diffuse editing.** In Fig. 15, we show the results of diffuse shading editing by preserving the estimated albedo and normal but replacing the low-frequency diffuse lighting by five different SH coefficients setups with different color and intensity distributions. Thanks to our accurate estimations, the results are quite realistic to represent each virtual diffuse lighting. Compared with SfS-Net [21], our relighted face appears to be more yellow under the first warm lighting and to be white under the second cool lighting. Furthermore, strong incident diffuse light leads our relighted face to have a stronger contrast in the corresponding directions, for example, the right part of the face in the fifth case.

**Specular editing.** In Fig. 16, we enhance three different distant lights (left, frontal, right) to virtually editing the specular reflection. And the results in the second/third row show reasonable magnitude and direction of specular highlights cause by the corresponding enhanced distant lights.

**Light transfer.** In Fig. 17, we show the results of light transfer based on our estimated face appearance model. Seven subjects under different illuminations are included in the first column and we present the corresponding relighting results on the same subject with these seven illuminations in each row of Fig. 17. Because our method can address the ambiguity between skin color and illumination color, the purplish illumination in Subject 5 is correctly estimated and transferred to other subjects, as demonstrated in the sixth column. Furthermore, we model both the low-frequency and high-frequency illumination in a hybrid model, so the strong specular highlights in Subject 4 are well modeled and the corresponding specular reflection appears in its relighting results (the fifth column). In contrast, the illumination in Subject 3 contains a less high-frequency component, so we do not observe significant specular highlights in its relighting results in the fourth column of Fig. 17.

### 7 Conclusion

We present a method to learn the hybrid face model from Internet face images. Compared to the conventional face modeling methods, the Internet face images have a wide range of unknown lighting information and camera settings, which makes it hard to compute the face components directly. We solve this challenging problem by proposing HyFris-Net – a three-branch network, integrating the hybrid face modeling in the architecture. We combine the reflectance consistency and facial identity constraints in our training losses to resolve the ambiguity between the skin and light color. We further develop a self-evolving training scheme for jointly using synthetic data and real data on the Internet and eliminating the gap between them. The effectiveness of the proposed method is verified both quantitatively and qualitatively using extensive comparisons on different datasets and tasks.

Currently, face modeling methods using Internet images mainly need to use 3DM [54] to constraint the face performance. With hybrid face modeling using Internet images introduced, it will be interesting to integrate with 3DM-based fitting and

---

**TABLE 5**

Ablation study on self-evolving training.

<table>
<thead>
<tr>
<th></th>
<th>Syn.-Only</th>
<th>Gen. 1</th>
<th>Gen. 2</th>
<th>Gen. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RMSE</td>
<td>14.97</td>
<td>12.57</td>
<td>7.52</td>
<td><strong>7.15</strong></td>
</tr>
<tr>
<td>Median RMSE</td>
<td>13.53</td>
<td>12.79</td>
<td>7.70</td>
<td><strong>6.53</strong></td>
</tr>
<tr>
<td>Mean SSIM</td>
<td>0.66</td>
<td>0.76</td>
<td>0.86</td>
<td><strong>0.90</strong></td>
</tr>
<tr>
<td>Median SSIM</td>
<td>0.68</td>
<td>0.72</td>
<td>0.87</td>
<td><strong>0.91</strong></td>
</tr>
</tbody>
</table>

---

**Fig. 13.** Reconstruction errors in training on real data with different numbers of light sources under the same optimization settings. We use overlaid darker lines to plot the smoothed error curves for better visualization.

**Fig. 14.** An example of the ablation study on adversarial loss. Red boxes indicate noticeable differences.

**Fig. 15.** Diffuse shading editing results. The top row shows the diffuse lighting hypothesis and the other rows below show the corresponding editing results based on estimated face albedo and shape of our model, compared with SfSNet [21].
improve its accuracy with the decomposing and appearance modeling tasks. Different from multiple images-based methods and other single image based deep learning methods, the proposed method uses only a single image as input and generates high-quality results at the same time. Although our method can handle a variety of lighting conditions, it has not yet considered the self-shadowing effects caused by occlusions. It might be possible to directly integrate such a constraint into our learning-based framework if training datasets with corresponding ground truth labels are available to further separate self-shadows from the estimated albedo. The results of our method are expected to get further improvement, if the high quality of initial 3D information is provided (3DMM fitting performs poorly if key point detection fails). What’s more, the misalignment of the warped map in texture space could not be eliminated in our method, because we use the reconstruction error to constrain the recovery, and it may be an interesting future topic to correct the misalignment caused by initialization fitting during the reconstruction process.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China under Grant No. 61872012, 62088102, and Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

Fig. 17. Light transfer results. Seven different illuminations are estimated from the images of Subject 1 to Subject 7 in the first column. The relighting results on the same subject with these seven illuminations are shown in each row.


[40] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In Proc. of International Conference on Computer Vision, 2017. 2


Yongjie Zhu received the B.E. degree from Beijing University of Posts and Telecommunications (BUPT) in 2019. He is currently pursuing his Master degree at Beijing University of Posts and Telecommunications. His research interest is physics based computer vision.

Chen Li is a senior researcher in WeChat, Tencent (Apr. 2020 - present) and leads the applied vision team of WeChat. Before he joined WeChat, he was a senior researcher in Youtu Lab of Tencent (Sep. 2019 - Apr. 2020) and an algorithm engineer in DJI (Jan. 2017 - Sep. 2017). He received the PhD and BS degree from Zhejiang University in 2017 and 2011, respectively. He worked as a research intern with the Internet Graphics group of Microsoft Research, Beijing, from Feb. 2012 to Dec. 2016. His research interests include computer vision, computer graphics and the relevant mobile applications.

Si Li received the Ph.D. degree from Beijing University of Posts and Telecommunications in 2012. Now she is an Associate Professor in School of Artificial Intelligence, Beijing University of Posts and Telecommunications. Her current research interests include computer vision and natural language processing.

Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Group. Before joining PKU, he did postdoctoral research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University from 2013 to 2016, and worked as a researcher in the National Institute of Advanced Industrial Science and Technology from 2016 to 2017. He won the Best Paper Runner-up award at International Conference on Computational Photography 2015. He has served as an editorial board member of IJCV and an area chair of CVPR/ICCV. He is a senior member of the IEEE.

Yu-Wing Tai is a senior research director of Kuaishou Technology, China (May 2020 - now). He is also an adjunct Professor of Computer Science and Engineering Department of Hong Kong University of Science and Technology (HKUST). Before he joined the current company, he was a research director of Tencent, China (Jan 2017 – April 2020), a principal researcher at SenseTime (Aug 2015 - Dec 2016), and an associate professor of KAIST, S. Korea (Jun 2009 - Aug 2015). His major research area is in Computer Vision and Deep Learning. He has served as an Area Chair of CVPR (’21), ICCV (’11, ’15-’19), and ECCV (’20). He is a senior member of the IEEE.