Face Inverse Rendering via Hierarchical Decoupling

Meng Wang, Xiaojie Guo, *Senior Member, IEEE*, Wenjing Dai, and Jiawan Zhang, *Senior Member, IEEE*

***Abstract*— Previous face inverse rendering methods often require synthetic data with ground truth and/or professional equipment like a lighting stage. However, a model trained on synthetic data or using pre-defined lighting priors is typically unable to generalize well for real-world situations, due to the gap between synthetic data/lighting priors and real data. Furthermore, for common users, the professional equipment and skill make the task expensive and complex. In this paper, we propose a deep learning framework to disentangle face images in the wild into their corresponding albedo, normal, and lighting components. Specifically, a decomposition network is built with a hierarchical subdivision strategy, which takes image pairs captured from arbitrary viewpoints as input. In this way, our approach can greatly mitigate the pressure from data preparation, and significantly broaden the applicability of face inverse rendering. Extensive experiments are conducted to demonstrate the efficacy of our design, and show its superior performance in face relighting over other state-of-the-art alternatives. Our code is available at https://github.com/AutoHDR/HD-Net.git.**

***Index Terms*— Face inverse rendering, face image decomposition, deep learning.**

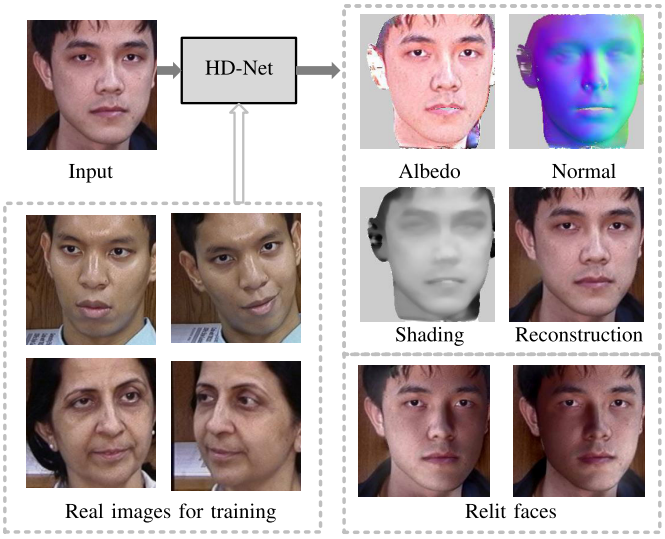


Fig. 1. Illustration of our proposed *HD-Net*, which decomposes real face images into several components. The faces can be relit through changing the lighting conditions.

I. Introduction

FACE inverse rendering can be viewed as a task of disentangling human face images into their albedo maps and shading maps, while the latter ingredient can be further decomposed into two components, *i.e.*, normal and illumination. The problem of face inverse rendering is severely ill-posed in nature, because the number of unknowns to be recovered is multiple times as many as that of the given measurements. A wide spectrum of applications could benefit from FIR, for instance, face relighting in virtual/augmented reality and style transfer, to name just a few.

To ease the ill-posedness, one technical line, with [9], [23], [38], [39] as representatives, seeks help from physical equipment on acquiring face images from different viewpoints and under varying lighting conditions. Though these methods can mitigate the difficulty of decomposition by providing ground-truth information, they demand professional photographing skills/tools and complicated preparation under well-controlled circumstances, significantly limiting the applicability to typical users. For the sake of releasing the professional requirement, several algorithms, *e.g.* [34] and [10], have been designed to learn face components (partially) on synthetic data. Despite the improvement, these methods still suffer from the gap left by the relatively simple distribution of synthetic data to real scenarios. Concerning the drawbacks of the aforementioned methods, it is highly desirable to design an equipment-free and real data-fitted model, which is the goal of this work. However, three main challenges impede moving towards the desire as follows:

1) ***Severe ill-posedness***. Separating an image in the wild into several components (one-to-three decomposition task under the Lambertian model in this work) is in nature heavily under-determined.
2) ***High variation***. The faces often appear to be loosely controlled. Pose and expression affect the appearance of faces in images, which, compared to global 2D changes, will result in higher complexity of the decomposition even in a supervised fashion.
3) ***Unavailable ground-truth***. In practice, it is expensive and complicated, if not impossible, to capture ground-truth information for respective components of real face images. Thus, solving the target problem without supervision seems to be extremely hard.

Manuscript received 18 July 2021; revised 25 June 2022 and 31 July 2022; accepted 8 August 2022. Date of publication 30 August 2022; date of current version 5 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62072327 and Grant 62172295 and in part by the National Key Research and Development Program of China under Grant 2019YFC1521200. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung. *(Corresponding author: Xiaojie Guo.)*

Meng Wang, Xiaojie Guo, and Jiawan Zhang are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: autohdr@gmail.com; xj.max.guo@gmail.com; jwzhang@tju.edu.cn).

Wenjing Dai is with the Department of Technology, Management and Economics Sustainability, Technical University of Denmark, 2800 Lyngby, Denmark (e-mail: weda@dtu.dk).

Digital Object Identifier 10.1109/TIP.2022.3201466

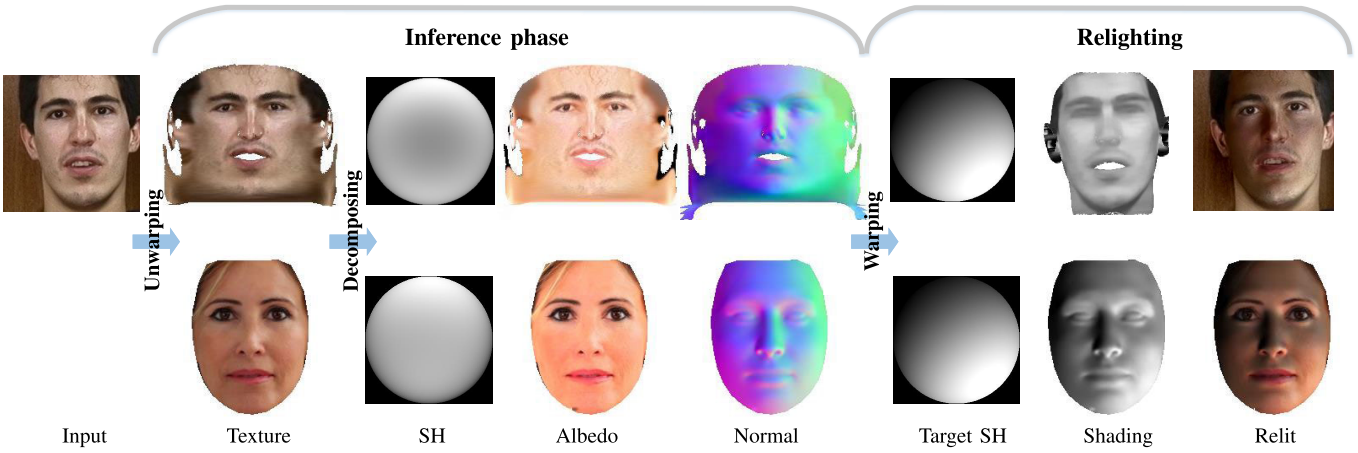


Fig. 2. The procedure for our face inverse rendering and relighting. The input is a single unwarped texture/image, which is then decomposed into lighting (or spherical harmonics, SH), albedo, and normal. For relighting, the decomposed SH lighting is replaced by a target SH provided by users. The shading map is constructed by recomposing the normal and target SH, and the relit shading further drags the decomposed albedo into a newly generated image. The first row tests on non-aligned face image datasets GT [1] with the help of unwarping and warping, and the second row tests on FFHQ [20].

A. Consideration & Contribution

Regarding the ill-posed characteristic, a one-to-three decomposition problem, like mapping a face image into its albedo, normal and lighting components (the target of this work, please see Figure 2), is technically much more difficult than a one-to-two problem, *e.g.* separating a face into its albedo and shading maps (*a.k.a.* intrinsic image decomposition), because the searching space exponentially expands as the number of unknowns to recover increases. To possibly relieve the difficulty, converting a one-to-three decomposition problem into two decoupled one-to-two sub-problems could be a good choice. Fortunately, under the Retinex theory [22] and the assumption of Lambertian reflectance [18], the albedo can play such a pivotal role in decoupling the original problem. Driven by this fact, we design a hierarchical strategy to achieve the goal. More concretely, a face image is firstly disassembled into its albedo and shading maps, then the shading is further decoupled into two elements, *i.e.* normal and lighting.

As for the high variation, thanks to the strong structure of faces, aligning different faces to a canonical status would effectively alleviate the issue. In the literature, a number of face alignment techniques have been proposed, among which the 3D Morphable Model (3DMM) and its follow-ups [42], [43] are arguably the most representative. Even with the above two points being properly disposed, another obstacle is that no ground-truth information is available to guide the training procedure. In other words, effective constraints need to be imposed on the desired solutions. We assume that, after alignment, the albedo and normal maps of the same person should be closely similar (consistency). In addition, the shading map, although it might be diverse for different face images, should be largely smooth.

Based on the above consideration, we customize a deep network to hierarchically decompose face images into three components, including albedo, normal and illumination/lighting. The main contributions of this work can be summarized as follows:

1) For tackling the one-to-three face decomposition, we propose a hierarchical strategy to alternatively solve two one-to-two smaller problems, which significantly reduces the complexity of original problem.
2) We employ the 3D face alignment technique to deal with the high variation of face appearance in terms of pose and expression, which further shrinks the freedom-degree of target space.
3) In an unconstrained setting, simple yet effective constraints, such as the albedo and normal consistency on aligned faces of the same person as well as the piece-wise smoothness on the shading map, are exploited to make the problem tractable.

Extensive experiments are conducted to reveal the effectiveness of our design, and show its superiority over other state-of-the-art methods both quantitatively and qualitatively.

II. Related Work

A variety of inverse rendering methods have been devised over the last decades. This section briefly reviews classic and contemporary techniques closely related to this work.

A. Equipment-Based Solutions

Debevec *et al.* [9] and Sun *et al.* [39] employed specialized light stages to capture reflectance information of human faces. Weyrich *et al.* [48], [49] introduced a face-scanning dome with 16 cameras and 150 light sources to capture sequence with two different exposure settings. Ghosh *et al.* [12], Wang *et al.* [47] and Lattas *et al.* [23] built a setup for multi-view face scanning with several cameras and light sources. Utilized a consistent environment or a specific device to capture images for estimating face appearance has been proposed in [29], [32], [39], and [51]. Though effective, the data they used is captured under a specific environment with professional equipment, and these techniques rely on complicated systems and computing, which can hardly be used for consumer-level usage.

Instead of following the physical equipment-based pipeline, a couple of works benefit from the generalized bas-relief (GBR) transformation (called photometric stereo). The early attempt via shape and surface reflectance decomposition based on the photometric stereo goes to [50]. This manner has been widely applied, especially in normal recovery, such as [4], [5], [7], [8], [15], [17], [35]. Ogun *et al.* [30] proposed a method to seek the surface reflectance without computing surface gradients. Shi *et al*. [35] developed a complete auto-calibration approach for estimating surface normals and albedos. Hauagge *et al.* [15] computed a per-pixel statistic over a stack of images, and combined local geometry at each point with illumination to decompose ambient occlusion, albedo and illumination. In [7], the authors applied color filters in front of 3 LED lights and estimated the surface normal under the constraint of piece-wise smooth albedo.

Recently, Antensteiner *et al.* [2] proposed a network trained with 3 differently colored illuminations to estimate albedos and normals from a single shot image. Hashimoto *et al.* [14] assumed the albedo is constant except for edges and calculated albedo and normal solely from the image sequence. In [13], a set-up with different wavelengths of light sources is used to capture images, and the multispectral photometric stereo and intrinsic image decomposition are adopted to solve the ambiguity of albedo and light. They can achieve the relighting effects, however, the goal of GBR is more like to estimate accurate normal. Besides, surface reflectance is often considered as an intermediate product. Besides, all methods based on photometric stereo need to face the camera and take images from a fixed viewpoint. Our proposed method is not restricted by viewpoints and poses.

B. Supervised and Unsupervised Solutions

Collecting large-scale decomposition datasets with fine labels from the real world is considerable expensive for supervised inverse rendering methods. Alternatively, several works use synthetic data, such as IIW dataset [26] for intrinsic image decomposition of indoor scenes and synthetic face dataset [34] for face inverse rendering based on the physical-based model, to do the job. In [19], a self-supervised method is customized for intrinsic image decomposition with a pre-trained shading model trained with synthetic data. Soumyadip *et al.* [34] learned low frequency from synthetic data. Meanwhile, high-frequency details in real data are captured using shading cues from 'pseudo-supervision'. The reconstruction loss to real data with the pseudo label would not be backward propagated accurately. Besides, the models trained on synthetic data often cannot work well on real face images.

As for unsupervised learning, several works explore the priors of individual components, and build shared-weight networks to decompose each component from multiple images. Lettry *et al.* [25] proposed an intrinsic image decomposition network to extract two local features by using the same weights with the help of unlabeled time-varying sequence images. Ma *et al.* [28] designed a network that requires neither ground truth nor priors. They drew connections between single image-based methods and multi-image-based approaches to show that one can benefit from the other. Besides, they presented a two-stream convolution neural network, which takes a pair of varying illumination images as input. Even though the models of [25] and [28] are powerful, they only fit for intrinsic image decomposition, thus cannot be used for lighting transfer on account of the constraint ability of intrinsic decomposition. Janner *et al.* [19] proposed a network with a shared convolutional encoder and three decoders for reflectance, shape, and lighting, respectively. However, they need the trained model with ground truth as an initialization. With the help of each component's specific physical meaning, Shu *et al.* [37] introduced a weakly unsupervised network by adding additional constraints. However, the ambiguity in the magnitude of lighting leads to unrealistic results. Besides, their results also lose high-frequency details, resulting in poor performance. Recently, some works can produce face decomposition individuals as intermediate components, such as [11], [40], [41], [43]. In contrast, our approach has a targeted approach to face inverse rendering in the wild in an unsupervised manner, obtaining high-frequency components via a hierarchical decoupling network.

III. Hierarchical Decoupling Network

Formally, we denote the input image, albedo, normal, and shading at position p as $I(p)$, $A(p)$, $N(p)$, and $S(p)$, respectively; L is represented as distant lighting. Then the rendering under Lambertian reflectance [18] can be formulated as follows:

$$I(p) = \mathcal{F}_{render}(A(p), N(p), L), \tag{1}$$

where $\mathcal{F}_{render}$ is a physical-based rendering function for reconstructing the input image $I(p)$. For intrinsic image decomposition, the reconstruction is typically written as:

$$I(p) = \mathcal{F}_{recons}(A(p), S(p)) = A(p) \cdot S(p), \tag{2}$$

where the operator $\cdot$ designates the element-wise product. From the above two formulations, it is clear to see that the albedo can be viewed as a bridge between intrinsic decomposition and physical-based inverse rendering.

Following previous works [34], [37], [46], [54], for position p, the normal is $N(p) = [x_p, y_p, z_p]^T$, the lighting L can be expressed as a 9-dimensional spherical harmonics coefficient $l = [l^1, l^2, \ldots, l^9]^T$. Accordingly, the spherical harmonic basis $h_p = [h_p^1, h_p^2, \ldots, h_p^9]^T$ can be represented as:

$$\begin{aligned}
&h_p^1 = \frac{1}{\sqrt{4\pi}}, \quad h_p^2 = \sqrt{\frac{3}{4\pi}} y_p, \quad h_p^3 = \sqrt{\frac{3}{4\pi}} z_p, \\
&h_p^4 = \sqrt{\frac{3}{4\pi}} x_p, \quad h_p^5 = 3\sqrt{\frac{5}{12\pi}} x_p y_p, \\
&h_p^6 = 3\sqrt{\frac{5}{12\pi}} y_p z_p, \quad h_p^7 = \frac{1}{2}\sqrt{\frac{5}{4\pi}} \left(3z_p^2 - 1\right), \\
&h_p^8 = 3\sqrt{\frac{5}{12\pi}} x_p z_p, \quad h_p^9 = \frac{3}{2}\sqrt{\frac{5}{12\pi}} \left(x_p^2 - y_p^2\right).
\end{aligned}$$

On the one hand, the shading S_p can be reconstructed from the normal $N(p)$ and lighting L, represented as:

$$S(p) = \mathcal{F}_{shading}(N(p), L) = h_p^T l. \tag{3}$$

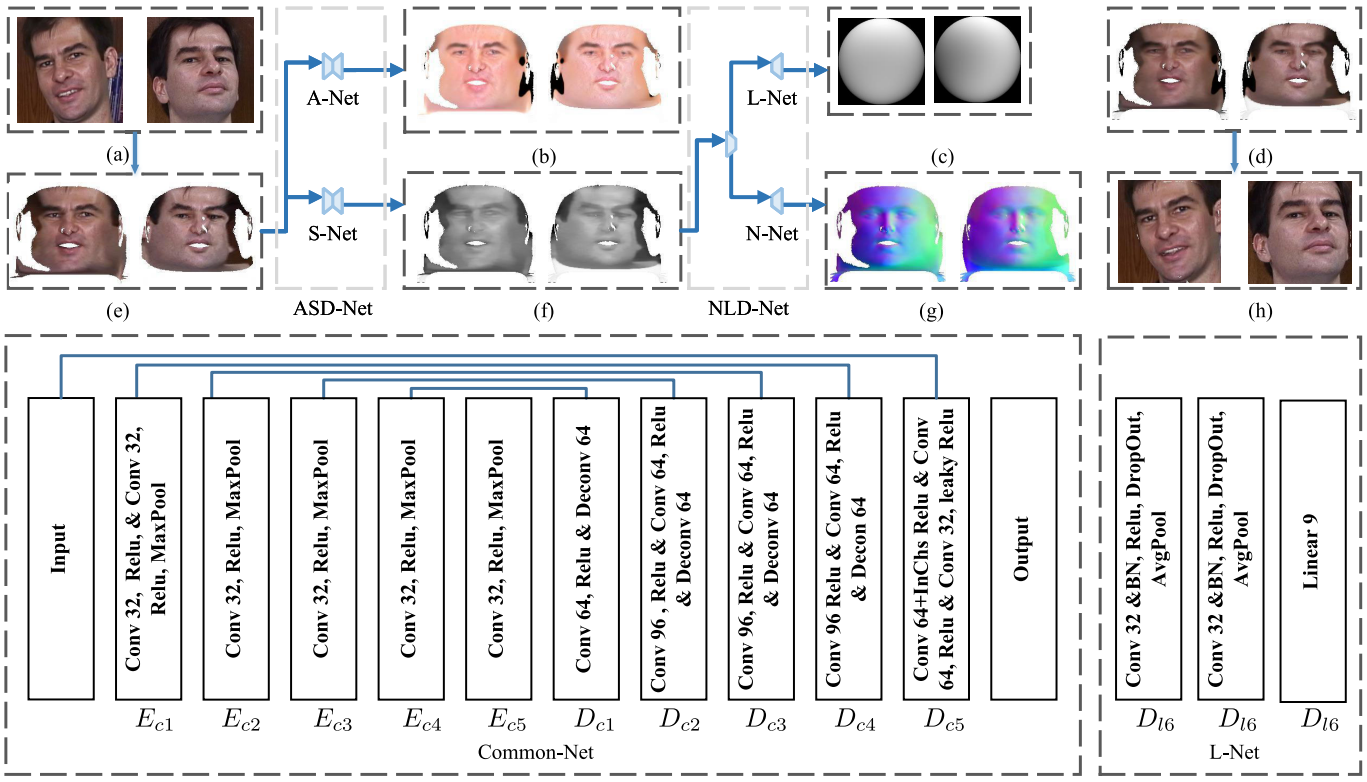


Fig. 3. Overview of our hierarchical decoupling network (*HD-Net*). The network consists of two decoupling sub-nets, *i.e.* the albedo and shading decoupling network (*ASD-Net*) and the normal and lighting decoupling (*NLD-Net*). First, the input images are unwarped, from (a) to (e). *ASD-Net* decomposes the unwarped textures (e) into albedo (b) and shading (f). *NLD-Net* disentangles shading (f) into normal (g) and light (c). Finally, the reconstructed unwarped textures are warped back from (d) to (h).

On the other hand, the lighting information can be simply computed from the predicted normal and shading maps via least square optimization.

A. Network Architecture

As discussed above, we regard the face inverse rendering as a recursive decoupling problem, which converts a one-to-three decomposition into two smaller/simpler one-to-two tasks. It is easy to find refined patterns with a simple, straightforward solution until a certain level of simplicity is achieved. The hierarchical decomposition can be independently applied to selected subsets in a divide and conquer manner, significantly reducing the overall computation cost and the difficulty of adjusting weights in the loss function during training. We emphasize that the goal of this paper is to learn a model that decomposes the unconstrained human face images into three components, namely, albedo maps, normal maps, and lightings. The overall architecture is illustrated in Figure 3.

The pivot of our method is the albedo, which is the bridge between physical-based inverse rendering and intrinsic image decomposition. Therefore, we take advantage of the relationship and build a connection between decoupling networks. Specifically, as shown in Figure 3, the input image I is firstly decomposed into the albedo A and shading S via the albedo and shading decoupling network, denoted as *ASD-Net*, which contains two branches including *A-Net* and *S-Net*. The shading S can be then decoupled into the normal N and light L by the normal and lighting decoupling network (*NLD-Net*), which also contains two branches, called *N-Net* and *L-Net*. The following details the architectures of the two sub-networks, say *ASD-Net* and *NLD-Net*.

The sub-networks of *A-Net*, *S-Net* and *N-Net*, except for the *L-Net* are of the same architecture, called *Common-Net*. The *L-Net* and *N-Net* share the same encoder in *NLD-Net*, the encoded features E_{c5} at the 5th layer of *NLD-Net* are fed into *L-Net* for lighting prediction. We employ the classical U-shaped based network from Noise2Noise [24] as our Common-Net. The main reason is to largely exclude other possible influences from sophisticated network architectures and focus on our proposed strategy. Even with such a simple net, our performance is promising, which reveals the effectiveness of our design and verifies the main claims.

1) Albedo and Shading Decoupling Network: Given an input pair of images, we firstly get the unwarped texture as the input of our *ASD-Net*. As illustrated in Figure 3, The *ASD-Net* takes an unwarped texture of image pair as input and gradually learns the decoupled parameters. *ASD-Net* has two branches, *A-Net* and *S-Net* predicting albedo and shading, respectively. The outputs are albedo map and shading map. Our subdivision process follows a simple update rule, directed by the intrinsic decomposition model as in Eqn. (2).

2) Normal and Lighting Decoupling Network: Based on the shading generated from the first decoupling network *ASD-Net*, we can further apply a subdivision strategy for shading by a second decoupling network *NLD-Net*, which is to decompose

the shading map into normal map and lighting. It also has two sub-networks, *N-Net* and *L-Net* for normal and lighting prediction, respectively.

Tuning parameters could be sensitive and difficult without ground truth. Our hierarchical decoupling network employs a step-wise subdivision strategy to reduce the sensitivity of parameter tuning and facilitate the rapid identification of suitable decomposition parameters.

B. Loss Design

As we do not have the ground-truth components of the input image, additional constraints need to be imposed to guide our hierarchical network. We can utilize the properties of individual components. Please notice that to avoid the disturbance from various expressions and poses of faces, we apply [42] to unwarp the face images for alignment.

1) Albedo and Shading Losses: For a certain person, the albedo should be closely similar or consistent. Thus, the consistency of albedo in paired images can offer a piece of information to constrain albedo learning. While shading is partially influenced by illumination, which should be piece-wise smooth [53]. Considering these two aspects, we introduce the shading smoothness loss in the gradient domain, and the consistent loss on albedo to *ASD-Net* for albedo (*A-Net*) and shading (*S-Net*) prediction, respectively. The loss function $\mathcal{L}_{ASD}$ can be expressed as follows:

$$\mathcal{L}_{ASD} = \lambda_a \mathcal{L}_a + \lambda_s \mathcal{L}_s. \tag{4}$$

Concretely, $\mathcal{L}_a$ is defined as $\|A_i - A_j\|_1$, where $\|\cdot\|_1$ means the ℓ_1 norm. This term regularizes the fidelity between the estimated albedos A_i and A_j from paired images I_i and I_j. The shading map should be piece-wise smooth under a distant light. Similar to [53], $\mathcal{L}_s$ is defined as $\lambda_s(\|\frac{\nabla S_i}{max(|\nabla \hat{S}_i|,\xi)}\|_1 + \|\frac{\nabla S_j}{max(|\nabla S_j|,\xi)}\|_1)$, where ∇S stands for the derivative operator of ∇S_x and ∇S_y in the first order on shading predicted from pair of image, and ξ is a small positive constant (0.01 in this work) for avoiding zero denominator. The non-negative coefficients λ_a and λ_s balance the importance of the corresponding terms.

2) Normal and Lighting Losses: Suppose that we have already had estimated albedo and shading from the *ASD-Net*, it is still hard to disentangle the normal and light due to the ambiguity between them. A number of strategies have been devised for normal estimation, such as [6], [42], [56]. To ease the normal estimation and the lighting, we introduce the estimated normal by 3D Morphable Model [6] as an initialization, which acts as a reference for our estimation. The loss function $\mathcal{L}_{NLD}$ is as follows:

$$\mathcal{L}_{NLD} = \lambda_n \mathcal{L}_n + \lambda_l \mathcal{L}_l, \tag{5}$$

where $\mathcal{L}_n$ is defined as $\|\bar{N}_i - N_i\|_2^2 + \|\bar{N}_j - N_j\|_2^2$ with $\bar{N}_i$ and $\bar{N}_j$ initialized normal maps for coarse training, N_i and N_j are the predicted normal maps from *N-Net*, and $\|\cdot\|_2$ stands for the ℓ_2 norm. Moreover, $\mathcal{L}_l$ adopts $\|l_i - \hat{l}_i\|_1 + \|l_j - \hat{l}_j\|_1$ for regularizing the lighting component. Please note that, under the Lambertian shading model, lights can be represented by 9-dimensional spherical harmonic coefficient vectors. In the term $\mathcal{L}_l$, $\hat{l}_i$ and $\hat{l}_j$ are computed from the predicted shading maps and initialized normal maps using least square optimization, and l_i and l_j are predicted lighting from *L-Net*.

3) Reconstruction and Adversarial Losses: The reconstruction loss function $\mathcal{L}_{rec}$ contains two terms. One is about image reconstruction via $\mathcal{L}_{Irec}$, and the other takes care of shading reconstruction by $\mathcal{L}_{Srec}$, which can be represented as follows:

$$\mathcal{L}_{rec} = \lambda_{Irec}\mathcal{L}_{Irec} + \lambda_{Srec}\mathcal{L}_{Srec} + \lambda_{adv}\mathcal{L}_{adv}. \tag{6}$$

The image reconstruction loss $\mathcal{L}_{Irec}$ is under the intrinsic decomposition model, while the shading reconstruction loss $\mathcal{L}_{Srec}$ is based on the physical-based shading reconstruction computed from the predicted normal and lighting. More specifically, $\mathcal{L}_{Irec} = \|I_i - A_i \cdot S_i\|_1 + \|I_j - A_j \cdot S_j\|_1$, where I_i and I_j are paired input images, A_i and A_j, and S_i and S_j are the albedo maps and shading maps predicted from *ASD-Net*. In addition, $\mathcal{L}_{Srec}$ is given as $\|S_i - \hat{S}_i\|_1 + \|S_j - \hat{S}_j\|_1$ with $\hat{S}_i$ and $\hat{S}_j$ are the shadings reconstructed by the predicted normal and lighting as in Eqn. (3). In this way, along with the reconstruction of the input, the shading reconstruction loss minimizes the gap between intrinsic image decomposition and physical-based rendering. Besides, the discriminator is frequently used in unsupervised learning to distinguish which ones are real from fake ones. In our hierarchical decoupling network, we introduce an adversarial loss $\mathcal{L}_{adv}$ as in [36] for the reconstruction under the Lambertian model, which can further guarantee the reasonable reconstruction.

C. Training Strategy

To learn each component, we introduce a rough initialization and subdivision strategy into our hierarchical network. This has the potential to improve the accuracy of predicted components significantly. In other words, we divide the training process into two stages.

1) Training Stage 1: The parameters are randomly initialized at the start of the training phase; it is not possible to decompose individual components without normal initialization. For this reason, we introduce coarse normals for initialization in order to guide the normal learning; otherwise, the initialization normals also conduct the light predictions with the help of predicted shading.

2) Training Stage 2: When the training of the network is converged in stage 1, our model is considered to be able to decompose the albedo, shading, normal, and lighting, effectively. However, due to the effects of coarse initialization, some errors will be backpropagated (*e.g.*, the loss propagation between the predicted light and the light computed from the initial normal maps and predicted shading maps via least-square optimization). As a result, the model training might be inaccurate. Thus, a few adjustments are made to refine the result, including 1) we fix the parameters of *A-Net*, *S-Net* and *N-Net* except for the decoder of *L-Net*; 2) we replace the initialization normal with the predicted normal to refine light; and 3) we train our hierarchical decoupling network with a lower learning rate. After several iterations, our network can gradually close the gap.

D. Implementation Details

The entire hierarchical decoupling network is schematically illustrated in Figure 3, which contains two sub-networks *ASD-Net* and *NLD-Net*. Specifically, the input is first passed through a $32 \times 3 \times 3$ convolutional layer without pooling, and then the obtained features are fed into *A-Net* and *S-Net* for albedo and shading prediction, respectively. Following that, the predicted shading is employed by *N-Net* and *L-Net* to predict normal and lighting. The network is trained on images with 256×256 size. For the GT dataset [1], we use a mask to expand the unwarped pair of images to 256×256 for training and testing, due to the aligned face images generated from unwarped function [42] is $192 * 224$. In the training phase, we set $\lambda_a = 0.25$, $\lambda_s = 0.1$, $\lambda_n = 0.5$, $\lambda_l = 0.01$, $\lambda_{Irec} = 0.25$ and $\lambda_{Srec} = 0.01$. Besides, an adversarial loss with $\lambda_{adv} = 0.001$ of the input face is added to reduce the reconstruction error. As the different distributions of different datasets influence the weights of loss terms, we set $\lambda_s = 0.01$, $\lambda_{Irec} = 0.25$, $\lambda_a = 0.15$ and $\lambda_{adv} = 0.0001$ for the DPR dataset [54].

Our training uses 0.001 as the learning rate for training stage 1, and 0.0001 for training stage 2 both with a batch size of 8, and the optimizer is Adam. We conducted all the experiments on a platform with CPU i5-9400F, 16G RAM and a single NVIDIA GPU with 11GB of RAM. The results provided in our paper are generated from a model spending about 250K/30K training iterations at the first/second stage. To better exhibit the effect of re-rendering, we utilize a Poisson blending algorithm [31] to combine the rebuilt faces with the backgrounds.

IV. Experiments

In this section, we evaluate our proposed method and compare it with state-of-the-art competitors on 5 different datasets including DPR [54], FFHQ [20], Photoface [52], CelebA [27] and Georgia Tech (GT) [1]. DPR [54] contains 138135 relit images that are generated by CelebA [27] under various lighting conditions (including harsh lighting). It is used to verify our method can apply the extreme condition. FFHQ [20] consists of 70,000 high-quality images with a range of age and ethnicity. It can be used to put our algorithm through its paces on a larger scale. Photoface [52] contains images of the same person under different lighting conditions, which we use to assess the ability in normal and albedo prediction. We also test our method on real unconstrained face images from the GT dataset [1], which has multiple pictures of the same person with different facial expressions, lighting and poses. Furthermore, FFHQ [20] and CelebA [27] do not have paired images, and we test the generalization capability of our method using a model trained on DPR [54]. In order to evaluate our method quantitatively and qualitatively, we also generate synthetic paired image data by randomly selecting of DPR lighting [54], normal and albedo from SfSNet [34].

A. Comparison With State-of-the-Art Methods

1) Evaluation on Synthetic Data: To quantitatively compare our model with prior works, we generate synthetic paired image data by randomly selecting lights from the DPR dataset [54], normal and albedo from the SfSNet dataset [34]. Table I reports the numerical results of our reconstruction and predicted albedo with the state-of-the-art inverse rendering method SfSNet [34]. The metrics used to measure the predicted albedo and reconstruction adopts MAE and RMSE. Please notice that SfSNet [34] is trained with the ground truth information, while ours is trained with normal initialization. As can be seen from the table, the MAE and RMSE values of our method are close to those of SfSNet [34] in terms of reconstruction error and albedo error. We again emphasize that our approach is able to decompose the desired components sufficiently well despite the fact that we do not train using labeled data. We provide several visual comparisons in Figure 4. From the figure, it is clear to see that our results can properly decompose the normal maps while recovering reasonable albedo maps from the inputs.

TABLE I

Face Image Reconstruction and Albedo Reconstruction Comparison Between Our Method and SfSNet [34] on Synthetic Data in Terms of MAE and RMSE

Method	Recon. Error		Albedo Error	
	MAE	RMSE	MAE	RMSE
SfSNet	0.003	0.007	0.019	0.027
Ours	0.014	0.021	0.051	0.071

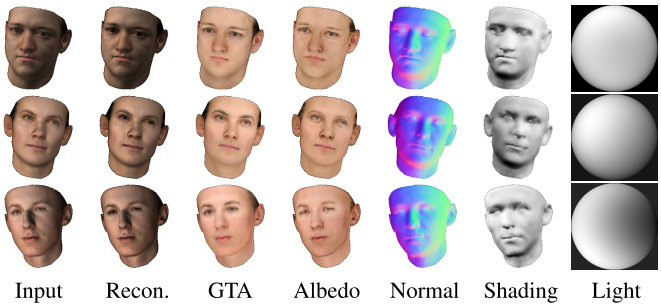


Fig. 4. Results on synthetic data. We compare our results with the ground truths. From left to right, they are the inputs, reconstruction faces (Recon.), ground truths of albedo maps (GTA), predicted albedo maps, predicted normal maps, shading maps and predicted light, respectively.

2) Evaluation of Normal Estimation: We compare the quality of our estimated normals with the state-of-the-art methods, which recover from a single image. Since we do not know the split of the training dataset, we select all faces of the same person and make the permutation of images to constitute pair of images without repetition. Then we randomly split the data for our training and testing. The evaluation is the mean angular error and percentage of pixels under angular error thresholds [45]. For a fair comparison, we also train our model on real face images. The compared model has been trained with a mixture of synthetic data and FFHQ dataset [20]. From Table II, the mean and std of predicted normal by 'Ours' (not trained on the Photoface dataset [52]) outperform those of the others, we also show the normal error maps comparison in Figure 5. Figure 5 and Table II show that our predicted normal is slightly inferior to the compared methods on angle error below 20°, while our model clearly outperforms the competitors when the angle error is above 25°. This is because coarse 3DMM normal initialization may misguide our normal

TABLE II

NORMAL COMPARISON RESULTS ON PHOTOFACE DATASET [52]. THE DATA COMES FROM SFSNET [34]. LOWER IS BETTER FOR MEAN ERROR, AND HIGHER IS BETTER FOR THE PERCENTAGE OF CORRECT PIXELS AT VARIOUS THRESHOLDS

Method	Mean ± std	$< 20°$	$< 25°$	$< 30°$
3DMM [6]	26.3±10.2	4.3%	56.1%	89.4%
Pix2Vertex [33]	33.9±5.6	24.8%	36.1%	47.6%
SfSNet [34]	25.5±9.3	43.6%	57.5%	68.7%
Ours	**23.1±8.9**	38.6%	54.6%	**72.3%**
Marr Rev. [3]	28.3±10.1	31.8%	36.5%	44.4%
UberNet [21]	29.1±11.5	30.8%	36.5%	55.2%
NiW [44]	22.0±6.3	36.6%	59.8%	79.6%
SfSNet-ft [34]	12.8±5.4	**83.7%**	90.8%	94.5%
Ours-ft	**11.2±9.7**	80.5%	**91.1%**	**96.4%**

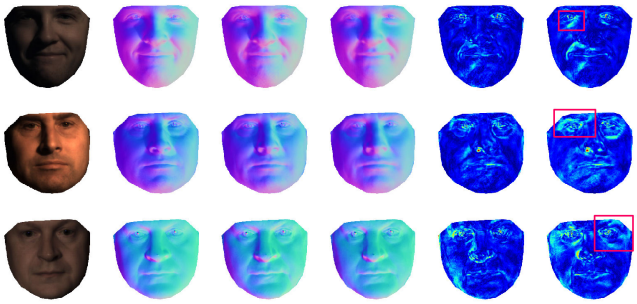


Fig. 5. Normal comparison with SfSNet [34] on the Photoface dataset [52]. From left to right, they are the inputs, ground truth normal maps (GTN), our normal maps (OurN), SfSNet normal maps (SfSNetN), our normal error maps (OurE) and SfSNet normal error maps (SfSNetE).

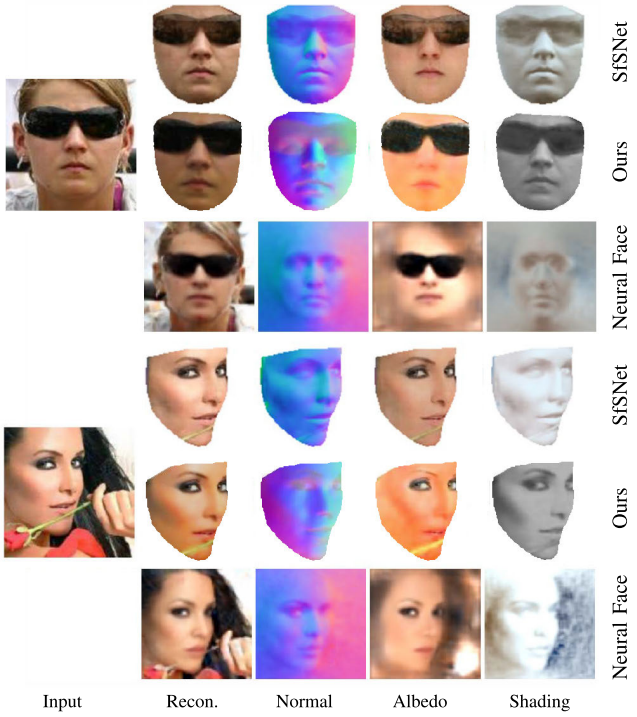


Fig. 6. Inverse rendering comparison with the state-of-the-art methods. Our results compare with the SfSNet [34] and Neural Face [37] on the data showcased by the authors. We outperform the unsupervised method Neural Face [37], which regards as the baseline. It should be noted that our model does not train on the CelebA dataset [27] due to a lack of paired images. Instead, we employ the model trained on the DPR dataset [54].

estimation when angles are small in the early steps. When our model acquires a reasonable decomposition ability, we refine our model with the predicted normal instead of initialization.

Similar to SfSNet [34], we also apply the ground truth normal from Photoface dataset [52] to refine our model (named 'Ours-ft') and use the refined model to compare with 'SfSNet-ft'. More specifically, we replace the initial normal map with the ground truth normal from the Photoface, and train the model until convergence. The mean and std error of our normal is more accurate than SfSNet [34] from the last row in Table II. This is because our training uses paired images, which have a stronger constraint than a single. During the test, our model still retains a strong ability. The standard deviation of our result is higher than SfSNet [34]. We deem that our method can accurately predict the direction of normal in most cases without using labeled data during training. However, in some data (e.g., image pairs with a great difference between illumination), it cannot accurately estimate albedo, leading to poor shading, further influencing the final normal prediction. As a result, the outlier data make our standard deviation bigger than the labeled data trained method, SfSNet [34].

3) Evaluation of Light Estimation: For lighting evaluation, SfSNet [34] used 27-dimensional spherical harmonic coefficient vectors as their light. Following DPR [54], the predicted lighting of our model is 9-dimensional spherical harmonic coefficient vectors. The lighting comparison with MAE/RMSE is unfair under different dimensions. For this reason, we regard lighting evaluation as a classification problem, as LDAN [55] and SfSNet [34] did. Similar to SfSNet [34], we evaluate the accuracy of estimated lighting with the accuracy of lighting classification. We compare the synthetic data after light clustering by K-Means. Specifically, we use K-Means to give lights to cluster 10 classes and then compare the lighting classification correctness. Ours accuracy is 89.46%, and SfSNet [34] is 94.32%. The comparisons show that our method is slightly less effective than the supervised method, SfSNet [34]. It is quite usual that we have a few percentage points less than SfSNet [34].

4) Inverse Rendering Comparison: In Figure 6, we show our results compared with the state-of-the-art inverse rendering methods on CelebA [27]. It can be seen that our method can produce high-frequency albedo maps and normal maps than 'Neural Face' [37], which is regarded as the baseline. Compared with SfSNet [34] trained with synthetic data and real data, our results were able to produce similar results to theirs. Significantly, the distribution gap between CelebA [27] and DPR [54] leads to relatively poor reconstruction results of our method, but it shows that our model has a general ability to disentangle face images.

In Figure 7, we show that our model can be applied to a broader range of scenarios, such as a face with glasses and scarves. This is because the coarse normal is simple without occlusion, and our model learns an ability to eliminate the occlusion parts.

5) Relighting Comparison: Figure 8 shows the portrait relighting results compare with the state-of-the-art methods SMFR [16], DPR [54], SIPR [39] and SfSNet [34] on

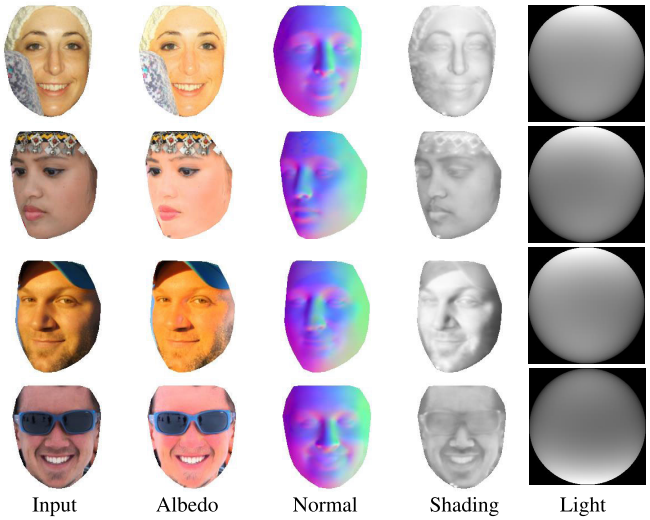


Fig. 7. Inverse rendering results on FFHQ [20] with occlusions, such as sunglasses and scarf.

FFHQ [20]. Compared to SMFR [16], our results produce a more realistic lighting effect on the faces because the light on face changes gradually to appear realistic without high-contrast shadows. As we all know, Single-sided glare increases the intensity of ambient light, which in turn affects the effect of light action. The high-contrast shadows on face only happen in the photography studio. Therefore, it would not be easy to see high-contrast shadows on face in the nature under the strong bright light. Furthermore, our model is a physical-based inverse rendering model, while the shadows produced by SMFR [16] is controlled by thresholds, which is the peak at the center and smoothly decays with the distance. An incorrect threshold will lead to unrealistic results. Give an example of shadows in the third row of the images, SMFR [16] concerns the high-contrast shadow effects on the nose. They do, however, disregard shadow effects on the lower lip. This erroneous lighting could be caused by the soft-shadow thresholds that are outside of the range. The outside thresholds will not work properly and will result in misjudgment results. The physical-based inverse rendering model based on lighting and geometry can accurately show the light effects on geometry. Our physical-based inverse rendering model can accurately calculate the interaction between lighting and normal. As a result, our relit faces can portray the shading more fully to provide realistic cast shadow effects. We can see from the comparison that our relighting outcomes are more natural and realistic results, as shown in Figure 8. In terms of light intensity and shadows, our illuminated faces outperform others.

Furthermore, we provide several results on the real face dataset, *i.e.* the GT dataset [1]. Figure 9 shows the results of our method on unconstrained real face images. The predicted albedo maps, normal maps, and lighting can reconstruct the original images and ensure that the reconstructed images obtain high-frequency details. However, unwarping and warping processes [42] will lose the face details due to the bilinear sampling related operation. Considering that the shading should be piece-wise smooth, we assume the loss caused by warping and unwarping [42] goes to the shading map. Thus, we use the input face I and predicted shading S to compute the albedo map and regard the computed albedo maps as the final output to relight a new face. More specifically, the final albedo map can be computed by $A = I/S$ under the intrinsic decomposition model. And we regard the A as the final albedo maps for relighting. In Figure 9, It is obvious that our method can predict the albedo maps with high-frequency. On the other hand, the high-frequency relit faces under a new light with the natural cast shadows demonstrate that our method is capable of decomposing each component accurately.

In addition, we present numerous face decomposition results on DPR [54] and FFHQ [20] in Figure 10. We also merge these decomposed components with the new light to obtain the new relit faces, as shown in Figure 10. We can see that our technique can properly estimate each component; *e.g.*, the predicted shading maps demonstrate the correctness of our predicted light. Besides, the new relit faces under the target SH can produce natural and realistic cast shadows. The decomposition components by the effect of new light can recompose natural cast shadows, indicating that our method is proper.

6) Runtime Comparison: The official codes of Neural Face [37] and SfSNet [34] are Lua and Matlab, respectively. A direct comparison is not fair. For this reason, we only compared with an unofficial implementation of Pytorch-based SfSNet[1] [34]. Without code optimization, the mean prediction of our model and SfSNet [34] are 59ms and 41ms with a batch size of 8.

B. Inverse Rendering Applications

In this section, we show several applications of inverse rendering. The instinctive applications are as follows:

1) Relighting: Figure 11, we show a relighting application our method. It is worth mentioning that the seventh row of the face has been graffitied, our approach is still capable of properly estimating the illumination information, and the re-rendered images look natural and striking. On the other hand, it shows that our method can be applied to faces with different skin tones and accurately decompose the lighting and albedo.

2) Light Transfer: We also exhibit light transfer results in Figure 12, where we apply the estimated light from the 'source image' to the 'target image'. The natural and realistic transfer results illustrate the accuracy of our decomposition.

3) De-Lighting: The accuracy of face detection is sensitive to the light on face images. We believe that another application is de-lighting (as shown in Figure 13) to improve the accuracy of face detection. Thus, we randomly selected 60,000 face images from the extreme lighting face dataset, DPR [54], to evaluate the actual application of our model. The accuracy of face detection with dlib face detector[2] is 94.59%. After that, we utilized the dlib face detector to detect face albedo maps generated from our model, and the accuracy increased to 97.85%.

[1]https://github.com/bhushan23/SfSNet-PyTorch
[2]http://dlib.net/

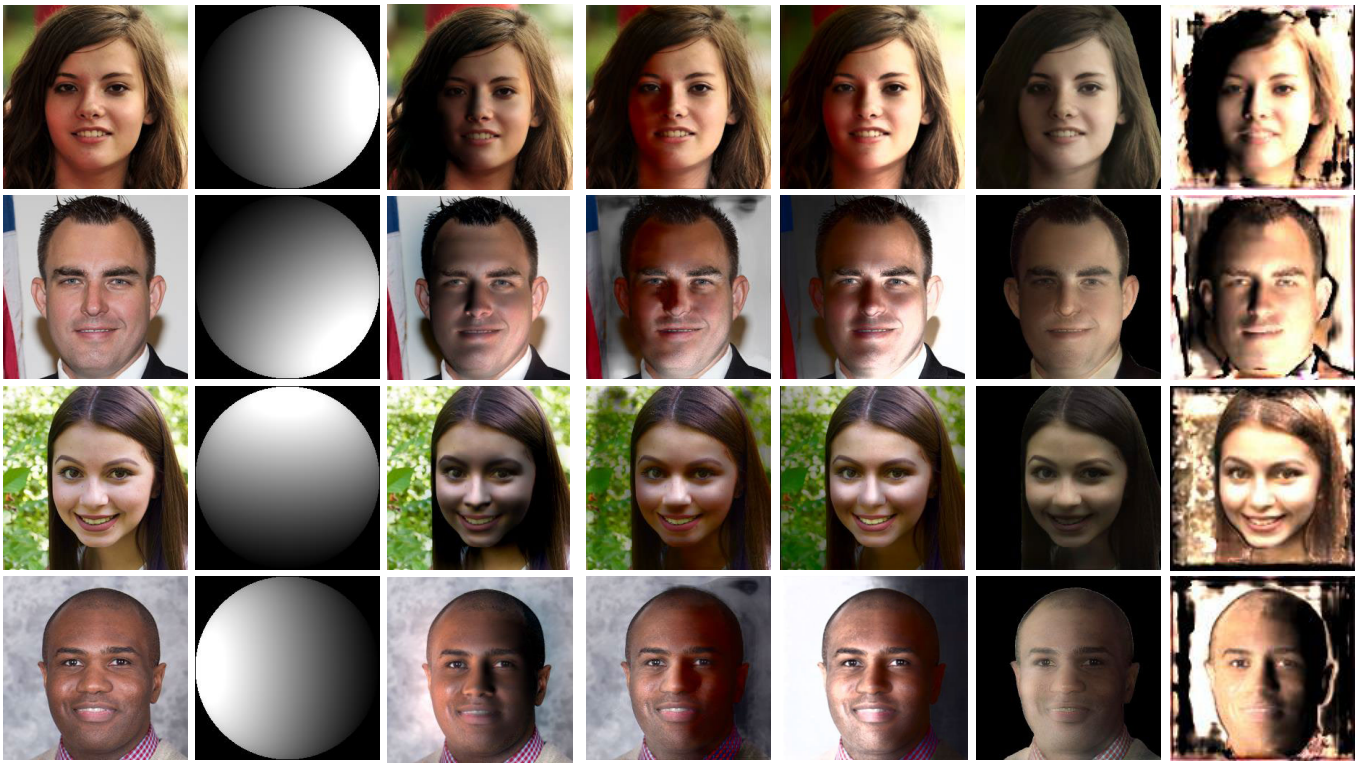


Fig. 8. Relighting comparison with the state-of-the-art methods. Our results compare with the data provided by SMFR [16] in qualitative on FFHQ [20]. Our model produces more natural and real cast shadows than prior works, especially around the nose, lip, and eyebrow.

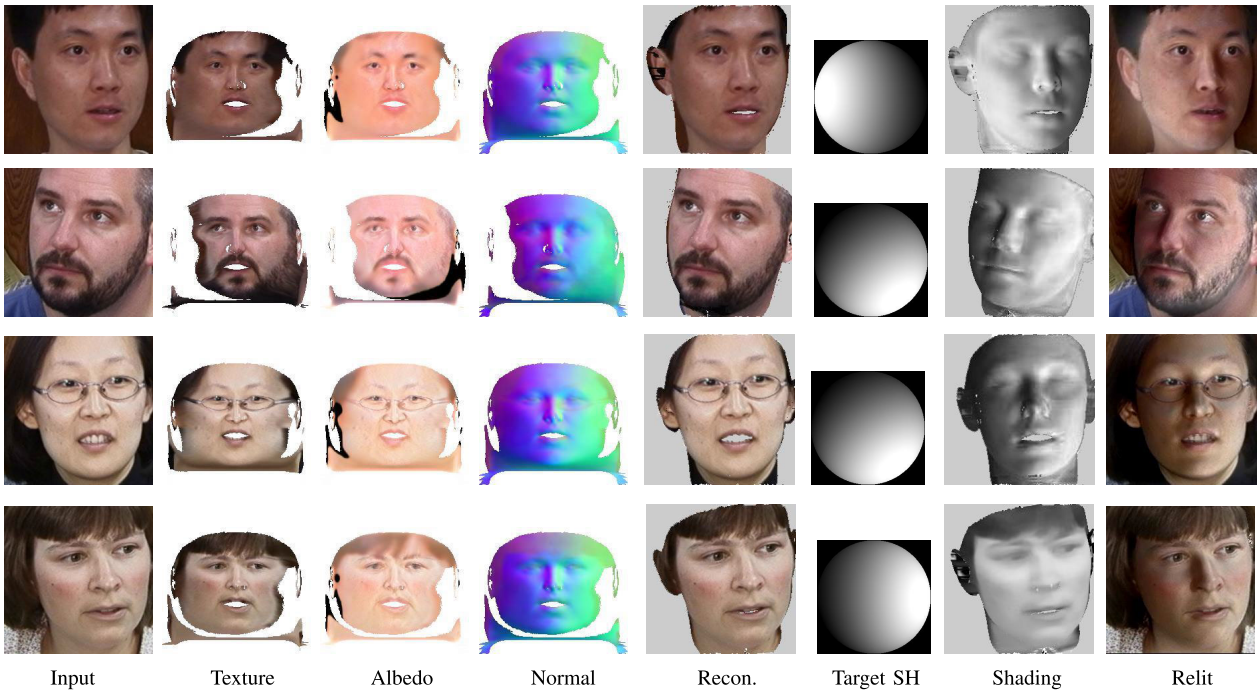


Fig. 9. Decomposition and relighting results on real unconstrained images from the GT dataset (GT) [1]. Our relit faces keep local facial details with natural cast shadows under the target SH, particularly around the nose.

V. Ablation Study

In the following, we perform several ablation studies to explore different aspects of our approach in more detail.

A. Ablation of Network Architectures

To demonstrate the effectiveness of our hierarchical decoupling network, we train two additional models based on

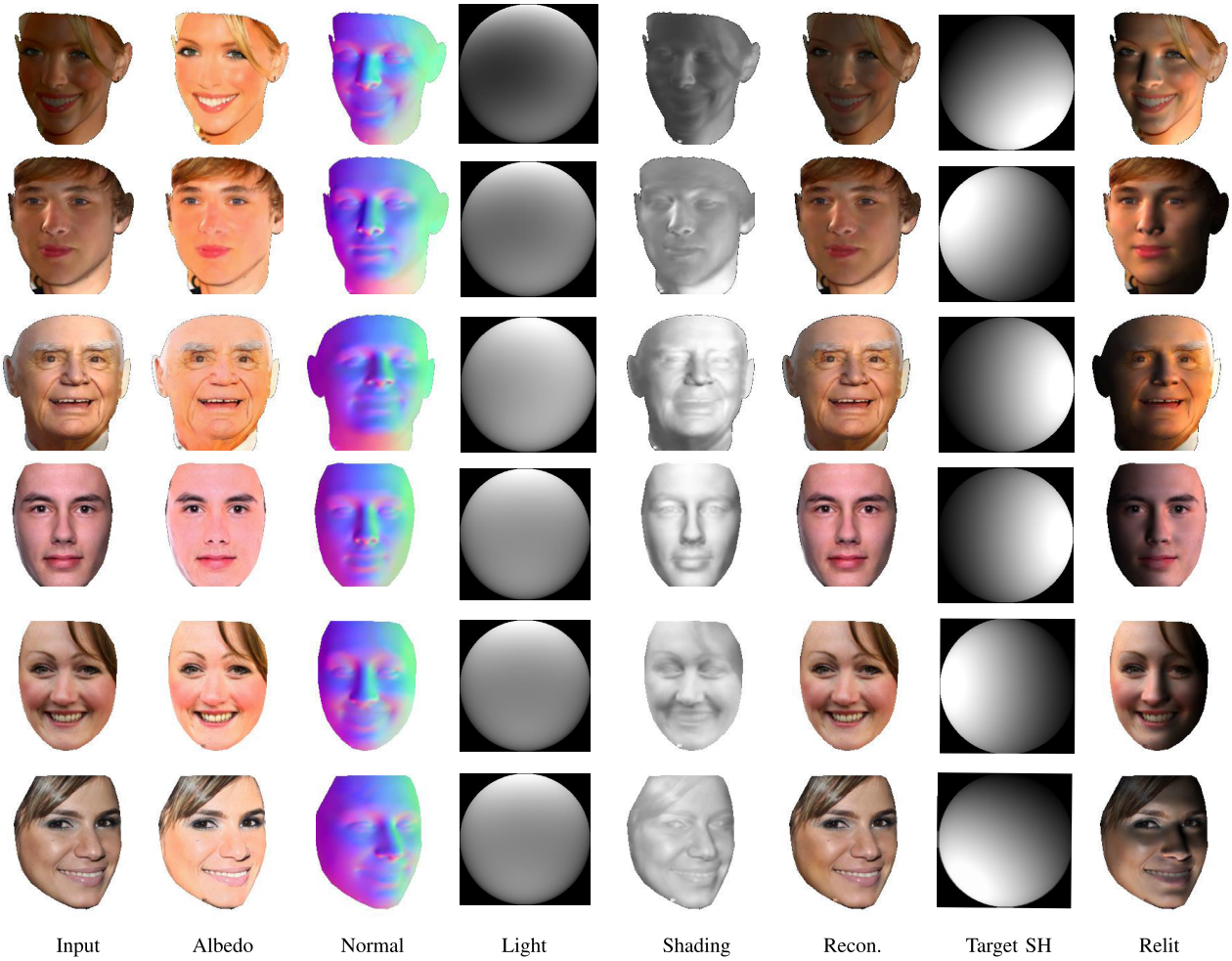


Fig. 10. Decomposition and relighting results on DPR [54] (first three rows) and FFHQ [20] (last three rows). Our albedo maps and normal maps with the target SH can relight new faces. The first three lines are DPR [54] results, and the last three lines are FFHQ [20] results.

TABLE III
NORMAL AND ALBEDO (MAE/RMSE) EVALUATION ON PHOTOFACE DATASET [52] OF DIFFERENT NETWORK ARCHITECTURES

Setting	Mean ± std	$< 20°$	$< 25°$	$< 30°$	MAE	RMSE
(a)	8.9 ±11.6	87.4%	93.5%	96.5%	0.051	0.071
(b)	9.0±12.0	86.4%	93.1%	96.3%	0.053	0.074
(c)	10.1±13.4	81.9%	90.0%	94.3%	0.056	0.083

Lehtinen *et al.* [24] (see Figure 14). The results are shown in Table III. From the table, it can be seen that the accuracy of normal and individual albedo components predicted by *HD-Net* (a) is much better than the other two ((b) and (c)). The comparisons confirm the effectiveness of our hierarchical design. In addition, the number of trainable parameters is used to evaluate the complexity of a model. For neural networks, the higher the number of parameters, the more complex the model. Here, we report the complexity of our model with the number of trainable parameters of our model. The models of (b) and (c) have approximately 6.8M parameters and 7M parameters, whereas our *HD-Net* model has about 5M parameters, which is pretty small. We can achieve better results with a smaller model than the other two network architectures. When the batch size is 8, our *HD-Net* only requires 6.3GB of GPU memory on $3 \times 256 \times 256$ inputs.

To ease the one-to-three decomposition task, we employ a similar network architecture from [53] for our shading prediction. The benefits of splitting shading prediction network *S-Net* are twofold: one is to speed up the convergence of the network and reduce the training time, and the other is that the predicted shading guides our normal and lighting estimation.

B. Ablation of Network Detaching

Table IV shows the normal reconstruction error about different settings of network detaching. As can be seen from the table, detaching light (w/ DL) is able to predict a more accurate normal map, although the MAE and RMSE of albedo increase. We consider that the optimal model is the predicted normal with a smaller error. For there is no real albedo for constraint, we assume that the albedo has converged as long as the two albedo maps are consistent. In addition, it can be seen that detaching light is able to reach the convergence state faster than the other two settings in Figure 15. Obviously, detaching light is the best setting for *HD-Net*.

Fig. 11. More relighting results under varied lighting conditions by our method on samples from the FFHQ [20]. The input and 7 relit faces are shown from left to right. (Best viewed by zooming in.)

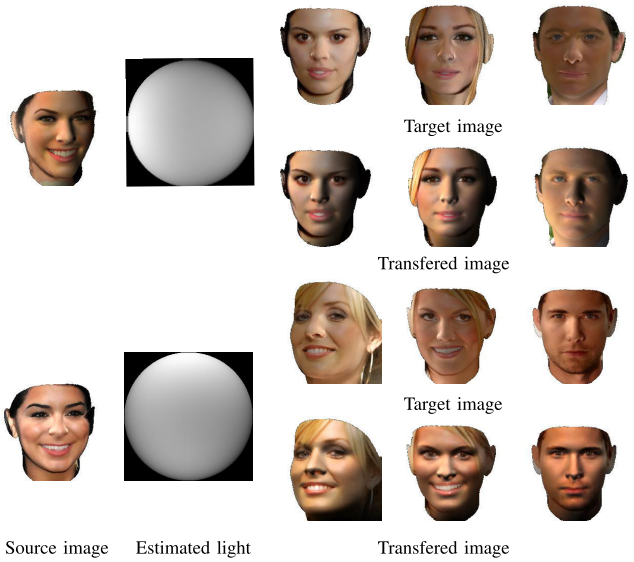


Fig. 12. Light transfer results from source images to the target images from the DPR dataset [54].

C. Ablation of Loss Functions

To demonstrate the effectiveness of our loss design, we train an additional model without adversarial loss $\mathcal{L}_{adv}$ (w/o AD) and with a normal consistent loss (w/ NC) between N_i and N_j. As can be seen from Table VI, our proposed model is able to produce significantly better results for albedo and normal prediction than the models trained without $\mathcal{L}_{adv}$ and with a normal consistent loss. Our model outperforms the model without $\mathcal{L}_{adv}$ and with normal consistent loss. The normal consistent loss can be regarded as a prior during training. However, it would increase the complexity of calibrating weights for balancing the influence among loss functionss. As a trade-off, we forego incorporating normal consistent loss.

Fig. 13. Face de-lighting on DPR dataset [54]. Dlib face detector fails in face detection on the input images (Input). It can be a success after de-lighting (De-light) with our decomposition model.

TABLE IV

NORMAL AND ALBEDO (MAE/RMSE) EVALUATION ON PHOTOFACE [52] WITH/WITHOUT DETACH. W/O D, W/ DLS AND W/ DL STAND FOR TRAINING NETWORKS WITHOUT PREDICTED SHADING AND COMPUTED LIGHT DETACHING, WITH PREDICTED SHADING, AND COMPUTED LIGHT DETACHING AND WITH COMPUTED LIGHT DETACHING, RESPECTIVELY

Setting	Mean ± std	$< 20°$	$< 25°$	$< 30°$	MAE	RMSE
w/o D	9.1±11.8	86.8%	93.1%	96.1%	0.053	0.079
w/ DLS	10.7±16.1	73.0%	84.9%	91.5%	0.046	0.068
w/ DL	8.9±11.6	87.4%	93.5%	96.5%	0.051	0.071

TABLE V

NORMAL AND ALBEDO (MAE/RMSE) EVALUATION ON PHOTOFACE [52] OF DIFFERENT TRAINING STRATEGIES. (a) AND (b) ARE TRAINING WITH TWO STEPS AND WITH AN END-TO-END WAY, RESPECTIVELY

Setting	Mean ± std	$< 20°$	$< 25°$	$< 30°$	MAE	RMSE
(a)	8.9 ±11.6	87.4%	93.5%	96.5%	0.051	0.071
(b)	9.2±12.1	87.3%	93.1%	95.6%	0.057	0.081

TABLE VI

NORMAL COMPARISON ON PHOTOFACE DATASET [52] WITH DIFFERENT LOSS SETTINGS

Setting	Mean ± std	$< 20°$	$< 25°$	$< 30°$	MAE	RMSE
w/ NC	10.2±13.2	82.6%	90.2%	94.4%	0.052	0.077
w/o AD	9.5±12.1	86.2%	92.4%	95.6%	0.145	0.183
Ours	8.9 ±11.6	87.4%	93.5%	96.5%	0.051	0.071

D. Ablation of Training Strategies

Table V shows the normal reconstruction and albedo error about different training strategies, (a) and (b). In addition, we also evaluate lighting classification correctness with different training strategies. The lighting classification correctness of training in two steps (a) and training with an end-to-end way (b) are 89.46% and 87.23%, respectively.

E. Ablation of Loss Parameter Settings

There are seven loss items to control the performance of *HD-Net*, as can be seen in Table VII, Table VIII and Figure 16.

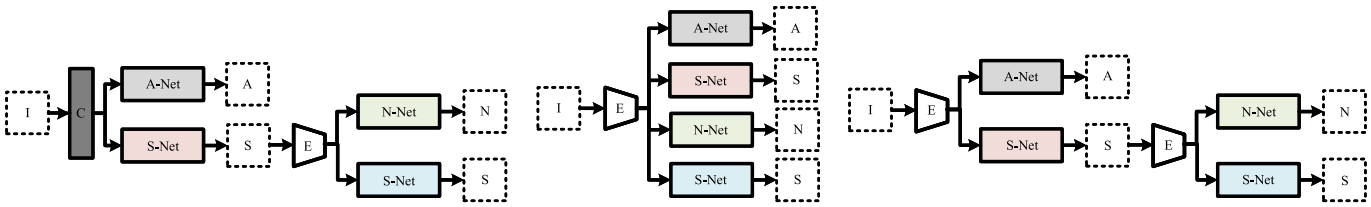


Fig. 14. Three different decoupling network architectures. (a) Our proposed hierarchical decoupling network (*HD-Net*), (b) An autoencoder network and (c) A share-encoder-based hierarchical decoupling network. The letters *I, C, E, A, S, N*, and *L* represent the input, convolution layer, encoder, albedo, shading, normal and lighting, respectively.

TABLE VII

DIFFERENT WEIGHT SETTINGS OF LOSS ITEMS

Setting	λ_{Irec}	λ_s	λ_{Srec}	λ_a	λ_n	λ_l	λ_{adv}
1	0.25	0.1	0.01	0.25	0.25	0.01	0.001
2	0.25	0.1	0.01	**0.1**	**0.5**	0.01	0.001
3	0.25	0.1	**0.1**	0.25	**0.5**	0.01	0.001
4	**0.5**	0.1	0.01	0.25	**0.5**	0.01	0.001
5	0.25	0.1	0.01	0.25	**0.5**	0.01	0.001

TABLE VIII

NORMAL AND ALBEDO (MAE/RMSE) EVALUATION ON PHOTOFACE DATASET [52] OF DIFFERENT WEIGHT SETTINGS

Setting	Mean ± std	$< 20°$	$< 25°$	$< 30°$	MAE	RMSE
1	13.9±22.8	50.2%	65.7%	77.0%	0.079	0.119
2	12.0±17.1	83.3%	90.4%	94.3%	0.109	0.149
3	10.8±14.8	76.9%	86.2%	92.1%	0.088	0.117
4	10.2±12.9	83.9%	90.3%	94.6%	0.082	0.107
5	8.9 ±11.6	87.4%	93.5%	96.5%	0.051	0.071

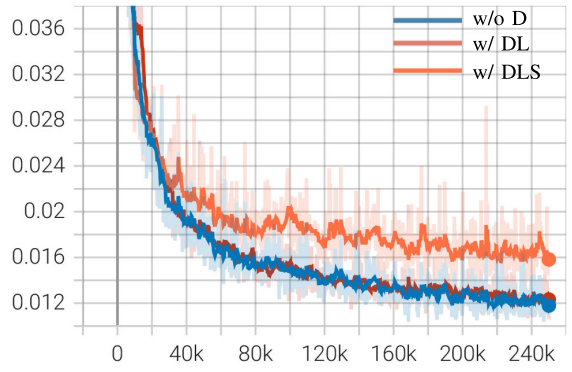


Fig. 15. The normal loss of different detaching settings. w/o D, w/DLS and w/DL stand for training networks without predicted shading and computed light detaching, with predicted shading and computed light detaching, and with computed light detaching, respectively.

In Figure 16, for example, Setting 1 is able to decompose the individual components, but the normal estimation deviates greatly. The predicted albedo losses some details in Setting 2, the predicted normal contains some light in Setting 3 and the reconstruction face loses the original appearance in Setting 4. With suitable weight parameters, our network is able to accomplish the task, which can be found in Setting 5.

F. Ablation of Gaussian/Salt-Pepper Noises

We show the performance of our model trained on the face with Gaussian noises ($\sigma = 0.1$) or salt-pepper noises ($SNR = 0.01$) to evaluate the robustness of our method.

TABLE IX

NORMAL AND ALBEDO (MAE/RMSE) EVALUATION ON PHOTOFACE DATASET [52] WITH GAUSSIAN AND SALT-PEPPER NOISES

Noise	Mean ± std	$< 20°y$	$< 25°$	$< 30°$	MAE	RMSE
GN	9.9±12.8	86.3%	91.6%	95.3%	0.076	0.104
SN	12.4±18.9	61.8%	76.3%	85.9%	0.081	0.116
Ours	8.9±11.6	87.4%	93.5%	96.5%	0.051	0.071

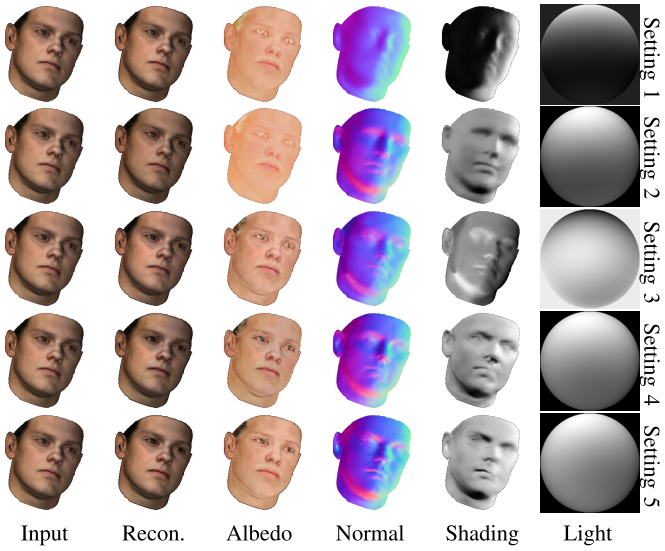


Fig. 16. Visualized results with different weight settings of loss items.

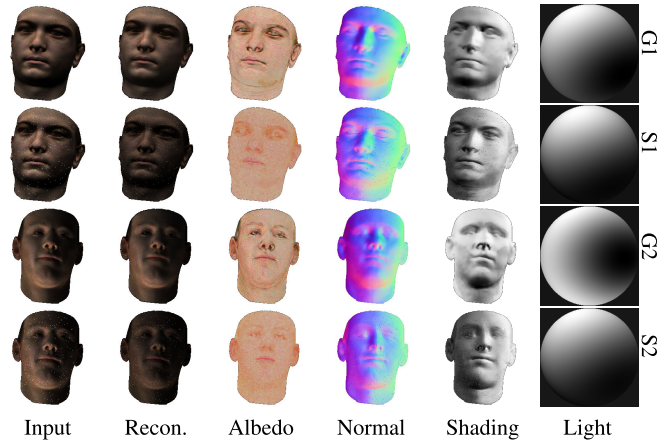


Fig. 17. Visualized results with gaussian noises (G1/G2) and salt-peppers noises (S1/S2).

Table IX shows the normal reconstruction and ableldo error with different noise on face images, and Figure 17 presents several visualized results of our model. Face images with

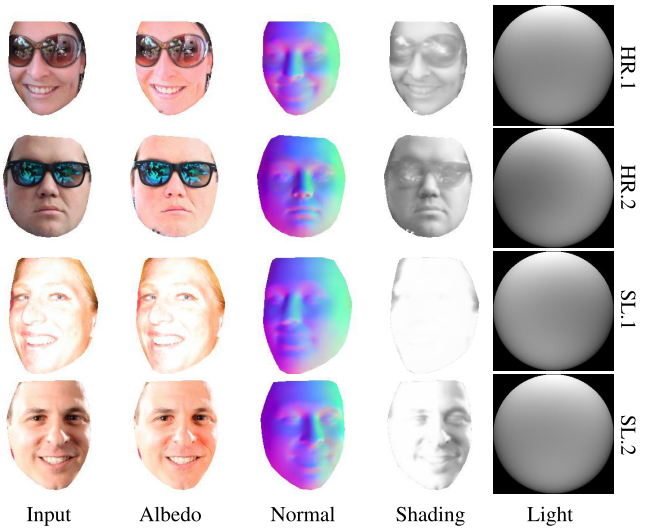


Fig. 18. Failure cases. Highly reflective glasses (HR.1/HR.2), severe lighting (SL.1/SL.2).

Gaussian noises (G1/G2 in Figure 17) will treat the noise as part of the albedo, while face images with salt-pepper noises (S1/S2 in Figure 17) will treat the noise as part of the face geometry. Regardless of how the model handles the noise, the model is able to decompose each component. It is worth mentioning that face with noise increases training time. The model is not optimal at the current number of 250k iterations, especially with the influence of salt-pepper noises.

VI. Concluding Remarks

In this paper, we have presented a novel hierarchical decoupling network to solve the problem of face inverse rendering in the wild. The relief of data preparation can significantly broaden the applicable range of face inverse rendering. Our hierarchical decoupling network is inspired by a divide-and-conquer strategy, which converts a one-to-three decomposition problem into two decoupled one-to-two sub-problems. Our model can learn the implicit relationship of the light between the paired images during training, which improves its generalization ability to unseen data. The hierarchical decoupling network has revealed its advantages over other state-of-the-art inverse rendering approaches through extensive experiments.

The proposed method still has limitations, some of which are shown in Figure 18. These belong to extreme situations, such as a face with highly reflective glasses (HR.1/HR.2 in Figure 18) or extreme lighting (SL.1/SL.2 in Figure 18).

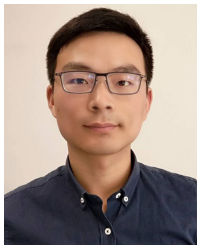
Our method can predict plausible components without ground truths. The relit faces are pretty well in different lighting directions (as shown in Figure 8, Figure 9, Figure 10 and Figure 11), while the results has some artifacts compared to original images. There are two reasons, the first one is that our model is trained with the masks. In order to compare with other methods, we utilize a Poisson blending [31] to rebuild relit faces with backgrounds, resulting in the inconsistency of the whole image. Second, predicted smooth shading losses the details, and it is hard for *NLD-Net* to produce normal details.

In addition, it is worth mentioning that the unwarping and warping functions limit the quality of final results. For example, the predicted albedo is used to relight a new face with predicted normal and new light. However, warping and unwarping processes will cause detail loss because of using bilinear sampling on image pixels, thus leading to unsatisfactory quality when the warped components are applied to relight new faces. For real face images in the wild, our network is limited by the size of warped and unwarped images, and if the size of unwarped face images is $512 * 512$ or larger, then the estimated face components will perform more plausible. Besides, our network does not train in an end-to-end way. Therefore, we expect future work to explore further robust face features that can be used in extreme situations. We would like to build an end-to-end model for learning 3DMM to align unwarped face pairs of images and simultaneously predict inverse renderings components.

References

[1] *Georgia Tech Face Database*, Georgia Inst. Technol., Atlanta, GA, USA, 2007. Accessed: Feb. 2, 2013.
[2] D. Antensteiner, S. Stolc, and D. Soukup, "Single image multi-spectral photometric stereo using a split U-shaped CNN," in *Proc. CVPRW*, Jun. 2019, pp. 1–3.
[3] A. Bansal, B. Russell, and A. Gupta, "Marr revisited: 2D–3D alignment via surface normal prediction," in *Proc. CVPR*, Jun. 2016, pp. 5965–5974.
[4] S. Barsky and M. Petrou, "The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1239–1252, Oct. 2003.
[5] R. Basri, D. W. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *Int. J. Comput. Vis.*, vol. 72, no. 3, pp. 239–257, May 2007.
[6] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. SIGGRAPH*, 1999, pp. 187–194.
[7] A. Chakrabarti and K. Sunkavalli, "Single-image RGB photometric stereo with spatially-varying albedo," in *Proc. 3DV*, Oct. 2016, pp. 258–266.
[8] C.-Y. Chen, R. Klette, and C.-F. Chen, "Recovery of coloured surface reflectances using the photometric stereo method," Citeseer, Princeton, NJ, USA, Tech. Rep. CITR-TR-117, 2002.
[9] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *Proc. SIGGRAPH*. Reading, MA, USA: Addison-Wesley, 2000, pp. 145–156.
[10] B. Egger *et al.*, "Occlusion-aware 3D morphable models and an illumination prior for face image analysis," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1269–1287, Dec. 2018.
[11] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction," in *Proc. CVPR*, Jun. 2019, pp. 1155–1164.
[12] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec, "Multiview face capture using polarized spherical gradient illumination," in *Proc. SIGGRAPH Asia*, 2011, pp. 1–10.
[13] K. Hamaen, D. Miyazaki, and S. Hiura, "Multispectral photometric stereo using intrinsic image decomposition," in *Proc. Int. Workshop Frontiers Comput. Vis.* Singapore: Springer, 2020, pp. 289–304.
[14] S. Hashimoto, D. Miyazaki, and S. Hiura, "Uncalibrated photometric stereo constrained by intrinsic reflectance image and shape from silhoutte," in *Proc. MVA*, May 2019, pp. 1–6.
[15] D. Hauagge, S. Wehrwein, K. Bala, and N. Snavely, "Photometric ambient occlusion for intrinsic image decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 639–651, Apr. 2016.
[16] A. Hou, Z. Zhang, M. Sarkis, N. Bi, Y. Tong, and X. Liu, "Towards high fidelity face relighting with realistic shadows," in *Proc. CVPR*, Jun. 2021, pp. 14719–14728.
[17] O. Ikeda and Y. Duan, "Color photometric stereo for albedo and shape reconstruction," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2008, pp. 1–6.

[18] D. W. Jacobs and R. Basri, "Lambertian reflectance and linear subspaces," U.S. Patent 6 853 745, Feb. 8, 2005.
[19] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. Tenenbaum, "Self-supervised intrinsic image decomposition," in *Proc. NeurIPS*, 2017, pp. 5936–5946.
[20] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, Jun. 2019, pp. 4401–4410.
[21] I. Kokkinos, "UberNet: Training a universal convolutional neural network for Low-, Mid-, and high-level vision using diverse datasets and limited memory," in *Proc. CVPR*, Jul. 2017, pp. 6129–6138.
[22] E. H. Land and J. J. McCann, "Lightness and Retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, pp. 1–11, 1971.
[23] A. Lattas *et al.*, "AvatarMe: Realistically renderable 3D facial reconstruction 'in-the-wild,'" in *Proc. CVPR*, Jun. 2020, pp. 760–769.
[24] J. Lehtinen *et al.*, "Noise2Noise: Learning image restoration without clean data," 2018, *arXiv:1803.04189*.
[25] L. Lettry, K. Vanhoey, and L. Van Gool, "Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences," *Comput. Graph. Forum*, vol. 37, pp. 409–419, Oct. 2018.
[26] Z. Li and N. Snavely, "CGIntrinsics: Better intrinsic image decomposition through physically-based rendering," in *Proc. ECCV*, 2018, pp. 371–387.
[27] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, Dec. 2015, pp. 3730–3738.
[28] W.-C. Ma, H. Chu, B. Zhou, R. Urtasun, and A. Torralba, "Single image intrinsic decomposition without a single intrinsic image," in *Proc. ECCV*, 2018, pp. 201–217.
[29] T. Nestmeyer, J.-F. Lalonde, I. Matthews, and A. Lehrmann, "Learning physics-guided face relighting under directional light," in *Proc. CVPR*, Jun. 2020, pp. 5124–5133.
[30] P. S. Ogun, M. R. Jackson, and R. M. Parkin, "Determination of the surface reflectance properties of timber using photometric stereo technique," in *Proc. CEEC*, Sep. 2010, pp. 1–5.
[31] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *Proc. ACM SIGGRAPH Papers*, 2003, pp. 313–318.
[32] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li, "Photorealistic facial texture inference using deep neural networks," in *Proc. CVPR*, Jul. 2017, pp. 5144–5153.
[33] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proc. ICCV*, Oct. 2017, pp. 1576–1585.
[34] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "SfSNet: Learning shape, reflectance and illuminance of faces 'in the Wild,'" in *Proc. CVPR*, Jun. 2018, pp. 6296–6305.
[35] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan, "Self-calibrating photometric stereo," in *Proc. CVPR*, Jun. 2010, pp. 1118–1125.
[36] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. CVPR*, Jul. 2017, pp. 2107–2116.
[37] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proc. CVPR*, Jul. 2017, pp. 5541–5550.
[38] W. A. P. Smith, A. Seck, H. Dee, B. Tiddeman, J. B. Tenenbaum, and B. Egger, "A morphable face albedo model," in *Proc. CVPR*, Jun. 2020, pp. 5011–5020.
[39] T. Sun *et al.*, "Single image portrait relighting," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–79, 2019.
[40] A. Tewari *et al.*, "FML: Face model learning from videos," in *Proc. CVPR*, Jun. 2019, pp. 10812–10822.
[41] A. Tewari *et al.*, "MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proc. ICCV*, Oct. 2017, pp. 1274–1283.
[42] L. Tran and X. Liu, "Nonlinear 3D face morphable model," in *Proc. CVPR*, Jun. 2018, pp. 7346–7355.
[43] L. Tran and X. Liu, "On learning 3D face morphable model from in-the-wild images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 157–171, Jan. 2019.
[44] G. Trigeorgis, P. Snape, I. Kokkinos, and S. Zafeiriou, "Face normals 'in-the-wild' using fully convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 38–47.
[45] G. Trigeorgis, P. Snape, S. Zafeiriou, and I. Kokkinos, "Normal estimation for 'in-the-wild' faces using fully convolutional networks," in *Proc. CVPR*, vol. 2, 2017, p. 5.
[46] Y. Wang *et al.*, "Face relighting from a single image under arbitrary unknown lighting conditions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1968–1984, Nov. 2009.
[47] Z. Wang, X. Yu, M. Lu, Q. Wang, C. Qian, and F. Xu, "Single image portrait relighting via explicit multiple reflectance channel modeling," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–13, Dec. 2020.
[48] T. Weyrich *et al.*, "Analysis of human faces using a measurement-based skin reflectance model," in *Proc. TOG*, 2006, pp. 1013–1024.
[49] T. Weyrich *et al.*, "A measurement-based skin reflectance model for face rendering and editing," *None TR*, vol. 71, no. 4, pp. 1–13, 2005.
[50] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Opt. Eng.*, vol. 19, no. 1, p. 191139, 1980.
[51] S. Yamaguchi *et al.*, "High-fidelity facial reflectance and geometry inference from an unconstrained image," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, 2018.
[52] S. Zafeiriou *et al.*, "The photoface database," in *Proc. CVPRW*, Jun. 2011, pp. 132–139.
[53] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. ACM MM*, Oct. 2019, pp. 1632–1640.
[54] H. Zhou, S. Hadap, K. Sunkavalli, and D. Jacobs, "Deep single-image portrait relighting," in *Proc. ICCV*, Oct. 2019, pp. 7194–7202.
[55] H. Zhou, J. Sun, Y. Yacoob, and D. W. Jacobs, "Label denoising adversarial network (LDAN) for inverse lighting of faces," in *Proc. CVPR*, Jun. 2018, pp. 6238–6247.
[56] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, Jan. 2019.



Meng Wang received the B.E. degree from the School of Information Science and Engineering, Henan University of Technology, Henan, China, in 2012, and the M.S. degree in software engineering from the College of Information, Liaoning University, Liaoning, China, in 2015. He is currently pursuing the Ph.D. degree with the College of Intelligence and Computing, Tianjin University. His research interests include computer vision, machine learning, and pattern recognition.

Xiaojie Guo (Senior Member, IEEE) is currently a tenured Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. He was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition in 2010, the IEEE ICME Best Student Paper Runner-Up Award in 2018, and the PRCV Best Student Paper Runner-Up Award in 2020.

Wenjing Dai received the B.S. and M.S. degrees from the College of Intelligence and Computing, Tianjin University, Tianjin, China, in 2016 and 2019, respectively. Her research interests include visualization, visual analytics, and machine learning.

Jiawan Zhang (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from Tianjin University, Tianjin, China, in 2001 and 2004, respectively. He is currently a Full Professor with the College of Intelligence and Computing, Tianjin University. He served for academic events, including the General Co-Chair of ChinaVis in 2015 and 2016 and PacificVis in 2019 and 2020. He also served as the Program Committee Member or a Reviewer for many conferences and journals, including CVPR, ICCV, AAAI, VIS, PacificVis, EuroVis, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, and IEEE TRANSACTIONS ON IMAGE PROCESSING.