SfSNet: Learning Shape, Reflectance and Illuminance of Faces ‘in the wild’ Appendix

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. SfSNet Architecture

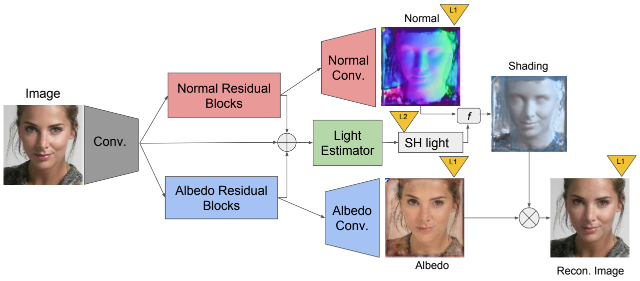

Figure 1: **SfSNet Architecture**.

The schematic diagram of our SfSNet is again shown in Figure 1 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)’ denotes convolution layers with N $S \times S$ filters with stride 1, followed by Batch Normalization and ReLU. ‘C*N(kS)’ denotes only convolution layers with N $S \times S$ filters with stride 2, without batch Normalization. The output of ‘Conv’ layer produces a blob of spatial resolution 128 × 64 × 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 × 64 × 64, followed by Batch Normalization (BN) and ReLU. Each ‘ResBLK’ consists of BN - ReLU - C128 - BN - ReLU - C128.

‘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 that converts 128 × 64 × 64 to 128 × 128 × 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.

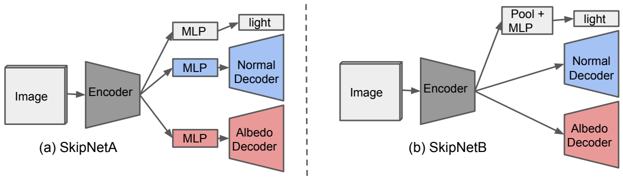

Figure 2: **SkipNet and SkipNet+ Network Architectures**.

‘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 × 64 × 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. Our model and code is available for research purposes at https://senguptaumd.github.io/SfSNet/.

2. SkipNet Architecture

The schematic diagram of SkipNet is shown in Figure 2(a). SkipNet is based on the network used in [3] 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 1. 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 up-sampled 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 [1, 2].

3. SkipNet+

SkipNet+ is very similar to SkipNet, but with larger capacity and without a fully connected bottleneck 'MLP' as shown in Figure 2(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 resolution $256 \times 4 \times 4$.
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.

4. Spherical Harmonics

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

$$Y = [Y_{00}, Y_{10}, Y^e_{11} Y^0_{11}, Y_{20}, Y^e_{21}, Y^o_{21}, Y^e_{22}, Y^o_{22}]^T, \qquad (1)$$

where

$$\begin{array}{llll} Y_{00} & = \frac{1}{\sqrt{4\pi}} & Y_{10} & = \sqrt{\frac{3}{4\pi}} z \\ Y^e_{11} & = \sqrt{\frac{3}{4\pi}} x & Y^o_{11} & = \sqrt{\frac{3}{4\pi}} y \\ Y_{20} & = \frac{1}{2}\sqrt{\frac{5}{4\pi}}(3z^2 - 1) & Y^e_{21} & = 3\sqrt{\frac{5}{12\pi}} xz \\ Y^o_{21} & = 3\sqrt{\frac{5}{12\pi}} yz & Y^e_{22} & = \frac{3}{2}\sqrt{\frac{5}{12\pi}}(x^2 - y^2) \\ Y^0_{22} & = 3\sqrt{\frac{5}{12\pi}} xy & & \end{array} \qquad (2)$$

Then the intensity at pixel p is defined as:

$$I(p) = f_{render}(A(p), N(p), L) = A(p)(Y(p)^T L), \qquad (3)$$

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:

$$E_{recon} = \sum_p |I(p) - f_{render}(\tilde{A}(p), \tilde{N}(p), \tilde{L})|. \qquad (4)$$

5. More Qualitative Comparisons

SfSNet on CelebA: In Figures 3 and 4 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 5 we compare SfSNet to Pix2Vertex [2]. These images contain non-ambient illuminations and expressions, where surface normal recovery is much more robust for SfSNet than for Pix2Vertex. Figures 6, 7 and 8 also compares performance of SfSNet and Pix2Vertex on the images showcased by Sela *et al*. in [2]. Since these images mostly contain ambient illumination, SfSNet performs comparable to Pix2Vertex.

SfSNet vs MoFA: We also provide more comparison results with MoFA [4] on the images provided by the authors in Figures 10, 11 and 12. 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 more detailed shape and reflectance than MoFA.

SfSNet vs Neural Face: Similarly comparison with 'Neural Face' [3] in Figure 13 on the images showcased by the authors, show that SfSNet obtains more realistic reconstruction than 'Neural Face'.

References

[1] 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. 2

[2] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. *arXiv preprint arXiv:1703.10131*, 2017. 2, 3, 6, 7, 8, 9

[3] 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. 1, 3, 14

[4] 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. 3, 11, 12, 13

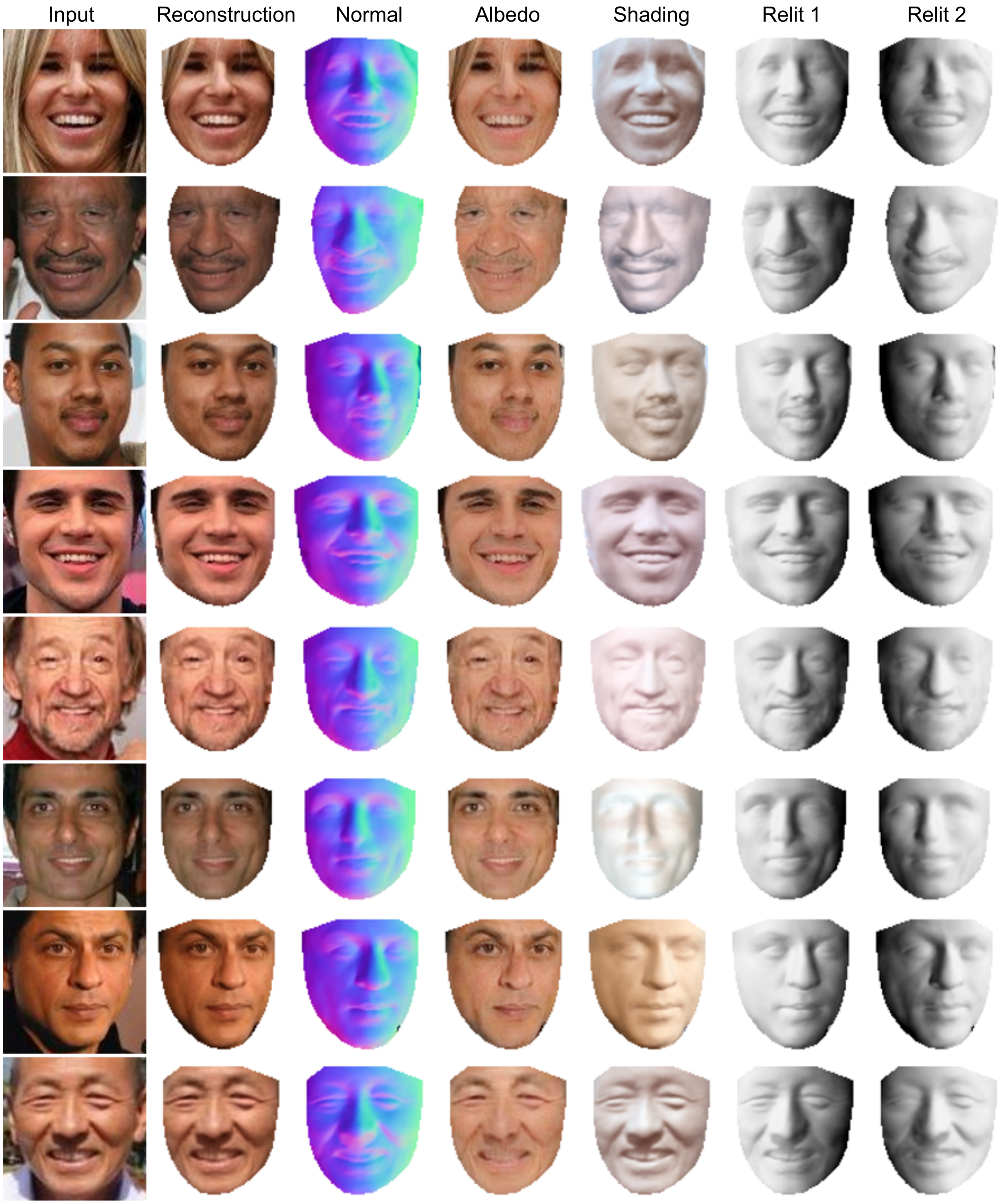

Figure 3: Results of SfSNet on CelebA. 'Relit' images are generated by directional lighting and uniform albedo to highlight the quality of the reconstructed normals. (Best viewed in color)

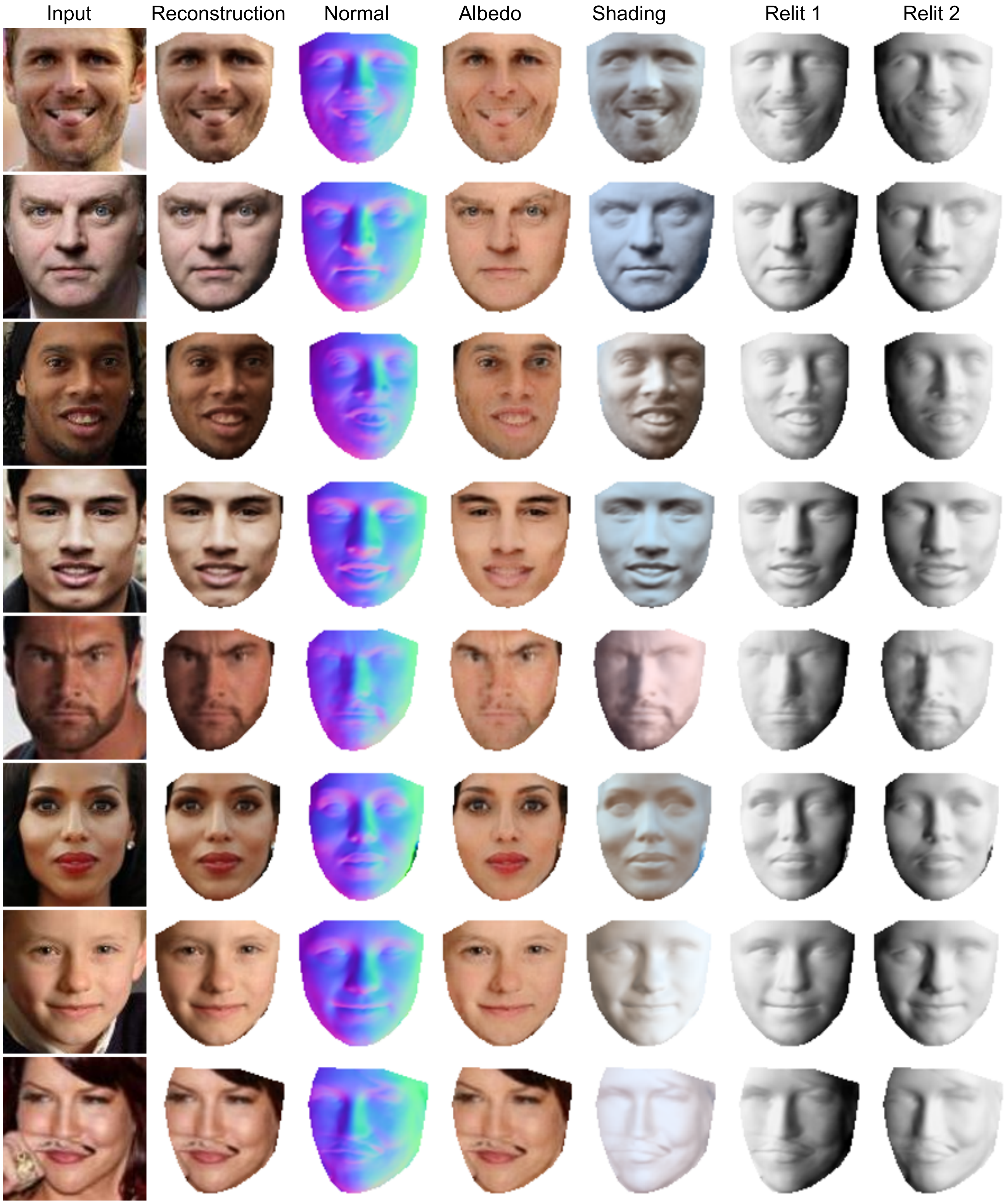

Figure 4: Results of SfSNet on CelebA. ‘Relit’ images are generated by directional lighting and uniform albedo to highlight the quality of the reconstructed normals. (Best viewed in color)

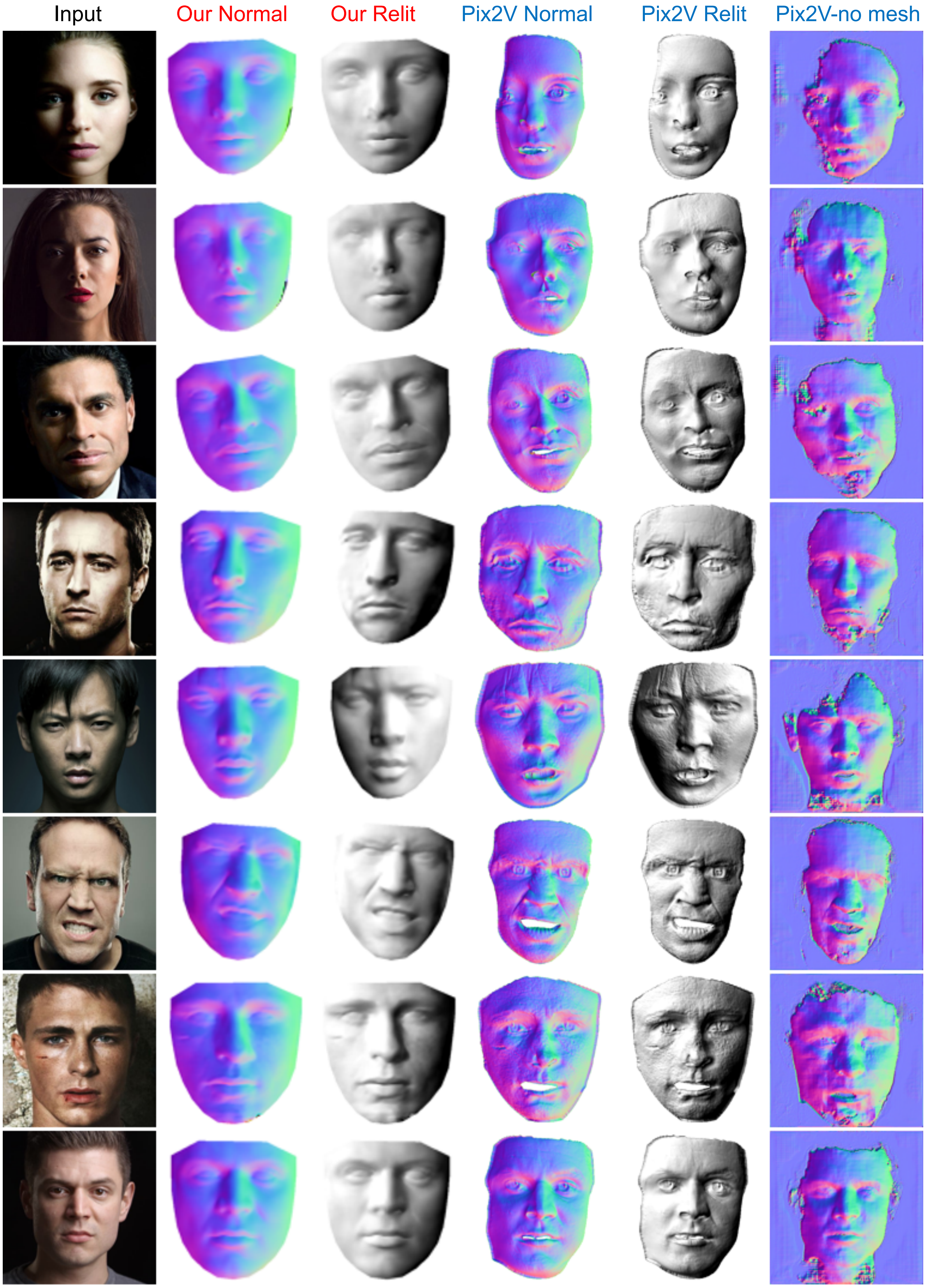

Figure 5: **SfSNet vs Pix2Vertex** [2] on images selected by us with non-ambient illumination and expression. 'Relit' images are generated by directional lighting and uniform albedo selected to highlight the quality of the reconstructed normals. (Best viewed in color)

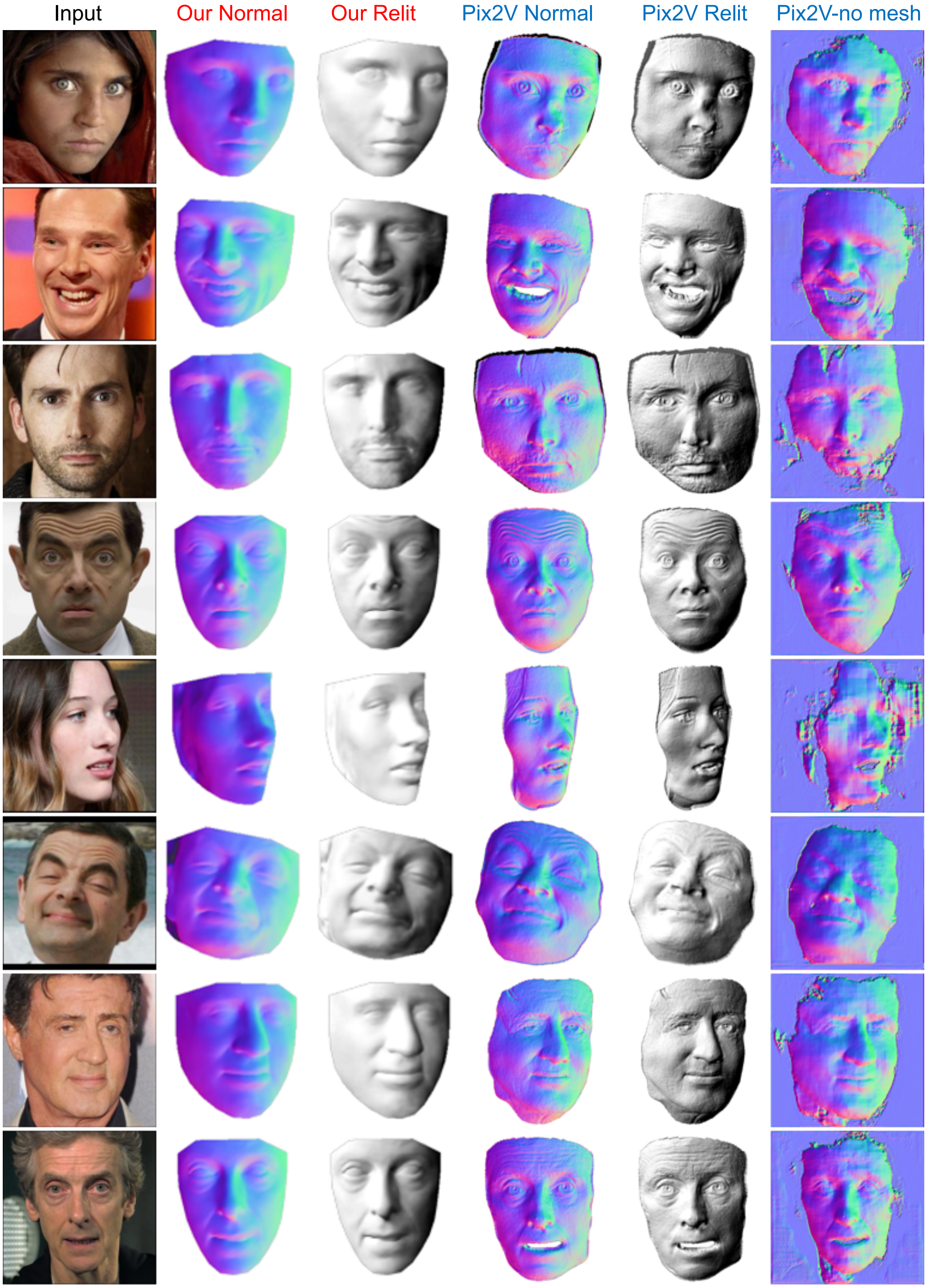

Figure 6: **SfSNet vs Pix2Vertex** [2] on the images showcased by Sela *et al*. in [2]. ‘Relit’ images are generated by directional lighting and uniform albedo selected to highlight the quality of the reconstructed normals. (Best viewed in color)

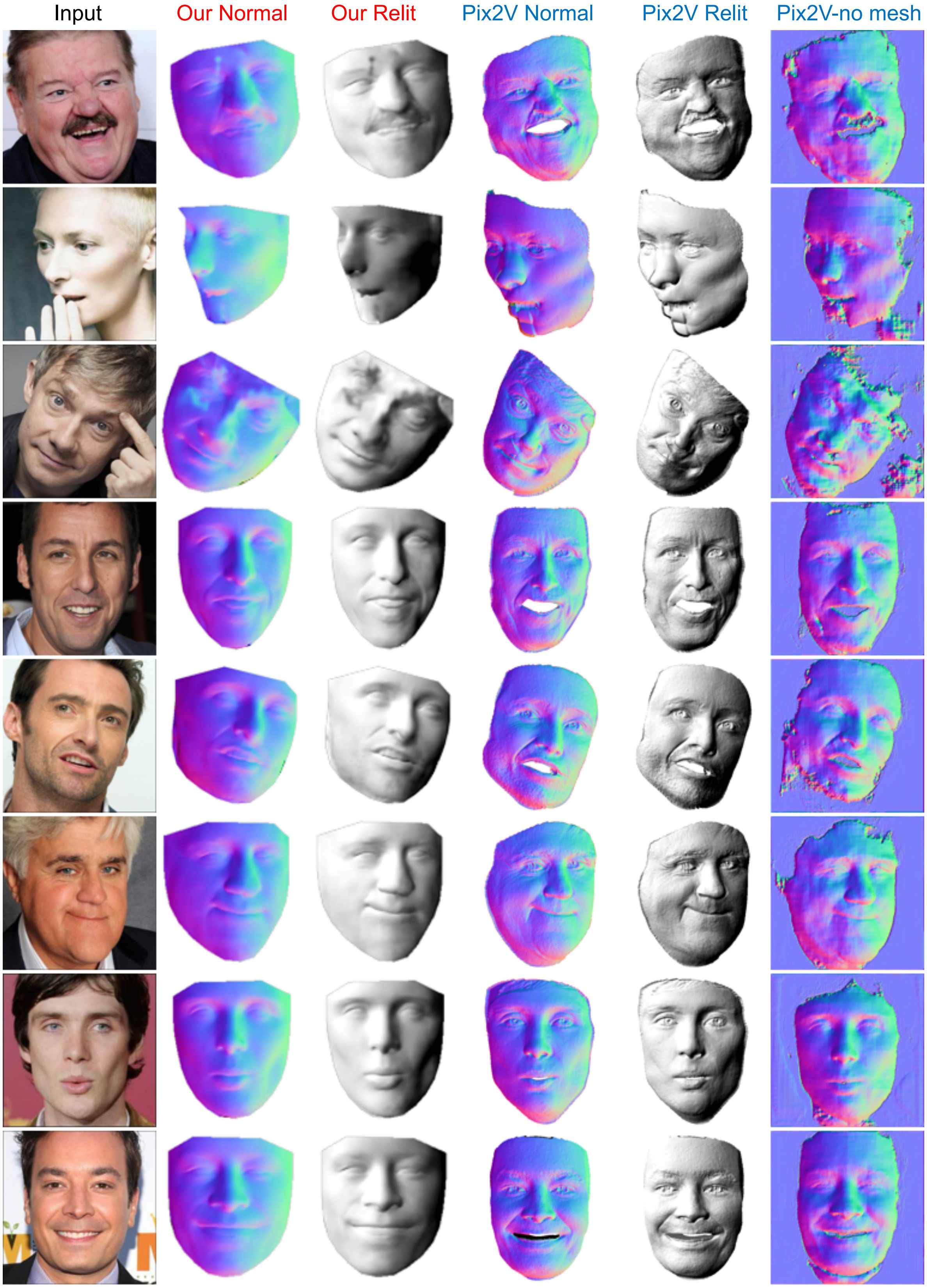

Figure 7: **SfSNet vs Pix2Vertex** [2] on the images showcased by Sela *et al*. in [2]. ‘Relit’ images are generated by directional lighting and uniform albedo selected to highlight the quality of the reconstructed normals. (Best viewed in color)

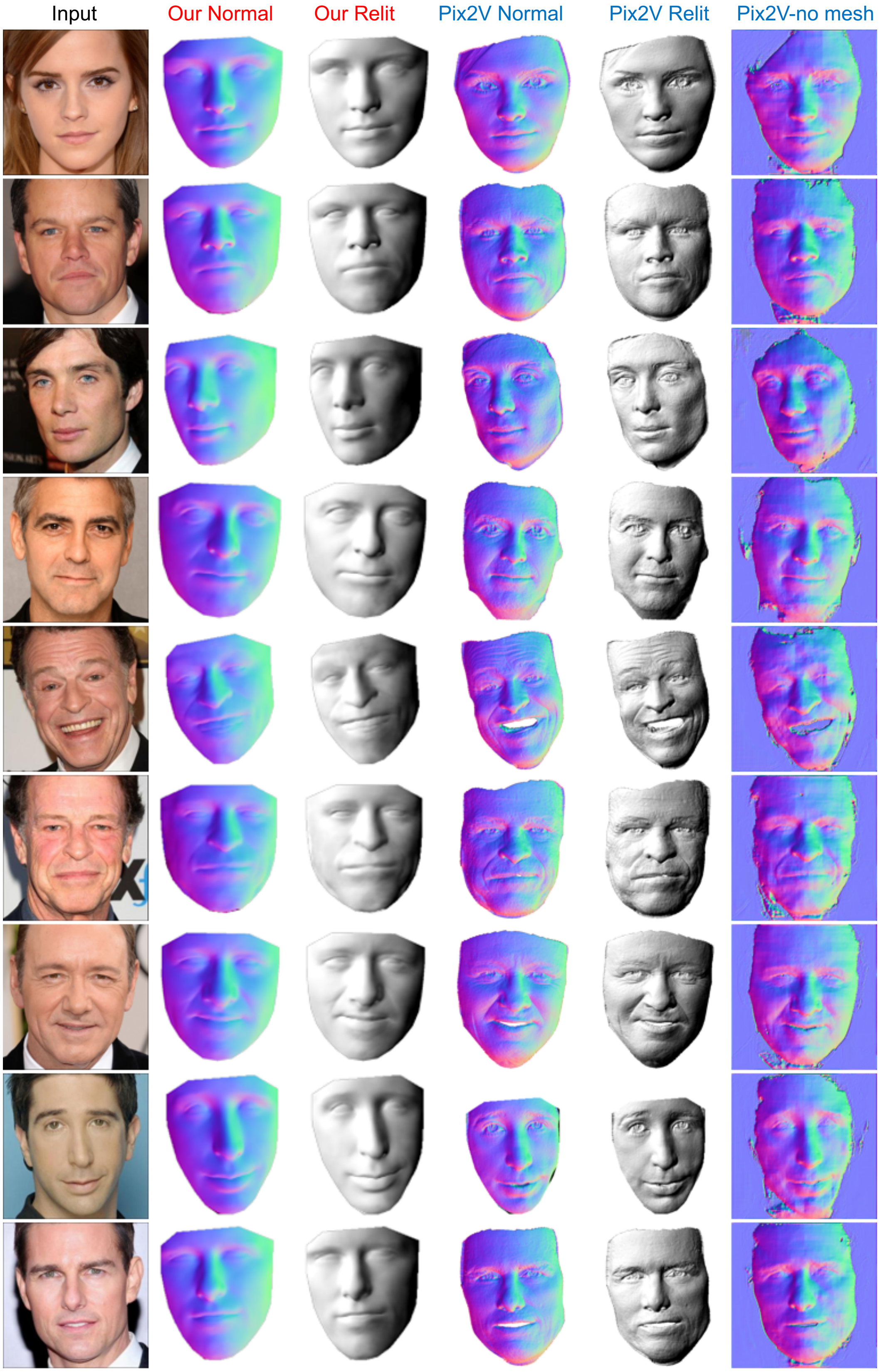

Figure 8: **SfSNet vs Pix2Vertex** [2] on the images showcased by Sela *et al*. in [2]. 'Relit' images are generated by directional lighting and uniform albedo selected to highlight the quality of the reconstructed normals. (Best viewed in color)

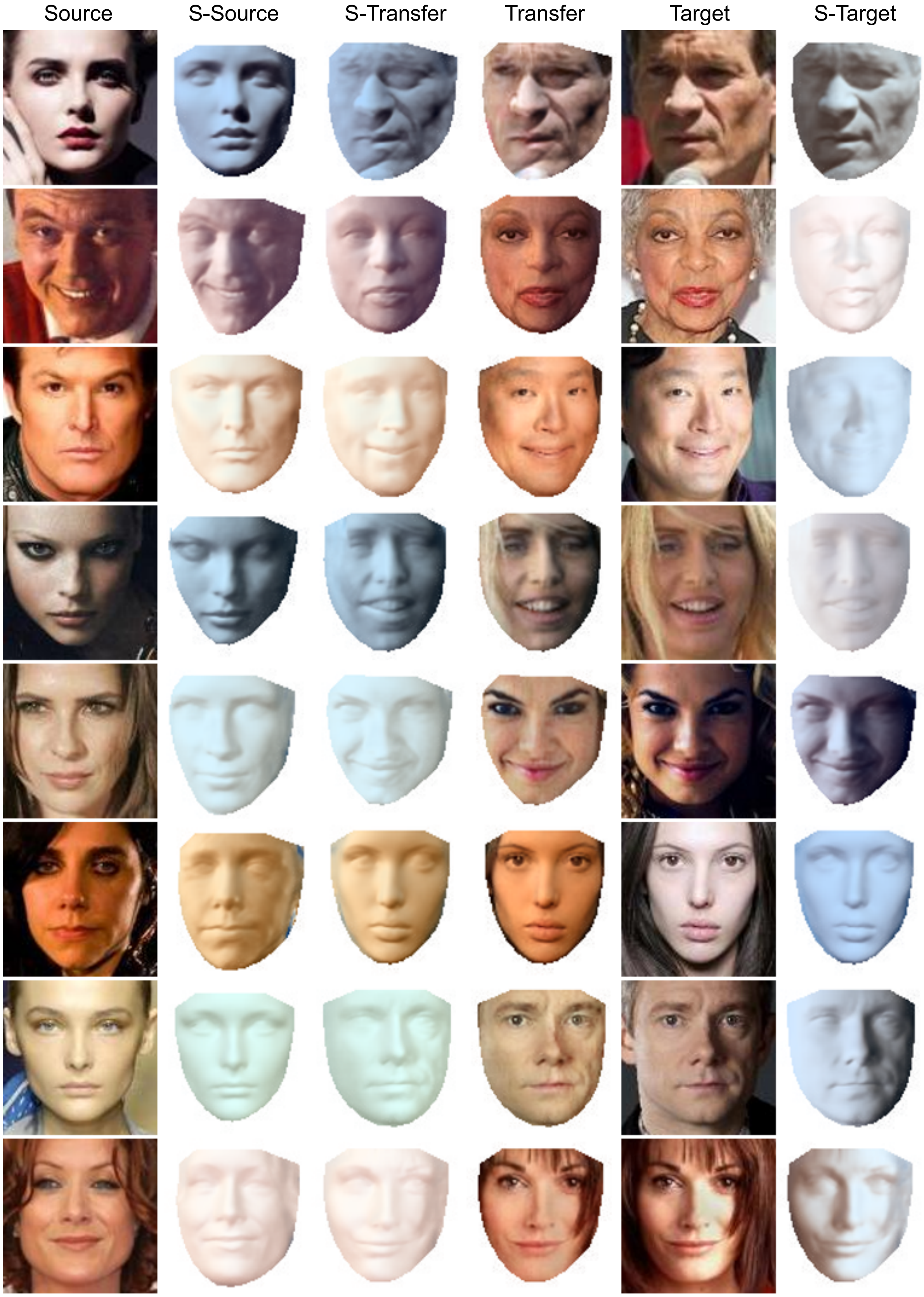

Figure 9: **Light transfer.** Our SfSNet allows us to transfer lighting of the 'Source' image to the 'Target' image to produce 'Transfer' image. 'S' refers to shading. (Best viewed in color)

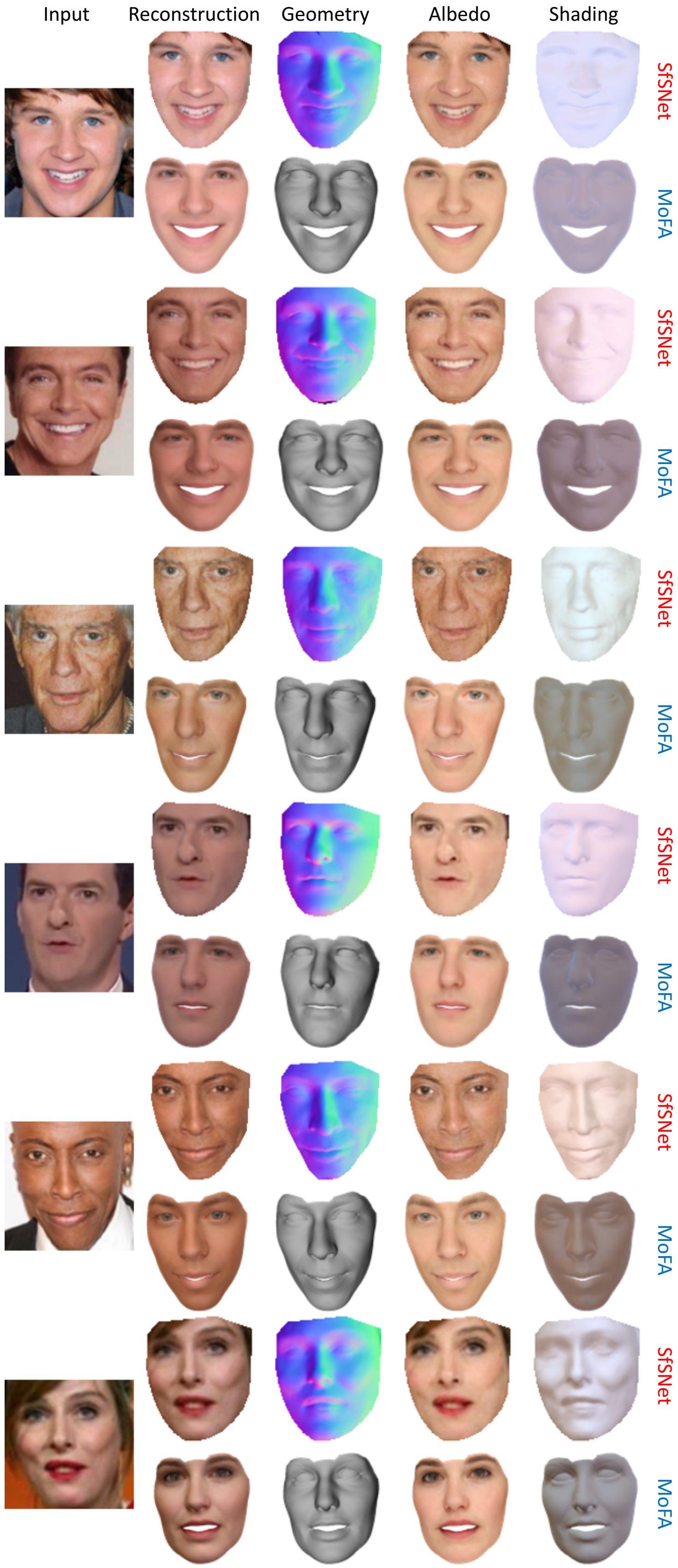

Figure 10: **Inverse Rendering. SfSNet vs 'MoFA'** [4] on the data provided by the authors. (Best viewed in color)

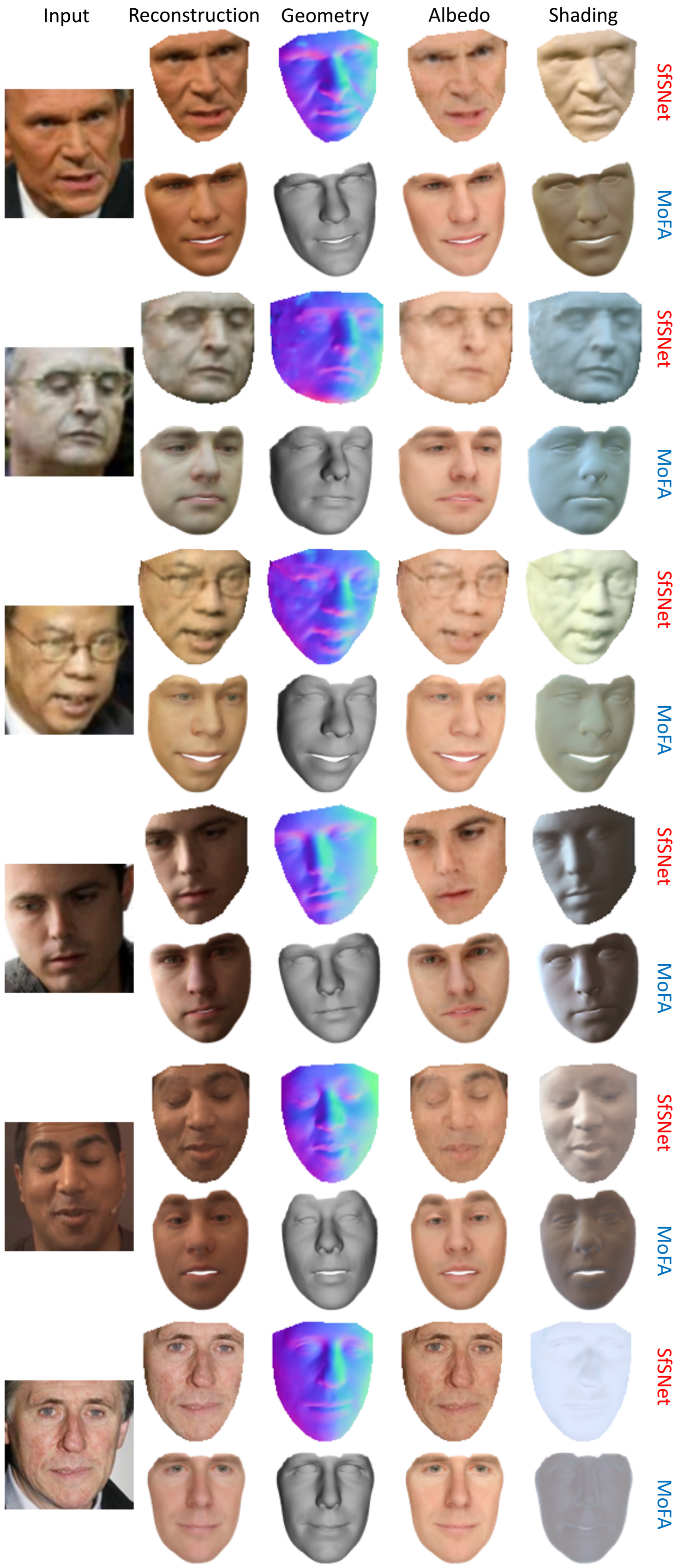

Figure 11: **Inverse Rendering. SfSNet vs 'MoFA'** [4] on the data provided by the authors. (Best viewed in color)

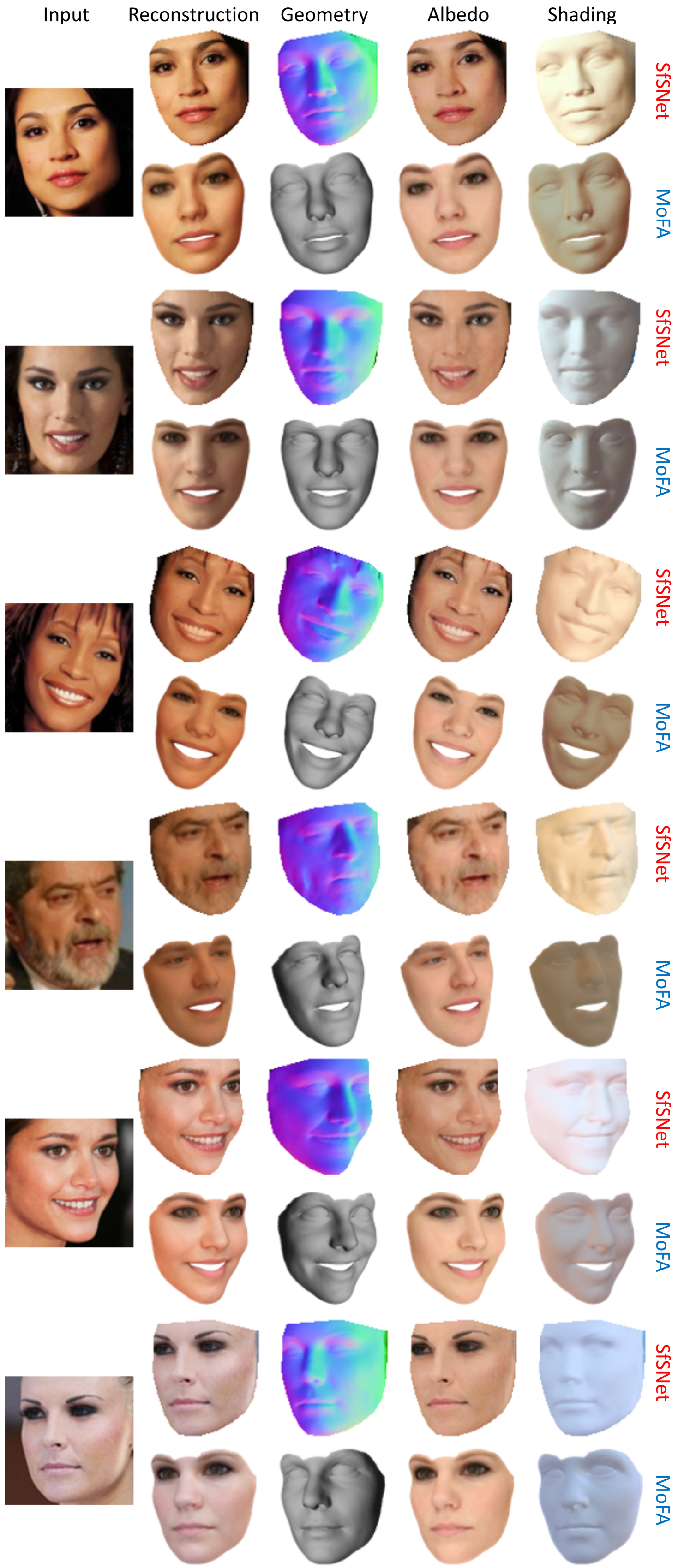

Figure 12: **Inverse Rendering. SfSNet vs 'MoFA'** [4] on the data provided by the authors. (Best viewed in color)

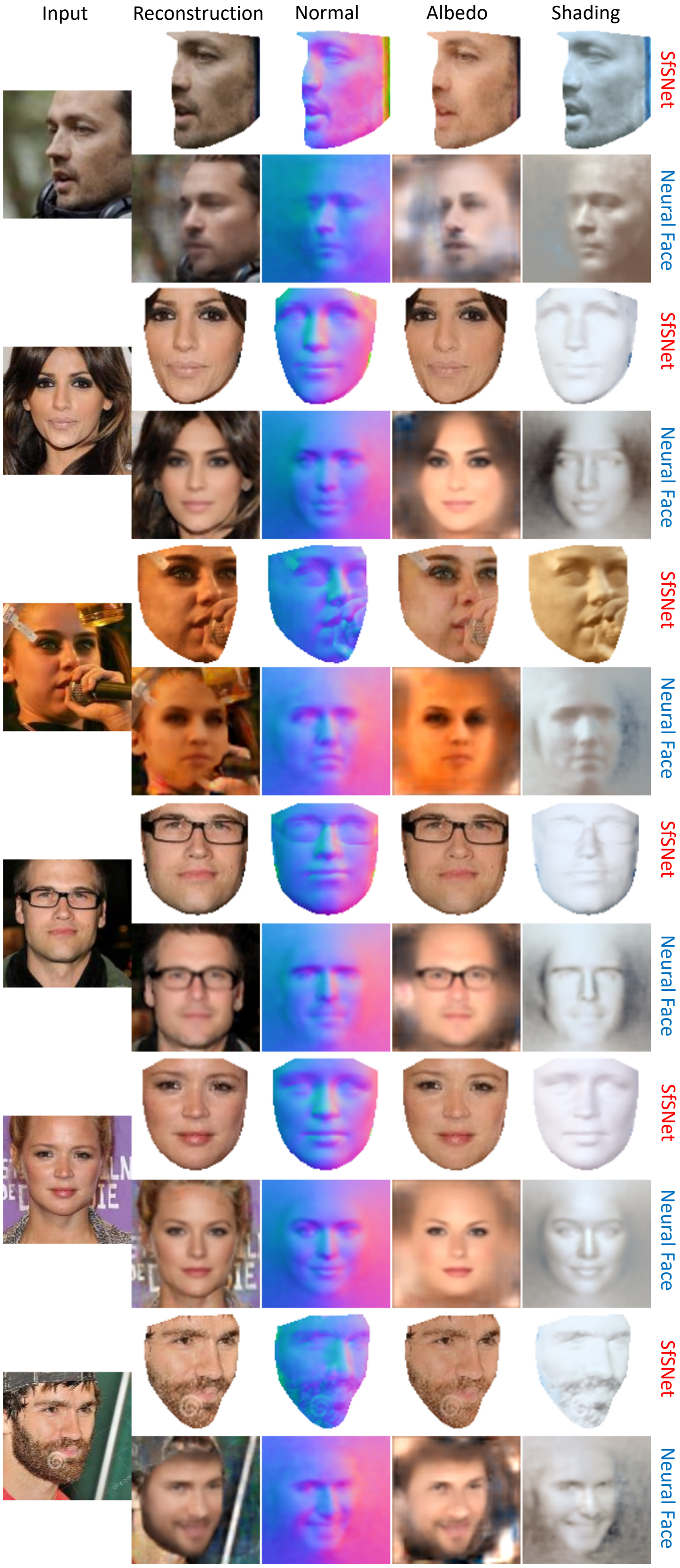

Figure 13: **Inverse Rendering. SfSNet vs 'Neural Face'** [3] on the images showcased by the authors. (Best viewed in color)