SfSNet: Learning Shape, Reflectance and Illuminance of Faces 'in the wild'

Soumyadip Sengupta[1], Angjoo Kanazawa[2], Carlos D. Castillo[1], and David W. Jacobs[1]

[1]University of Maryland, College Park, [2]University of California, Berkeley.

Abstract

We present SfSNet, an end-to-end learning framework for producing an accurate decomposition of an unconstrained human face image into shape, reflectance and illuminance. SfSNet is designed to reflect a physical lambertian rendering model. SfSNet learns from a mixture of labeled synthetic and unlabeled real world images. This allows the network to capture low frequency variations from synthetic and high frequency details from real images through the photometric reconstruction loss. SfSNet consists of a new decomposition architecture with residual blocks that learns a complete separation of albedo and normal. This is used along with the original image to predict lighting. SfSNet produces significantly better quantitative and qualitative results than state-of-the-art methods for inverse rendering and independent normal and illumination estimation.

1. Introduction

In this work, we propose a method to decompose unconstrained real world faces into shape, reflectance and illuminance assuming lambertian reflectance. This decomposition or inverse rendering is a classical and fundamental problem in computer vision [32, 22, 21, 2]. It allows one to edit an image, for example with re-lighting and light transfer [37]. Inverse rendering also has potential applications in Augmented Reality, where it is important to understand the illumination and reflectance of a human face. A major obstacle in solving this decomposition or any of its individual components for real images is the limited availability of ground-truth training data. Even though it is possible to collect real world facial shapes, it is extremely difficult to build a dataset of reflectance and illuminance of images in the wild at a large scale. Previous works have attempted to learn surface normal from synthetic data [35, 28], which often performs imperfectly in the presence of real world variations like illumination and expression. Supervised learning can generalize poorly if real test data comes from a different distribution than the synthetic training data.

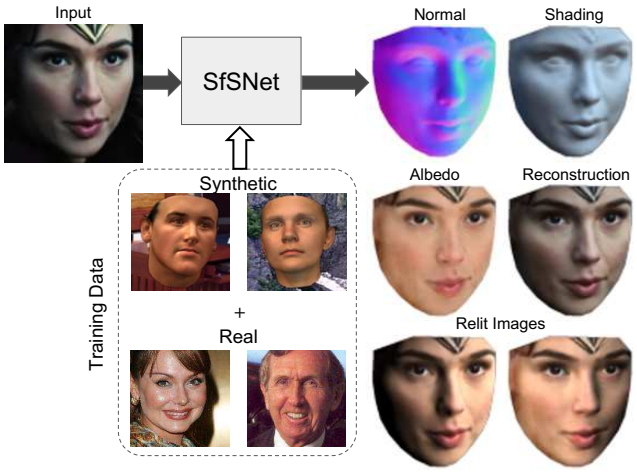

Figure 1: **Decomposing real world faces into shape, reflectance and illuminance.** We present SfSNet that learns from a combination of labeled synthetic and unlabeled real data to produce an accurate decomposition of an image into surface normals, albedo and lighting. Relit images are shown to highlight the accuracy of the decomposition. (Best viewed in color)

We propose a solution to this challenge by jointly learning all intrinsic components of the decomposition from real data. In the absence of ground-truth supervision for real data, photometric reconstruction loss can be used to validate the decomposition. This photometric consistency between the original image and inferred normal, albedo and illuminance provide strong cues for inverse rendering. However it is not possible to learn from real images only with reconstruction loss, as this may cause the individual components to collapse on each other and produce trivial solutions. Thus, a natural step forward is to get the best of both worlds by simultaneously using supervised data when available and real world data with reconstruction loss in their absence. To this end we propose a training paradigm 'SfS-supervision'.

To achieve this goal we propose a novel deep architecture called SfSNet, which attempts to mimic the physical model of lambertian image generation while learning from a mixture of labeled synthetic and unlabeled real world images.

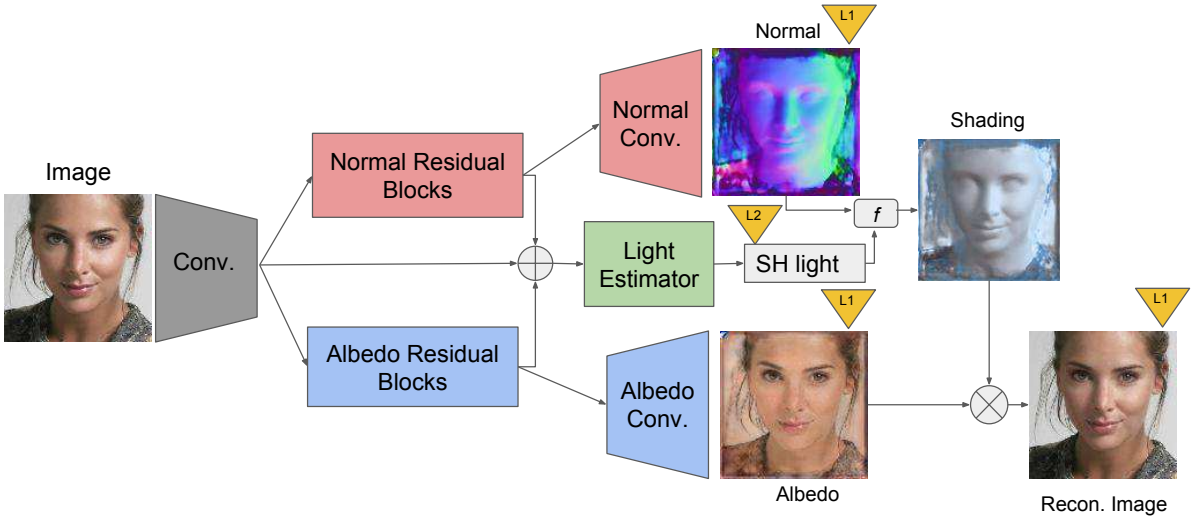

Figure 2: : **Network Architecture.** Our SfSNet consists of a novel decomposition architecture that uses residual blocks to produce normal and albedo features. They are further utilized along with image features to estimate lighting, inspired by a physical rendering model. f combines normal and lighting to produce shading. (Best viewed in color)

Training from this mixed data allows the network to learn low frequency variations in facial geometry, reflectance and lighting from synthetic data while simultaneously understanding the high frequency details in real data using shading cues through reconstruction loss. This idea is motivated by the classical works in the Shape from Shading (SfS) literature where often a reference model is used to compensate for the low frequency variations and then shading cues are utilized for obtaining high frequency details [13]. To meet this goal we develop a decomposition architecture with residual blocks that learns a complete separation of image features into normals and albedo. Then we use normal, albedo and image features to regress the illumination parameters. This is based on the observation that in classical illumination modeling, lighting is estimated from image, normal and albedo by solving an over-constrained system of equations. Our network architecture is illustrated in Figure 2. Our model and code is available for research purposes at https://senguptaumd.github.io/SfSNet/.

We evaluate our approach on the real world CelebA dataset [19] and present extensive comparison with recent state-of-the-art methods [30, 33], which also perform inverse rendering of faces. SfSNet produces significantly better reconstruction than [30, 33] on the same images that are showcased in their papers. We further compare SfSNet with state-of-the-art methods that aim to solve for only one component of the inverse rendering such as normals or lighting. SfSNet outperforms a recent approach that estimates normal independently [36], by improving normal estimation accuracy by 47% (37% to 84%) on the Photoface dataset [38], which contains faces captured under harsh lighting. We also compare against Pix2Vertex [28], which only estimate high resolution meshes. We demonstrate that SfSNet reconstructions are significantly more robust to expression and illumination variation compared to Pix2Vertex. This results from the fact that we are jointly solving for all components, which allows us to train on real images through reconstruction loss. SfSNet outperforms 'Pix2Vertex' (before meshing) by 19% (25% to 44%) without training on the Photoface dataset. We also outperform a recent approach on lighting estimation 'LDAN' [39] by 12.5% (65.9% to 78.4%).

In summary our main contributions are as follows:
(1) We propose a network, SfSNet, inspired by a physical lambertian rendering model. This uses a decomposition architecture with residual blocks to separate image features into normal and albedo, further used to estimate lighting.
(2) We present a training paradigm 'SfS-supervision', which allows learning from a mixture of labeled synthetic and unlabeled real world images. This allows us to jointly learn normal, albedo and lighting from real images via reconstruction loss, outperforming approaches that only learn an individual component.
(3) SfSNet produces remarkably better visual results compared to state-of-the-art methods for inverse rendering [30, 33]. In comparison with methods that obtain one component of the inverse rendering [36, 28, 39], SfSNet is significantly better, especially for images with expression and non-ambient illumination.

2. Related Work

Classical approaches for inverse rendering: The problem of decomposing shape, reflectance and illuminance from a single image is a classical problem in computer vi-

sion and has been studied in various forms such as intrinsic image decomposition [32] and Shape from Shading (SfS) [22, 21]. Recent work of SIRFS [2] performs decomposition of an object into surface normal, albedo and lighting assuming lambertian reflection by formulating extensive priors in an optimization framework. The problem of inverse rendering in the form of SfS gained particular attention in the domain of human facial modeling. This research was precipitated by the advent of the 3D Morphable Model (3DMM) [5] as a potential prior for shape and reflectance. Recent works used facial priors to reconstruct shape from a single image [14, 13, 6, 26] or from multiple images [25]. Classical SfS methods fail to produce realistic decomposition on unconstrained real images. More recently, Saito *et al*. proposes a method to synthesize a photorealistic albedo from a partial albedo obtained by traditional methods [27].

Learning based approaches for inverse rendering: In recent years, researchers have focused on data driven approaches for learning priors rather than hand-designing them for the purpose of inverse rendering. Attempts at learning such priors were presented in [31] using Deep Belief Nets and in [16] using a convolutional encoder-decoder based network. However these early works were limited in their performance on real world unconstrained faces. Recent work from Shu *et al*. [30] aims to find a meaningful latent space for normals, albedo and lighting to facilitate various editing of faces. Tewari *et al*. [33] solves this facial disentanglement problem by fitting a 3DMM for shape and reflectance and regressing illumination coefficients. Both [30, 33] learn directly from real world faces by using convolutional encoder-decoder based architectures. Decompositions produced by [30] are often not realistic; and [33] only captures low frequency variations. In contrast, our method learns from a mixture of labeled synthetic and unlabeled real world faces using a novel decomposition architecture. Although our work concentrates on decomposing faces, the problem of inverse rendering for generic objects in a learning based framework has also gained attention in recent years [4, 20, 29, 11].

Learning based approaches for estimating individual components: Another direction of research is to estimate shape or illumination of a face independently. Recently many research works aim to reconstruct the shape of real world faces by learning from synthetic data; by fitting a 3DMM [35, 17, 34], by predicting a depth map and subsequent non-rigid deformation to obtain a mesh [28] and by regressing a normal map [36]. Similarly [39] proposed a method to estimate lighting directly from a face. These learning based independent component estimation methods can not be trained with unlabeled real world data and thus suffer from the ability to handle unseen face modalities. In contrast our joint estimation approach performs the complete decomposition while allowing us to train on unlabeled real world images using our 'SfS-supervision'.

Architectures for learning based inverse rendering: In [30], a convolutional auto-encoder was used for disentanglement and generating normal and albedo images. However recent advances in skip-connection based convolutional encoder-decoder architectures for image to image translations [24, 10, 40] have also motivated their use in [29]. Even though skip connection based architectures are successful in transferring high frequency informations from input to output, they fail to produce meaningful disentanglement of both low and high frequencies. Our proposed decomposition architecture uses residual block based connections that allow the flow of high frequency information from input to output while each layer learns both high and low frequency features. A residual block based architecture was used for image to image translation in [12] for style transfer and in a completely different domain to learn a latent subspace with Generative Adversarial Networks [18].

3. Approach

Our goal is to use synthetic data with ground-truth supervision over normal, albedo and lighting along with real images with no ground-truth. We assume image formation under lambertian reflectance. Let $N(p)$, $A(p)$ and $I(p)$ denote the normal, albedo and image intensity at each pixel p. We represent lighting L as nine dimensional second order spherical harmonics coefficients for each of the RGB channels. The image formation process under lambertian reflectance, following [3] is represented in equation (1), where $f_{render}(.)$ is a differentiable function.

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

3.1. 'SfS-supervision' Training

Our 'SfS-supervision' consists of a multi-stage training as follows: (a) We train a simple skip-connection based encoder-decoder network on labeled synthetic data. (b) We apply this network on real data to obtain normal, albedo and lighting estimates. These elements will be used in the next stage as 'pseudo-supervision'. (c) We train our SfSNet with a mini-batch of synthetic data with ground-truth labels and real data with 'pseudo-supervision' labels. Along with supervision loss over normal, albedo and lighting we use a photometric reconstruction loss that aims to minimize the error between the original image and the reconstructed image following equation (1).

This reconstruction loss plays a key role in learning from real data using shading cues while 'pseudo-supervision' prevents the collapse of individual components of the decomposition that produce trivial solutions. In Section 6 we show that 'SfS-supervision' significantly improves inverse rendering over training on synthetic data only. Our idea of 'SfS-supervision' is motivated by the classical methods in

SfS, where a 3DMM or a reference shape is first fitted and then used as a prior to recover the details [13, 14]. Similarly in 'SfS-supervision', low frequency variations are obtained by learning from synthetic data. Then they are used as priors or 'pseudo-supervision' along with photometric reconstruction loss to add high frequency details.

Our loss function is described in equation (2). For E_N, E_A and E_{recon} we use L_1 loss over all pixels of the face for normal, albedo and reconstruction respectively; E_L is defined as the L_2 loss over 27 dimensional spherical harmonic coefficients. We train with a mixture of synthetic and real data in every mini-batch. We use λ_{recon}, λ_N and $\lambda_A = 0.5$ and $\lambda_L = 0.1$. Details of reconstruction loss under lambertian reflectance are presented in the Appendix.

$$E = \lambda_{recon} E_{recon} + \lambda_N E_N + \lambda_A E_A + \lambda_L E_L \qquad (2)$$

3.2. Proposed Architecture

A common architecture in image to image translation is skip-connection based encoder-decoder networks [24, 10]. In the context of inverse rendering, [29] used a similar skip-connection based network to perform decomposition for synthetic images consisting of ShapeNet [7] objects. We observe that in these networks most of the high frequency variations are passed from encoder to decoders through the skip connections. Thus the networks do not have to necessarily reason about whether high frequency variations like wrinkles and beards come from normal or albedo. Also in these networks the illumination is estimated only from the image features directly and is connected to normal and albedo through reconstruction loss only. However since illumination can be estimated from image, normal and albedo by solving an over-constrained system of equations, it makes more sense to predict lighting from image, normal and albedo features.

The above observations motivate us to develop an architecture that learns to separate both low and high frequency variations into normal and albedo to obtain a meaningful subspace that can be further used along with image features to predict lighting. Thus we use a residual block based architecture as shown in Figure 2. The decomposition with 'Normal Residual Blocks' and 'Albedo Residual Blocks' allows complete separation of image features into albedo and normal features as shown in Figure 3b. The skip connections (shown in red) allow the high frequency information to flow directly from input feature to output feature while the individual layers can also learn from the high frequency information present in the skip connections. This lets the network learn from both high and low frequency information and produce a meaningful separation of features at the output. In contrast a skip connection based convolutional encoder-decoder network as shown in Figure 3a consists of skip connections (shown in red) that bypass all the intermediate layers and flow directly to the output. This architecture

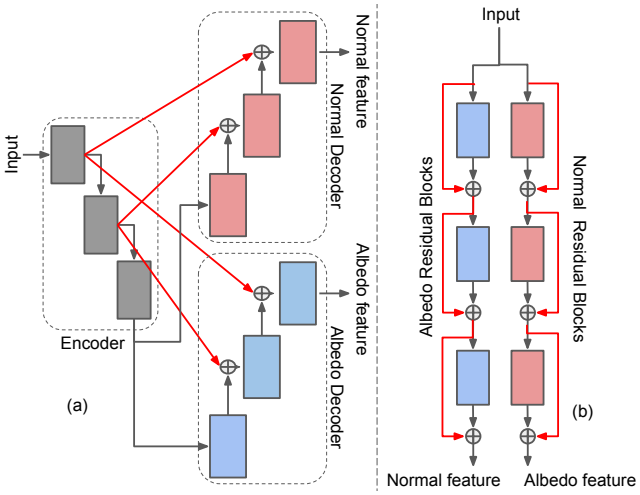

Figure 3: **Decomposition architectures.** We experiment with two architectures: (a) skip connection based encoder-decoder; (b) proposed residual block based network. Skip connections are shown in red.

allows us to estimate lighting from a combination of image, normal and albedo features. In Section 6 we show that using a residual block based decomposition improves lighting estimation by 11% (67.7% to 78.4%) compared to a skip connection based encoder-decoder.

The network uses few layers of convolution to obtain image features, denoted by I_f which is the output of the 'Conv' block in Figure 2. I_f is the input to two different residual blocks denoted as 'Normal Residual Blocks' and 'Albedo Residual Blocks', which take the image features and learns to separate them into normal and albedo features. Let the output of 'Normal Residual Blocks' and 'Albedo Residual Blocks' be N_f and A_f respectively. N_f and A_f are further processed through 'Normal Conv' and 'Albedo Conv' respectively to obtain normal and albedo aligned with the original face. To estimate lighting we use image (I_f), normal (N_f) and albedo (A_f) features in the 'Light Estimator' block of Figure 2 to obtain 27 dimensional spherical harmonic coefficients of lighting. The 'Light Estimator' block simply concatenates image, normal and albedo features followed by 1x1 convolutions, average pooling and a fully connected layer to produce lighting coefficients. The details of the network are provided in the Appendix.

3.3. Implementation Details

To generate synthetic data we use 3DMM [5] in various viewpoints, reflectance and illumination. We render these models using 27 dimensional spherical harmonics coefficients (9 for each RGB channel), which comes from a distribution estimated by fitting 3DMM over real images from the CelebA dataset using classical methods. We use CelebA

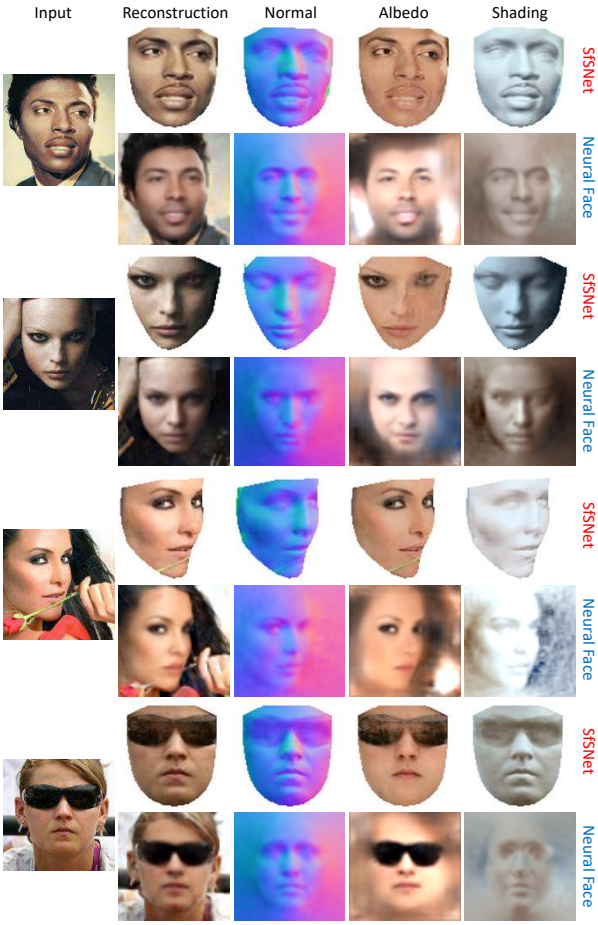

Figure 4: **Inverse Rendering. SfSNet vs 'Neural Face'** [30] on the data showcased by the authors. (Best viewed in color)

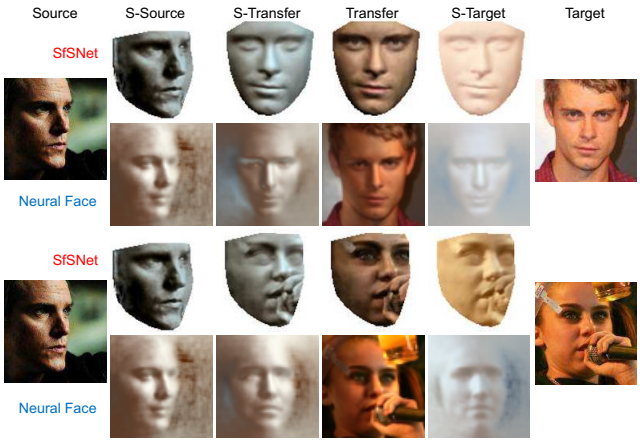

Figure 5: **Light Transfer. SfSNet vs 'Neural Face'** [30] on the image showcased by the authors. We transfer the lighting of the 'Source' image to the 'Target' image to produce 'Transfer' image. *S* denotes shading. Both 'Target' images contain an orangey glow, which is not present in the 'Source' image. Ideally in the 'Transfer' image, the orangey glow should be removed. 'Neural Face' fails to get rid of the orangey lighting effect of the 'Target' image in the 'Transfer' image. (Best viewed in color)

[19] as real data for both training, validation and testing, following the provided protocol. For real images we detect keypoints using [23] and create a mask based on these keypoints. Each of the Residual Blocks consists of 5 residual blocks based on the structure proposed by [9]. Our network is trained with input images of size 128×128 and the residual blocks all operate at 64×64 resolution. The 'pseudo-supervision' for real world images are generated by training a simple skip-connection based encoder-decoder network, similar to [30], on synthetic data. This network is also referred to as 'SkipNet' in Section 6 and details are provided in the Appendix.

4. Comparison with State-of-the-art Methods

We compare our SfSNet with [30, 33] qualitatively on unconstrained real world faces. As an application of inverse rendering we perform light transfer between a pair of images, which also illustrates the correctness of the decomposition. We quantitatively evaluate the estimated normals on the Photoface dataset [38] and compare with the state-of-the-art [36, 28]. Similarly we also evaluate the accuracy of estimated lighting on the MultiPIE dataset [8] and compare with [39]. We outperform state-of-the-art methods by a large margin both qualitatively and quantitatively.

4.1. Evaluation of Inverse Rendering

In Figures 4 and 5 we compare performance of our SfSNet with 'Neural Face' [30] on inverse rendering and light transfer respectively. The results are shown on the same images used in their paper. The results clearly show that SfSNet performs more realistic decomposition than 'Neural Face'. Note that in light transfer 'Neural Face' does not use their decomposition, but rather recomputes the albedo of the target image numerically. Light transfer results in Figure 5, show that SfSNet recovers and transfers the correct ambient light compared to 'Neural Face', which fails to get rid of the orangey lighting from the target images. We also compare inverse rendering results of SfSNet on the images provided to us by the authors of [33] in Figure 6. Since [33] aims to fit a 3DMM that can only capture low frequency variations, we obtain more realistic normals, albedo and lighting than them.

4.2. Evaluation of Facial Shape Recovery

In this section we compare the quality of our reconstructed normals with that of current state-of-the-art methods that only recover shape from a single image. We use the Photoface dataset [38], which provides ground-truth normals for images taken under harsh lighting. First we compare with algorithms that also train on the Photoface dataset.

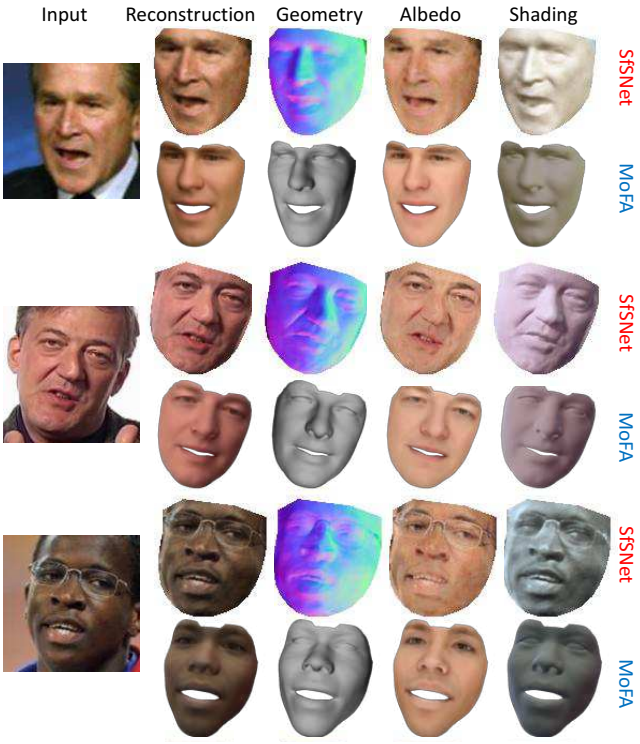

Figure 6: **Inverse Rendering. SfSNet vs 'MoFA'** [33] on the data provided by the authors of the paper. (Best viewed in color)

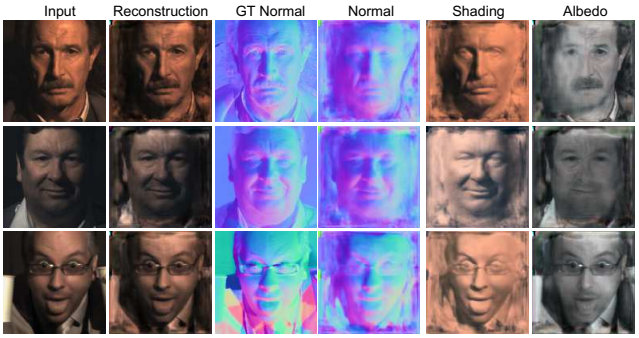

Figure 7: **Inverse Rendering on the Photoface dataset [38] with 'SfSNet-finetuned'.** The ground-truth albedo is in gray-scale and it encourages our network to also output gray-scale albedo.

We finetune our SfSNet on this dataset using ground truth normals and albedo as supervision since they are available. We compare our 'SfSNet-ft' with 'NiW' [36] and other baseline algorithms, 'Marr Rev.' [1] and 'UberNet' [15], reported in [36] in Table 1. The metric used for this task is mean angular error of the normals and the percentage of pixels at various angular error thresholds as in [36]. Since the exact training split of the dataset is not provided by the authors, we create a random split based on identity with 100 individuals in test data as mentioned in their paper. Our 'SfSNet-ft' improves normal estimation accuracy by more than a factor of two for the most challenging threshold of

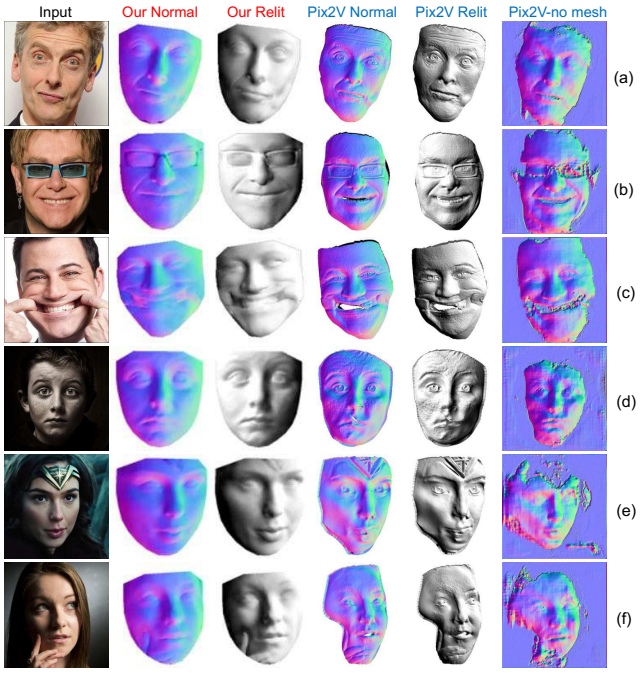

Figure 8: **SfSNet vs Pix2Vertex** [28]. Normals produced by SfSNet are significantly better than Pix2Vertex, especially for non-ambient illumination and expression. 'Relit' images are generated by directional lighting and uniform albedo to highlight the quality of the reconstructed normals. Note that (a), (b) and (c) are the images showcased by the authors. (Best viewed in color)

20 degrees accuracy. In Figure 7 we show visual results of decomposition on test data of the Photoface dataset.

Algorithm	Mean ±std	$< 20°$	$< 25°$	$< 30°$
3DMM	26.3 ±10.2	4.3%	56.1%	**89.4%**
Pix2Vertex[28]	33.9 ±5.6	24.8%	36.1%	47.6%
SfSNet	**25.5 ±9.3**	**43.6%**	**57.5%**	68.7%
Marr Rev.[1]	28.3 ±10.1	31.8%	36.5%	44.4%
UberNet[15]	29.1 ±11.5	30.8%	36.5%	55.2%
NiW[36]	22.0 ±6.3	36.6%	59.8%	79.6%
SfSNet-ft	**12.8 ±5.4**	**83.7%**	**90.8%**	**94.5%**

Table 1: **Normal reconstruction error on the Photoface dataset**. 3DMM, Pix2Vertex and SfSNet are not trained on this dataset. Marr Rev., UberNet, NiW and SfSNet-finetuned (SfSNet-ft) are trained on the training split of this dataset. Lower is better for mean error (column 1), and higher is better for the percentage of correct pixels at various thresholds (columns 3-5).

Next we compare our algorithm with 'Pix2Vertex' [28], which is trained on higher resolution 512×512 images. 'Pix2Vertex' learns to produce a depth map and a deformation map that are post-processed to produce a mesh. In contrast our goal is to perform inverse rendering. Since we are able to train on real data, unlike 'Pix2Vertex', which is trained on synthetic data, we can better capture real world variations. Figure 8 compares normals produced by SfSNet

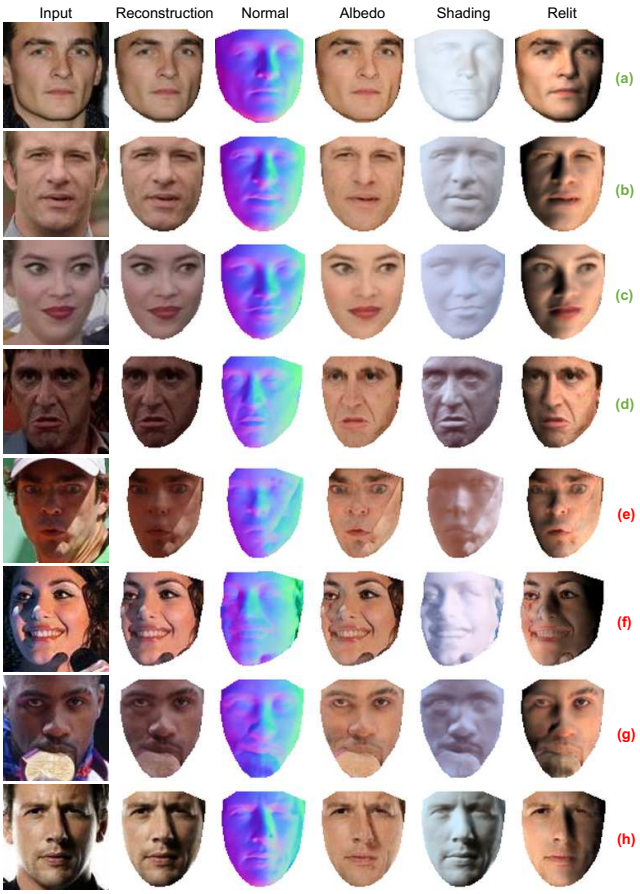

Figure 9: Selected results from **top 5% (a,b,c,d)** and **worst 5% (e,f,g,h)** reconstructed images. (Best viewed in color)

with that of 'Pix2Vertex' both before and after meshing on the images showcased by the authors. Since 'Pix2vertex' handles larger resolution and produces meshes, their normals can capture more details than ours. But with more expression and non-ambient illumination like (c), (d), (e) and (f) in Figure 8, we produce fewer artifacts and more realistic normals and shading. SfSNet is around 2000× faster than 'Pix2Vertex' due to the expensive mesh generation post-processing. These results show that learning all components of inverse rendering jointly allows us to train on real images to capture better variations than 'Pix2Vertex'. We further compare SfSNet with the normals produced by 'Pix2Vertex' quantitatively before meshing on the Photoface dataset. SfSNet, 'Pix2Vertex' and 3DMM are not trained on this dataset. The results shown in Table 1 shows that SfSNet outperforms 'Pix2Vertex' and 3DMM by a significant margin.

4.3. Evaluation of Light Estimation

We evaluate the quality of the estimated lighting using MultiPIE dataset [8] where each of the 250 individuals is photographed under 19 different lighting conditions. We perform 19-way classification, to check the consistency of the estimated lighting as described in [39] and compare with their proposed algorithm 'LDAN'. 'LDAN' estimates lighting independently from a single face image using adversarial learning. Results in Table 2 shows that we improve top-1% classification accuracy by 12.6% over 'LDAN'.

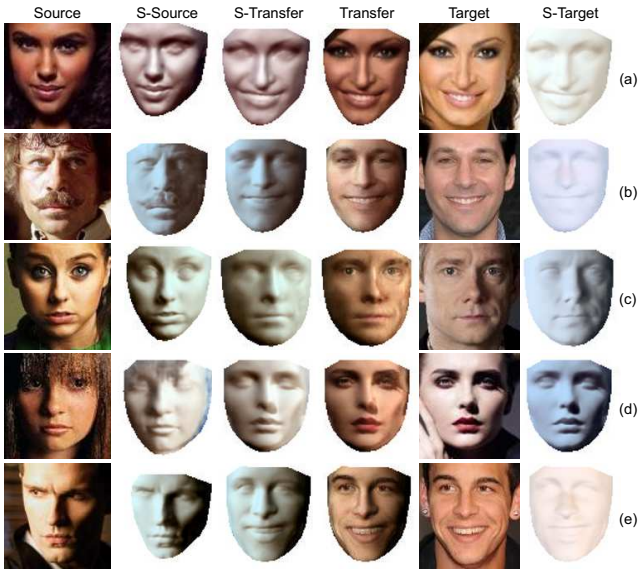

Figure 10: **Light transfer.** Our SfSNet allows us to transfer lighting of the 'Source' image to the 'Target' image to produce 'Transfer' image. 'S' refers to shading. (Best viewed in color)

Algorithm	top-1%	top-2%	top-3%
SIRFS log [2]	60.72	79.65	87.27
LDAN [39]	65.87	85.17	92.46
SfSNet	**78.44**	**89.44**	**92.64**

Table 2: **Light Classification Accuracy on MultiPIE dataset.** SfSNet significantly outperforms 'LDAN'.

5. Results on CelebA

In Figure 9 we provide sample results on CelebA test data from the best 5% and worst 5% reconstructed images respectively. For every test face, we also relight the face using a directional light source that highlights the flaws in the decomposition. As expected the best results are for frontal faces with little or no expression and easy ambient lighting as shown in Figure 9 (a-d). The worst reconstructed images have large amounts of cast shadows, specularity and occlusions as shown in Figure 9 (e-h). However, the recovered normal and lighting are still reasonable. We also show interesting results on light transfer in Figure 10, which also highlights the quality of the decomposition. Note that the examples shown in (c) and (d) are particularly hard as source and target images have opposite lighting directions. More qualitative results on CelebA and comparison with [30, 33, 28]

are provided in the Appendix.

6. Ablation Studies

We analyze the relative importance of mixed data training with 'SfS-supervision' compared to learning from synthetic data alone. We also contrast the SfSNet architecture with skip-connection based networks. For ablation studies, we consider photometric reconstruction loss (Recon. Error) and lighting classification accuracy (Lighting Acc.) as performance measures.

Role of 'SfS-supervision' training: To analyze the importance of our mixed data training we consider the SfSNet architecture and compare its performance using different training paradigms. We consider the following:
SfSNet-syn: We train SfSNet on synthetic data only.
SkipNet-syn: We observe that our residual block based network can not generalize well on unseen real world data when trained on synthetic data, as there is no direct skip connections that can transfer high frequencies from input to output. However a skip connection based encoder-decoder network can generalize on unseen real world data. Thus we consider a skip connection based network, 'SkipNet', which is similar in structure with the network presented in [30], but with increased capacity and skip connections. We train 'SkipNet' on synthetic data only and this training paradigm is similar to [29], which also uses a skip-connection based network for decomposition in ShapeNet objects.
SfSNet: We use our 'SfS-supervision' to train our SfSNet, where 'pseudo-supervision' is generated by 'SkipNet'.

Training Paradigm	Recon. Error		Lighting Acc.		
	MAE	RMSE	Rank 1	Rank 2	Rank 3
SkipNet-syn	42.83	48.22	54.86%	76.78%	85.76%
SfSNet-syn	48.54	58.13	63.88%	80.52%	87.24%
SfSNet	**10.99**	**13.55**	**78.44%**	**89.52%**	**92.64%**

Table 3: **Role of 'SfS-supervision' training.** 'SfS-supervision' outperforms training on synthetic data only.

Note that another alternative is training on synthetic data and fine-tuning on real data. It has been shown in [30] that it is not possible to train the network on real data alone by using only reconstruction loss, as the ambiguities in the decomposition can not be constrained, leading to a trivial solution. We also find that the same argument is true in our experiments. Thus we compare our 'SfS-supervision' training paradigm with only synthetic data training in Table 3. The results show that our 'SfS-supervision' improves significantly over the 'pseudo-supervision' used from SkipNet, indicating that we are successfully using shading information to add details in the reconstruction.

Role of SfSNet architecture: We evaluate the effectiveness of our proposed architecture against a skip connection based architecture. Our proposed architecture estimates lighting from image, normal and albedo, as opposed to a skip connection based network which estimates lighting directly from the image only. SkipNet described in the Appendix based on [30] does not produce a good decomposition because of the fully connected bottleneck. Thus we compare with a fully convolutional architecture with skip connection, similar to Pix2Pix [10], which we refer to as Skipnet+. This network has one encoder, two decoders for normal and albedo and a fully connected layer from the output of the encoder to predict light (see Appendix for details).

In Table 4 we show that our SfSNet outperforms SkipNet+, also trained using the 'SfS-supervision' paradigm. Although reconstruction error is similar for both networks, SfSNet predicts better lighting than 'SkipNet+'. This improved performance can be attributed to the fact that SfSNet learns an informative latent subspace for albedo and normal, which is further utilized along with image features to estimate lighting. Whereas in the case of the skip connection based network, the latent space is not informative as high frequency information is directly propagated from input to output bypassing the latent space. Thus lighting parameters estimated only from the latent space of the image encoder fail to capture the illumination variations.

Training Paradigm	Recon. Error		Lighting Acc.		
	MAE	RMSE	top-1%	top-2%	top-3%
SkipNet+	11.33	14.42	67.70%	85.08%	90.34%
SfSNet	**10.99**	**13.55**	**78.44%**	**89.52%**	**92.64%**

Table 4: **SfSNet vs SkipNet+.** Proposed SfSNet outperforms a skip connection based SkipNet+ which estimates lighting directly from the image.

7. Conclusion

In this paper we introduce a novel architecture SfSNet, which learns from a mixture of labeled synthetic and unlabeled real images to solve the problem of inverse face rendering. SfSNet is inspired by a physical rendering model and utilizes residual blocks to disentangle normal and albedo into separate subspaces. They are further combined with image features to estimate lighting. Detailed qualitative and quantitative evaluations show that SfSNet significantly outperforms state-of-the-art methods that perform inverse rendering and methods that only estimate the normal or lighting.

Acknowledgment This research is supported by the National Science Foundation under grant no. IIS-1526234. We thank Hao Zhou and Rajeev Ranjan for helpful discussions, Ayush Tewari for providing visual results of MoFA, and Zhixin Shu for providing test images of Neural Face.

References

[1] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5965–5974, 2016. 6

[2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2015. 1, 3, 7

[3] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 3

[4] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014. 3

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 3, 4

[6] M. Chai, L. Luo, K. Sunkavalli, N. Carr, S. Hadap, and K. Zhou. High-quality hair modeling from a single portrait photo. *ACM Transactions on Graphics (TOG)*, 34(6):204, 2015. 3

[7] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 4

[8] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 5, 7

[9] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016. 5

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 3, 4, 8

[11] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. Tenenbaum. Self-supervised intrinsic image decomposition. In *Advances in Neural Information Processing Systems*, pages 5938–5948, 2017. 3

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 3

[13] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):394–405, 2011. 2, 3, 4

[14] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1746–1753. IEEE, 2011. 3, 4

[15] I. Kokkinos. Ubernet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[16] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015. 3

[17] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen. Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 10. ACM, 2017. 3

[18] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017. 3

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2, 5

[20] T. Narihira, M. Maire, and S. X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2992–2992, 2015. 3

[21] G. Oxholm and K. Nishino. Shape and reflectance from natural illumination. In *European Conference on Computer Vision*, pages 528–541. Springer, 2012. 1, 3

[22] E. Prados and O. Faugeras. Shape from shading. *Handbook of mathematical models in computer vision*, pages 375–388, 2006. 1, 3

[23] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 17–24. IEEE, 2017. 5

[24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 3, 4

[25] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2016. 3

[26] S. Saito, T. Li, and H. Li. Real-time facial segmentation and performance capture from rgb input. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3

[27] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li. Photorealistic facial texture inference using deep neural networks. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017. 3

[28] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. *arXiv preprint arXiv:1703.10131*, 2017. 1, 2, 3, 5, 6, 7

[29] J. Shi, Y. Dong, H. Su, and X. Y. Stella. Learning non-lambertian object intrinsics across shapenet categories. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5844–5853. IEEE, 2017. 3, 4, 8

[30] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Conference on*, pages –. IEEE, 2017. 2, 3, 5, 7, 8

[31] Y. Tang, R. Salakhutdinov, and G. Hinton. Deep lambertian networks. In *Proceedings of the 29th International Conference on Machine Learning, 2012, Edinburgh, Scotland*, 2012. 3

[32] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. In *Advances in neural information processing systems*, pages 1367–1374, 2003. 1, 3

[33] A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3, 5, 6, 7

[34] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016. 3

[35] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. *arXiv preprint arXiv:1612.04904*, 2016. 1, 3

[36] G. Trigeorgis, P. Snape, S. Zafeiriou, and I. Kokkinos. Normal Estimation For "in-the-wild" Faces Using Fully Convolutional Networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 5, 6

[37] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1968–1984, 2009. 1

[38] S. Zafeiriou, M. Hansen, G. Atkinson, V. Argyriou, M. Petrou, M. Smith, and L. Smith. The photoface database. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 132–139. IEEE, 2011. 2, 5, 6

[39] H. Zhou, J. Sun, Y. Yacoob, and D. W. Jacobs. Label denoising adversarial network (ldan) for inverse lighting of face images. *arXiv preprint arXiv:1709.01993*, 2017. 2, 3, 5, 7

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 3