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,
[1]{*sengupta,carlos,djacobs*}*@umiacs.umd.edu* [2]*kanazawa@eecs.berkeley.edu*

Abstract

We present SfSNet, an end-to-end learning framework for producing an accurate decomposition of an unconstrained image of a human face into shape, reflectance and illuminance. Our network 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 images 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.

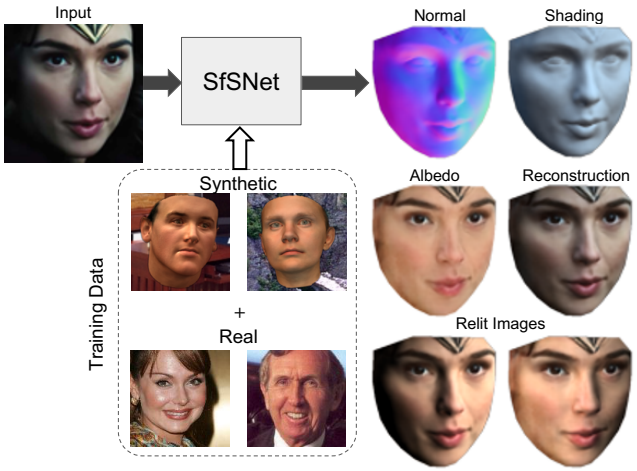

Figure 1: **Decomposing real world faces into shape, reflectance and illuminance.** We present SfSNet that learns from a combination of labeled synthetic and unlabeled real data to produce an accurate decomposition of an image into surface normals, albedo and lighting. Relit images are shown to highlight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

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 [30, 20, 19, 2]. It allows one to edit an image, for example with re-lighting and light transfer [34]. Inverse rendering also has potential applications in Augmented and Virtual 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 [32, 26], 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.

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. In the classical Shape from Shading (SfS) literature this reconstruction loss is often called as shading. 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 si-

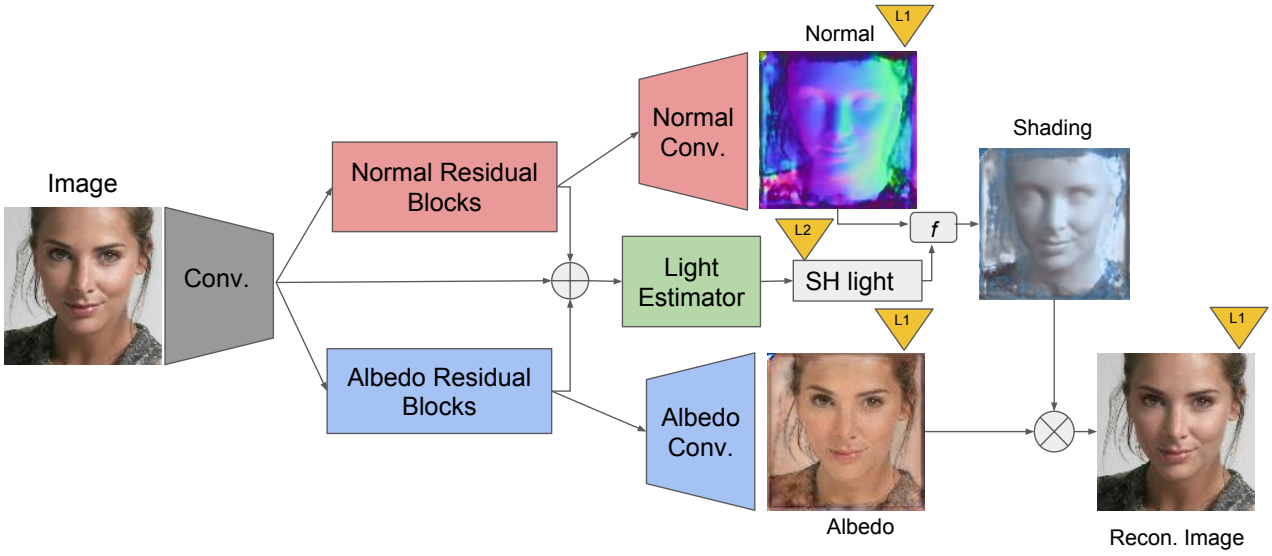

Figure 2: : **Network Architecture.** Our SfSNet consists of a novel decomposition architecture that uses residual blocks to produce normal and albedo features. They are further utilized along with image features to estimate lighting, inspired by a physical rendering model. f combines normal and lighting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

multaneously 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. 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 [11]. 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. We will make our network publicly available.

We evaluate our approach on the real world CelebA dataset [17] and present extensive comparison with recent state-of-the-art methods [28, 31], which learn from real world data using encoder-decoder networks. SfSNet produces significantly better reconstruction than [28, 31] 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 [33], by improving normal estimation accuracy by 47% (37% to 84%) on the Photoface dataset [35]. SfSNet is also compared with 'Pix2Vertex' [26] which produces higher resolution meshes by learning from synthetic data. 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 intrinsic components, which allows us to train on real images through reconstruction loss. On the Photoface dataset, which contains faces captured under harsh lighting, SfSNet outperforms 'Pix2Vertex' (before meshing) by 19% (25% to 44%) without training on this dataset. We also outperform a recent approach on lighting estimation 'LDAN' [36] by 12.5% (65.9% to 78.4%).

In summary our main contributions are as follow:
(1) We present a training paradigm 'SfS-supervision', which can learn from a mixture of labeled synthetic and unlabeled real world images. This allows us to jointly learn normal, albedo and lighting from real images using reconstruction loss, outperforming approaches that only learn an individual component.
(2) We propose a network SfSNet, inspired by a physical lambertian rendering model. This uses a novel decomposition architecture with residual blocks to separate image features into normal and albedo, further used to estimate lighting.
(3) SfSNet produces remarkably better visual results com-

pared to state-of-the-art methods for inverse rendering [28, 31]. In comparison with methods that obtain one component of the inverse rendering [33, 26, 36], 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 vision and has been studied in various forms such as intrinsic image decomposition [30] and Shape from Shading (SfS) [20, 19]. Recent work from Barron and Malik [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 [12, 11, 6, 24] or from multiple images [23]. Classical SfS methods fail to produce realistic decomposition on unconstrained real world images. More recently, Saito *et al*. proposes a method to synthesize a photorealistic albedo from a partial albedo obtained by traditional methods [25].

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 [29] using a Deep Belief Nets and in [14] 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*. [28] aims to find a meaningful latent space for normals, albedo and lighting to facilitate various editing of faces. Tewari *et al*. [31] solves this facial disentanglement problem by fitting a 3DMM for shape and reflectance and regressing illumination coefficients. Both [28, 31] learn directly from real world faces by using convolutional encoder-decoder based architectures. Decompositions produced by [28] are often not realistic; and [31] 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, 18, 27].

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 [32, 15], by predicting a depth map and subsequent non-rigid deformation to obtain a mesh [26] and by regressing a normal map [33]. Similarly [36] 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. However 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 [28], 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 [22, 10] have also motivated their use in [27]. 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 also recently used in a completely different domain to learn a latent subspace with Generative Adversarial Networks for image to image translation [16].

3. Our 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 image reconstructed following equation (1) and the original image.

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 [11, 12]. 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 image generation under lambertian reflectance and reconstruction loss are presented in Section 8.4.

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

3.2. Proposed Architecture

In image to image translation, skip connection based convolutional encoder-decoder networks as shown in [22, 10]. In the context of inverse rendering, [27] 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

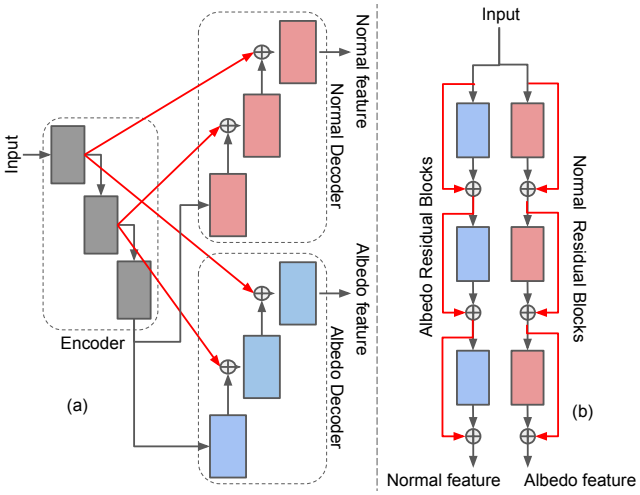

Figure 3: **Decomposition architectures.**We experiment with two architectures: (a) skip connection based encoder-decoder; (b) proposed residual block based network. Skip connections are shown in red.

'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 new decomposition architecture 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 skip connection based encoder-decoder.

The network uses a layer 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 us denote the output of 'Normal Residual Blocks' and 'Albedo Residual Blocks' as N_f and A_f respectively. N_f and A_f is 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 spher-

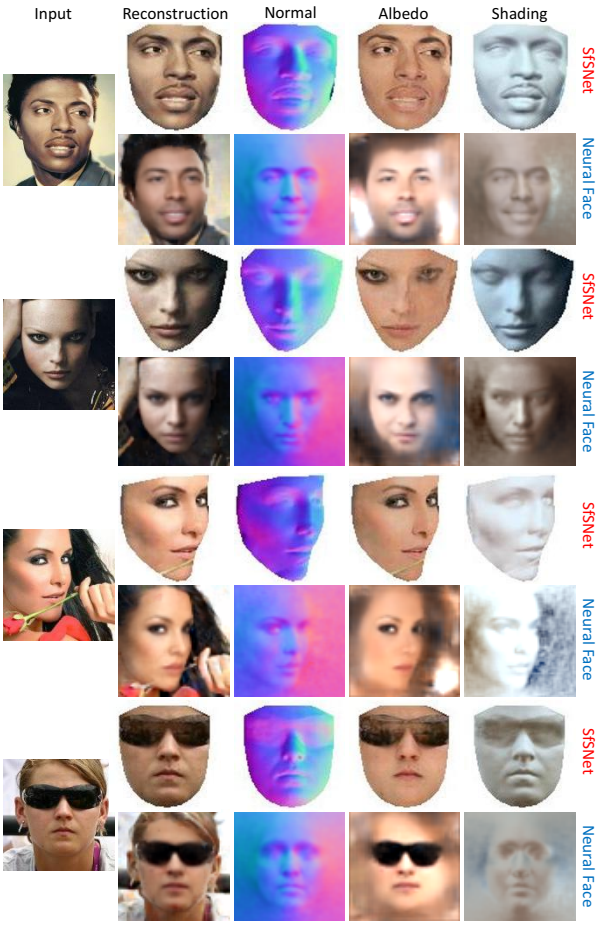

Figure 4: **Inverse Rendering. SfSNet vs 'Neural Face'** [28] on the data showcased by the authors. Note that the normals shown by SfSNet and 'Neural Face' have reversed color codes due to different choices in the coordinate system. (Best viewed in color)

ical harmonic coefficients of lighting. This 'Light Estimator' block simply concatenates image, normal and albedo features followed by 1x1 convolutions, averaging pooling and a fully connected layer to produce lighting coefficients. The details of the network architecture are provided in the Section 8.1.

3.3. Implementation Details

To generate synthetic data we use 3DMM [5] in various viewpoints and reflectance. 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 [17] as real data for both training, validation and testing, following the protocol provided by CelebA. For real images we detect keypoints using [21] and create a mask based on these keypoints. Since keypoint detections over 15 degree variation from frontal is not completely reliable, we do not use these

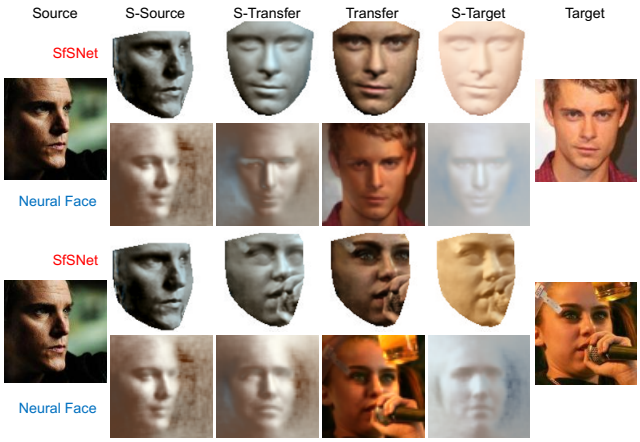

Figure 5: **Light Transfer. SfSNet vs 'Neural Face'** [28] on the data showcased by the authors. We transfer the lighting of the 'Source' image to the 'Target' image to produce 'Transfer' image. *S* denotes shading. Both 'Target' images contains an orangey glow, which is not present in the 'Source' image. 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)

larger pose variation images for training. We produce synthetic images with 45 degree pose variation from frontal, along with larger masks to generalize well on large pose variations in the test data.

For SfSNet we use 5 residual blocks in each of the '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. We will make our network publicly available for research purposes. The 'pseudo-supervision' for real world images are generated by training a simple skip-connection based encoder-decoder network, similar to [28] on synthetic data. This network is also referred to as 'SkipNet' in Section 6 and details are provided in Section 8.2.

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

Since there is no ground-truth normal, albedo and illumination for real world unconstrained images we evaluate our approach qualitatively (similar to [28, 31, 27]) by showing inferred normals, albedo, shading and reconstructed images. As an application of this inverse rendering we perform light transfer between a pair of images which also demonstrates the correctness of the decomposition. We numerically evaluate the quality of our estimated normals on the Photoface dataset [35] and compare with the state-of-the-art [33, 26]. Similarly we also evaluate the accuracy of estimated lighting on the MultiPIE dataset [8] and compare with [36]. We outperform state-of-the-art methods by a large margin both qualitatively and quantitatively.

Figure 6: **Inverse Rendering. SfSNet vs 'MoFA'** [31] on the data provided by the authors of the paper. (Best viewed in color)

4.1. Comparison with 'Neural Face' [28]

In Figures 4 and 5 we compare performance of our method SfSNet with 'Neural Face' [28] on inverse rendering and light transfer respectively. The results are shown on the images showcased by the authors in their paper[1]. 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 ambient lighting from the target images.

4.2. Comparison with 'MoFA' [31]

We compare inverse rendering results of SfSNet on the images provided to us by the authors of [31] in Figure 6. Since [31] aims to fit a linear 3DMM that can only capture low frequency variations, we certainly obtain more realistic normals, albedo and lighting than them.

4.3. 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 [35], which provides ground-truth normals for images taken under harsh lighting. First we compare with algorithms that also train on the Photoface dataset. We finetune our SfSNet on this dataset using ground truth normals and albedo as supervision since they are available. In equation (2), we use λ_N, λ_A and λ_{recon} as 0.5 and λ_L as 0. We compare our 'SfSNet-ft' with 'NiW' [33] and other baseline algorithms, 'Marr Rev.' [1] and 'UberNet' [13], reported by Trigeorgis *et al*. in their paper [33] 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 [33]. 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 without training on Photoface already performs comparable to 'NiW'. When we finetune on Photoface , 'SfSNet-ft' improves normal estimation accuracy by more than a factor of two for the most challenging threshold of 20 degrees accuracy. In Figure 7 we show visual results of decomposition on test data of the Photoface dataset.

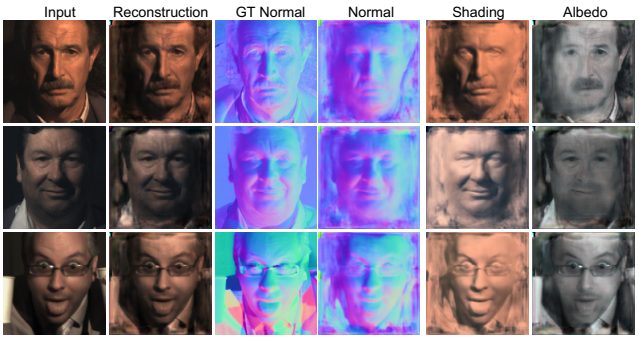

Figure 7: **Inverse Rendering on the Photoface dataset [35] with 'SfSNet-finetuned'.** Note that the ground-truth albedo is in gray-scale and encourages our network to also output gray-scale albedo. (Best viewed in color)

Algorithm	Mean ±std	$< 20°$	$< 25°$	$< 30°$
3DMM	26.3 ±10.2	4.3%	56.1%	**89.4%**
Pix2Vertex[26]	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[13]	29.1 ±11.5	30.8%	36.5%	55.2%
NiW[33]	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 pixels at various thresholds (columns 3-5).

[1] We do not yet have the permission to use their images. We show them here for convenience of the reviewers.

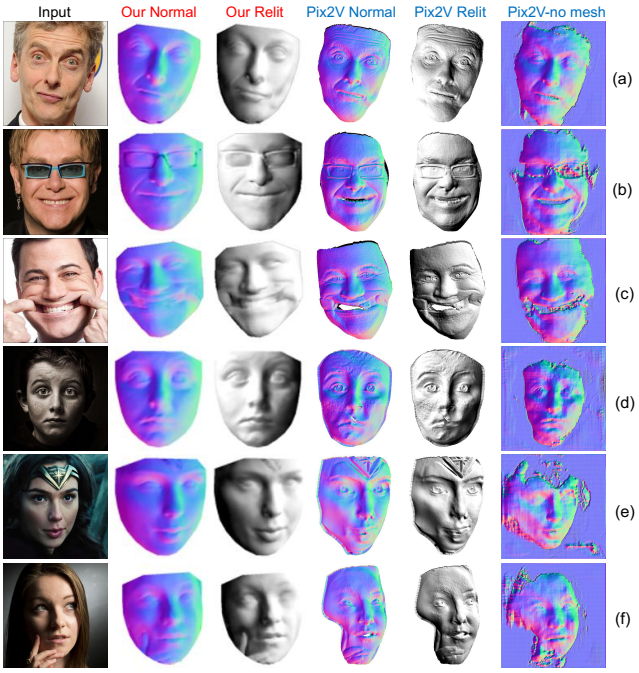

Figure 8: **SfSNet vs Pix2Vertex** [26]. Normals produced by SfSNet are significantly better than Pix2Vertex, especially for non-ambient illumination and expression. 'Relit' images are generated by directional lighting and uniform albedo to highlight the quality of the reconstructed normals. Note that (a), (b) and (c) are the images showcased by the authors. (Best viewed in color)

Next we compare our algorithm with 'Pix2Vertex' [26], 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 and we only produce normals for images of size 128×128. However 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 with that of 'Pix2Vertex' both before and after meshing on the images showcased by the authors[1] and us. Since 'Pix2vertex' handles larger resolution images 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 compared to them. SfSNet is around 2000× faster than 'Pix2Vertex' due to the expensive mesh generation post-processing. This shows that learning all components of inverse rendering jointly allows us to train on real images to capture better variations than 'Pix2Vertex'. We believe in future it is possible to add the novel meshing approach proposed by 'Pix2Vertex' in our framework to perform higher resolution inverse rendering with meshes. We further compare SfSNet with the normals produced by 'Pix2Vertex' 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.

4.4. Evaluation of Light Estimation

Algorithm	top-1%	top-2%	top-3%
SIRFS log [2]	60.72	79.65	87.27
LDAN [36]	65.87	85.17	92.46
SfSNet-L1	50.90	67.08	73.40
SfSNet	**78.44**	**89.44**	**92.64**

Table 2: **Light Classification Accuracy on Multi-PIE dataset.** SfSNet significantly outperforms 'LDAN'. SfSNet-L1 uses the inferred normal and albedo of SfSNet and recomputes lighting by solving a L1 optimization for each of the RGB channels.

To evaluate the quality of the estimated lighting, we use the MultiPIE dataset [8] where each of the 250 individuals is photographed under 19 different lighting conditions. We perform 19 class classification, to check the consistency of the estimated lighting as described in [36] 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'.

Once normal and albedo is inferred by SfSNet, lighting can also be solved as an over-constrained optimization problem. Thus we use the inferred normal and albedo from the output of SfSNet and estimate 9 dimensional spherical harmonics coefficients for each of the RGB channels independently by solving a L_1 optimization [11] we show that our SfSNet outperforms 'SfSNet-L1' in classification of lighting. This shows that learning illumination jointly with normal and albedo is significantly better than independently computing it from the inferred normals and albedo. Detailed inspection reveals that 'SfSNet-L1' is unable to properly disambiguate the correlation of lighting across the RGB channels, whereas our SfSNet actually learns this correlation from the data while training.

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 random directional light source which highlights the flaws in the decomposition of the face. As expected the best results are for faces that are frontal with no or little 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 we still perform a decent job in obtaining

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

normals and lighting. We also show some interesting results on light transfer from a source image to a target image 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 [28, 31, 26] is provided in Section 8.5.

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 novel mixed data training we consider SfSNet architecture and compare its performance on different training paradigms. We consider the following:

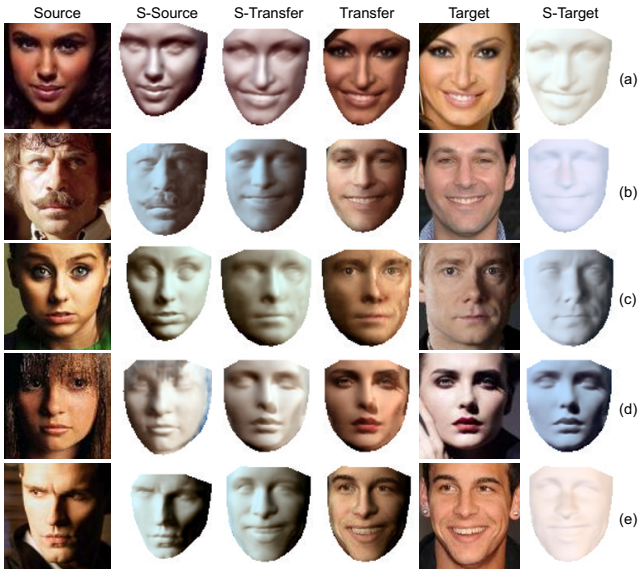

Figure 10: **Light transfer.** Our SfSNet allows us to transfer lighting of the 'Source' image to the 'Target' image to produce 'Transfer' image. 'S' refers to shading. (Best viewed in color)

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 it is well known that skip connection based encoder-decoder networks 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 [28], but with increased capacity and skip connections. We train 'SkipNet' on synthetic data only and this training paradigm is similar to [27], which also uses a skip-connection based network for decomposition in ShapeNet objects. Details of 'SkipNet' are provided in Section 8.2.
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 [28] 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 Section 8.2 based on [28] does not produce a good decomposition because of the fully connected bottleneck. Thus we compare with a fully convolutional skip connection based architecture, similar to Pix2Pix [10] and we refer to this 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 Section 8.3 for details). This network is less novel than SfSNet as it is produced by combining different existing knowledge in the community. However the architecture of SkipNet+ and its application in the context of inverse rendering is also formulated by us.

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 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 fails to capture the illumination variations.

Training Paradigm	Recon. Error		Lighting Acc.		
	MAE	RMSE	Rank 1	Rank 2	Rank 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+** (skip connection network). Proposed SfSNet outperforms a skip connection based network SkipNet+ which estimates lighting only from the image.

7. Conclusion

In this paper we introduce a novel architecture SfSNet, that learns from a mixture of labeled synthetic and 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 produce lighting. Detailed qualitative and quantitative evaluations show that SfSNet significantly outperforms state-of-the-art methods that perform inverse rendering and independent normal or lighting estimation.

Our results in Figure 8 shows that SfSNet is more robust than 'Pix2Vertex' as it is trained on real data using shading cues. A natural extension of our work will be to handle higher resolution images and produce meshes following the ideas presented by [26]. We also plan to extend beyond the lambertian world and handle specularity and cast shadows, which often appear in unconstrained faces.

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.

[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.

[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.

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

[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.

[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.

[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.

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

[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.

[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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. *arXiv preprint arXiv:1703.10131*, 2017.

[27] 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.

[28] 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.

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

[30] 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.

[31] A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

8. Appendix

8.1. SfSNet Architecture

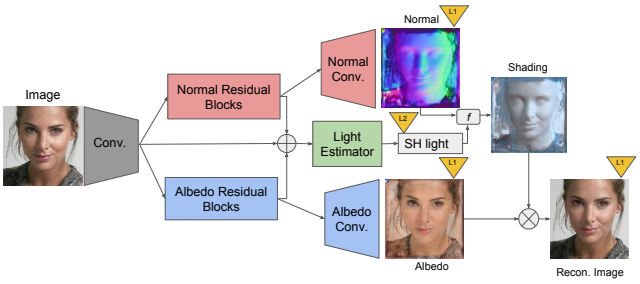

Figure 11: **SfSNet Architecture**.

The schematic diagram of our SfSNet is again shown in Figure 11 for reference. Our input, normal and albedo is of size 128×128. Below we provide the details of each of the blocks of SfSNet.

'Conv.': C64(k7) - C128(k3) - C*128(k3)
'CN(kS)' represents convolution layers with N $S \times S$ filters with stride 1, followed by Batch Normalization and ReLU. 'C*N(kS)' represents only convolution layers with N $S \times S$ filters with stride 2. The output of 'Conv' layer produces a blob of spatial resolution $128 \times 64 \times 64$.

'Normal Residual Blocks': 5 ResBLK - BN - ReLU
This consists of 5 Residual Blocks, 'BesBLK's, all of which operate at a spatial resolution of $128 \times 64 \times 64$, followed by Batch Normalization (BN) and ReLU. Each 'ResBLK' has an input X and output Y, where Y is obtained by the following structure : X - BN - ReLU - C128 - BN - ReLU - C128 - Z, $Y = Z + X$.

'Albedo Residual Blocks': Same as 'Normal Residual Blocks' (weights are not shared).

'Normal Conv'.: BU - CD128(k1) - C64(k3) - C*3(k1)
'BU' refers to Bilinear up-sampling to convert $128 \times 64 \times 64$ to $128 \times 128 \times 128$.'CN(kS)' represents convolution layers with N $S \times S$ filters with stride 1, followed by Batch Normalization and ReLU. 'C*N(kS)' represents only convolution layer with N $S \times S$ filters with stride 1. The network produces a normal map as output.

'Albedo Conv.': Same as 'Normal Conv.' (weights are not shared).

'Light Estimator': It first concatenates the responses of 'Conv', 'Normal Residual Blocks' and 'Albedo Residual Blocks' to produce a blob of spatial resolution $384 \times 64 \times 64$. This is further processed by 128 1×1 convolutions, Batch Normalization, ReLU, followed by Average Pooling over 64×64 spatial resolution to produce 128 dimensional features. This 128 dimensional feature is passed through a fully connected layer to produce 27 dimensional spherical harmonics coefficients of lighting.

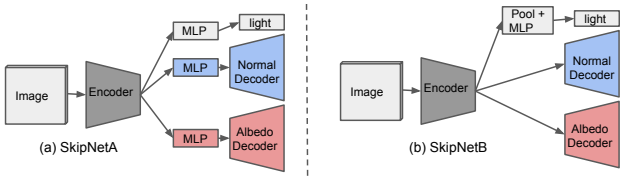

Figure 12: **SkipNet and SkipNet+ Network Architectures**.

8.2. SkipNet Architecture

The schematic diagram of SkipNet is shown in Figure 12(a). SkipNet is based on the network used in [28] with more capacity and skip connections. Similar to SfSNet the input is 128×128; 'Normal Decoder' and 'Albedo Decoder' produces normal and albedo maps. Normal, albedo and 'light' is also used to produce shading and the reconstructed image similar to Figure 11. Since that part of the architecture does not contain any trainable parameters we omit them in the figure for clarity. Note that the skip connections between encoder and decoder exist, which is also not shown in the figure. Details of SkipNet are provided below:

Encoder: C*64(k4) - C128(k4) - C256(k4) - C256(k4) - C256(k4) - fc256
'CN(kS)' represents convolution layers with N $S \times S$ filters with stride 2, followed by Batch Normalization and ReLU. 'C*N(kS)' is'CN(ks)' without Batch Normalization. All ReLUs are leaky with slope 0.2. 'fc256' is a fully connected layer that produces a 256 dimensional feature.
MLP: Contains a fully connected layer to take the response of Encoder and separate it into 256 dimensional features for 'Normal Decoder', 'Albedo Decoder' and 'light'. For 'Normal Decoder' and 'Albedo Decoder' a 256 dimensional feature is further upsampled to form a blob of shape $256 \times 4 \times 4$. For 'light' the 256 dimensional feature is passed through a fully connected network to produce 27 dimensional spherical harmonics coefficients.
Decoder (Normal and Albedo): CD256(k4) - CD256(k4) - CD256(k4) - CD128(k4) - CD64(k4) - C*3(k1) Both 'Normal Decoder' and 'Albedo Decoder' consists of the same architecture without weight sharing. 'CDN(kS)' represents a de-convolution layer with N $S \times S$ filters operated with stride 2, followed by Batch Normalization and ReLU. 'C*3(k1)' consists of 3 1×1 convolution filters with stride 1 to produce Normal or Albedo. Skip connections are present between encoders and decoders similar to [10, 26].

8.3. SkipNet+

SkipNet+ is very similar to SkipNet, but with larger capacity and without a fully connected bottleneck 'MLP' as shown in Figure 12(b). The Details of the network are shown below.
Encoder: Co64(k3) - Co64(k1) - C64(k3) - Co64(k1) - C128(k3) - Co128(k1) - C256(k3) - Co256(k1) - C256(k3) - Co256(k1) - C256(k3)
'CN(kS)' represents a convolution layer with N $S \times S$ filters with stride 2, followed by Batch Normalization and ReLU. 'CoN(kS)' is similar to 'CN(kS)' but with stride 1. All ReLUs are leaky with slope 0.3. The output of the Encoder is a feature of spatial resolu-

tion $256 \times 4 \times 4$.

Decoder (Normal and Albedo): C256(k1) - CD256(k4) - CD256(k4) - CD256(k4) - CD128(k4) - CD64(k4) - C*3(k1)
'CDN(kS)' represents a de-convolution layer with N $S \times S$ filters with stride 2, followed by Batch Normalization and ReLU. 'CN(kS)' represents a convolution layer with N $S \times S$ filters with stride 1, followed by Batch Normalization and ReLU. 'C*3(k1)' consists of 3 1×1 convolution filters to produce Normal or Albedo. Skip-connections exists between 'CN(k3)' layers of encoder and 'CDN(k4)' layers of decoder.

light: We perform Average pooling over 4×4 spatial resolution of the encoder output to produce a 256 dimensional feature. This feature is then passed through a fully connected layer to produce 27 dimensional spherical harmonics lighting.

8.4. Spherical Harmonics

In this section, we define the image generation process under lambertian reflectance following equation (1). Let the normal be $n(p) = [x, y, z]^T$ at pixel p. Then 9 dimensional spherical harmonics basis $Y(p)$ at pixel p is expressed as:

$$Y = [Y_{00}, Y_{10}, Y_{11}^e Y_{11}^0, Y_{20}, Y_{21}^e, Y_{21}^o, Y_{22}^e, Y_{22}^o]^T, \quad (3)$$

where

$$\begin{aligned}
Y_{00} &= \tfrac{1}{\sqrt{4\pi}} & Y_{10} &= \sqrt{\frac{3}{4\pi}}z \\
Y_{11}^e &= \sqrt{\tfrac{3}{4\pi}}x & Y_{11}^o &= \sqrt{\frac{3}{4\pi}}y \\
Y_{20} &= \tfrac{1}{2}\sqrt{\tfrac{5}{4\pi}}(3z^2-1) & Y_{21}^e &= 3\sqrt{\frac{5}{12\pi}}xz \\
Y_{21}^o &= 3\sqrt{\tfrac{5}{12\pi}}yz & Y_{22}^e &= \frac{3}{2}\sqrt{\frac{5}{12\pi}}(x^2-y^2) \\
Y_{22}^0 &= 3\sqrt{\tfrac{5}{12\pi}}xy
\end{aligned} \quad (4)$$

Then the intensity at pixel p is defined as :

$$I(p) = f_{render}(A(p), N(p), L) = A(p)(Y(p)^T L), \quad (5)$$

where $A(p)$ is the albedo at pixel p, and L is the lighting parameter denoting coefficients of spherical harmonics basis. Note that, the above equations are only for one of the RGB channels and can be repeated independently for 3 channels.

Next we define the reconstruction loss. Let $I(p)$ be the original image intensity and $\tilde{N}(p)$, $\tilde{A}(p)$ be the inferred normal and albedo by SfSNet at pixel p. Let $\tilde{L}$ be the 27 dimensional spherical harmonic coefficients also inferred by SfSNet. The reconstruction loss is defined as:

$$E_{recon} = \sum_p |I(p) - f_{render}(\tilde{A}(p), \tilde{N}(p), \tilde{L})|. \quad (6)$$

8.5. More Qualitative Comparisons

SfSNet on CelebA: In Figures 13 and 14 we present Inverse Rendering results on CelebA images with our SfSNet. To visualize the quality of the reconstructed normals, we use directional lights with uniform albedo to produce 'Relit' images.

SfSNet vs Pix2Vertex: In Figure 15 we compare SfSNet to Pix2Vertex [26] on images selected by us. These images contain non-ambient illuminations and expressions, where our SfSNet reconstructs much better normals than Pix2Vertex. Figures 16, 17 and 18 also compares performance of SfSNet and Pix2Vertex on the images showcased by Sela *et al*. in [26]. Since these images mostly contain ambient illumination, SfSNet performs comparable to Pix2Vertex.

SfSNet vs MoFA: We also provide more comparison results with MoFA [31] on the images provided by the authors in Figures 20, 21 and 22. MoFA aims to fit a 3DMM which is limited in its capability to represent real world shapes and reflectance, but can produce a full 3D mesh. Thus SfSNet reconstructs better shape and reflectance than MoFA.

SfSNet vs Neural Face: Similarly comparison with 'Neural Face' [28] in Figure 23 on the images showcased by the authors, show that SfSNet obtains more realistic reconstruction than 'Neural Face'.

Figure 13: Results of SfSNet on CelebA. 'Relit' images are generated by directional lighting and uniform albedo to highlight the quality of the reconstructed normals. (Best viewed in color)

Figure 14: Results of SfSNet on CelebA. 'Relit' images are generated by directional lighting and uniform albedo to highlight the quality of the reconstructed normals. (Best viewed in color)

Figure 15: **SfSNet vs Pix2Vertex** [26] on images selected by us with non-ambient illumination and expression. 'Relit' images are generated by directional lighting and uniform albedo selected to highlight the quality of the reconstructed normals. (Best viewed in color)

Figure 16: **SfSNet vs Pix2Vertex** [26] on the images showcased by Sela *et al*. in [26]. 'Relit' images are generated by directional lighting and uniform albedo selected to highlight the quality of the reconstructed normals. (Best viewed in color)

Figure 17: **SfSNet vs Pix2Vertex** [26] on the images showcased by Sela *et al*. in [26]. 'Relit' images are generated by directional lighting and uniform albedo selected to highlight the quality of the reconstructed normals. (Best viewed in color)

Figure 18: **SfSNet vs Pix2Vertex** [26] on the images showcased by Sela *et al*. in [26]. 'Relit' images are generated by directional lighting and uniform albedo selected to highlight the quality of the reconstructed normals. (Best viewed in color)

Figure 19: **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)

Figure 20: **Inverse Rendering. SfSNet vs 'MoFA'** [31] on the data provided by the authors. (Best viewed in color)

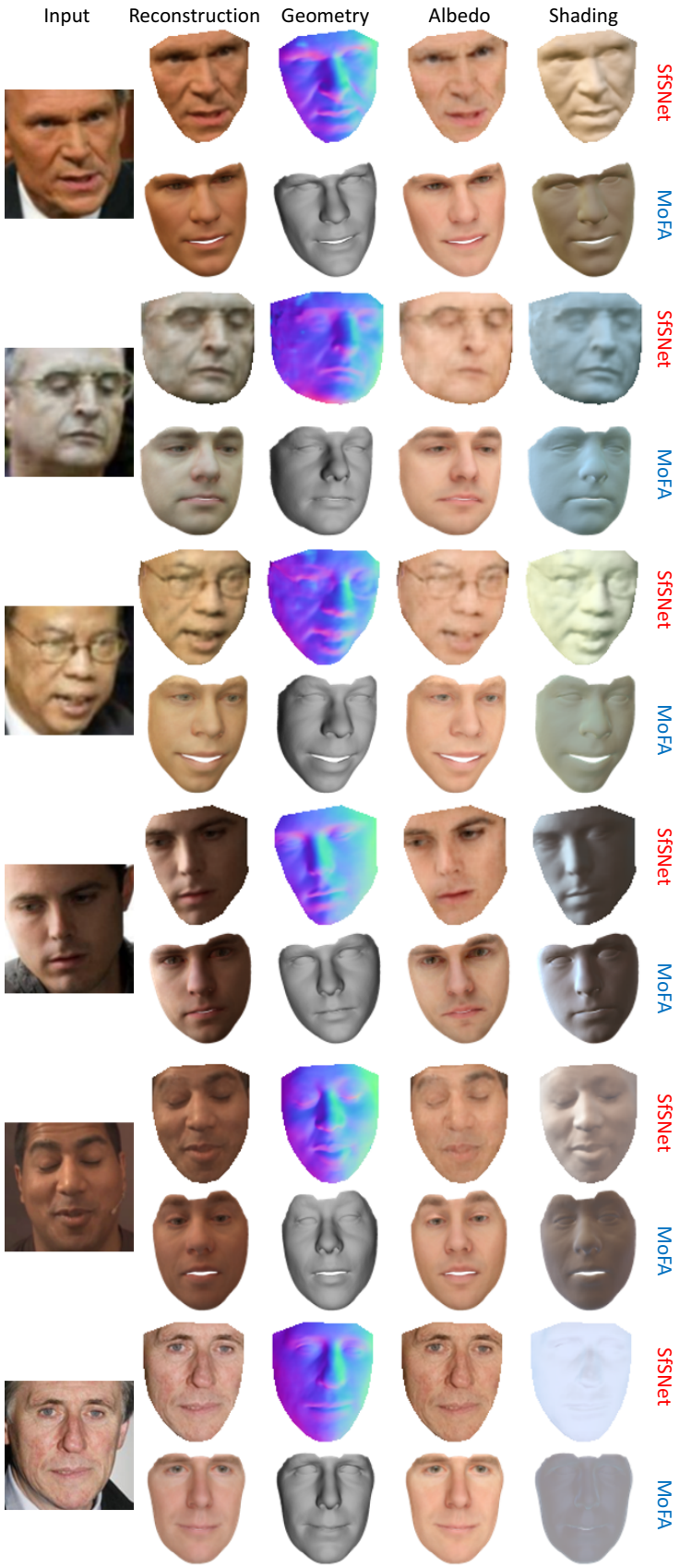

Figure 21: **Inverse Rendering. SfSNet vs 'MoFA'** [31] on the data provided by the authors. (Best viewed in color)

Figure 22: **Inverse Rendering. SfSNet vs 'MoFA'** [31] on the data provided by the authors. (Best viewed in color)

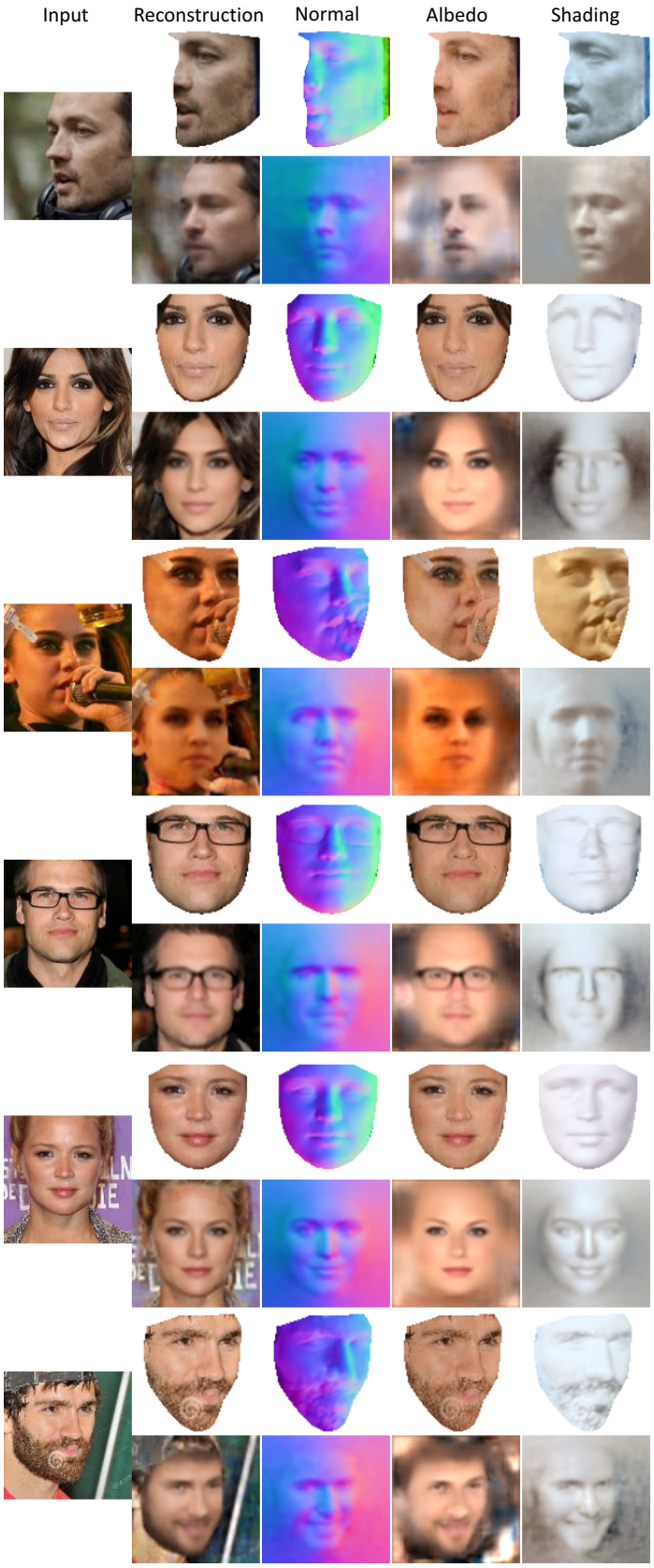

Figure 23: **Inverse Rendering. SfSNet vs 'Neural Face'** [28] on the images showcased by the authors. Note that the normals shown by SfSNet and 'Neural Face' have reversed color codes due to different choices in the coordinate system. (Best viewed in color)