Deep Learning for Inverse Rendering

Zhenyuan Lin, Zichan Yang

December 15, 2021

1 Problem Statement

1. Deep learning performed well on inverse rendering issues, for instance, the monocular depth estimation and intrinsic image decomposition. However, the inverse rendering of uncontrolled scenes is still an unsolved problem and labels for supervised learning cannot be used.
2. Most of the classical approaches which has been used for estimating intrinsic properties require multiple images.
3. The former research related to estimation of shape requires supervision by ground truth depth.

2 Solutions

2.1 Architecture of Network

Our inverse rendering network (see Fig. 1) is an image-to-image network that regresses albedo and normal maps from a single image and uses these to estimate lighting.At inference time, our network regresses diffuse albedo and normal maps from a single, uncontrolled image and then computes least squares optimal spherical harmonic lighting coefficients. At training time, we introduce self-supervision via an appearance loss computed using a differentiable renderer and the estimated quantities.

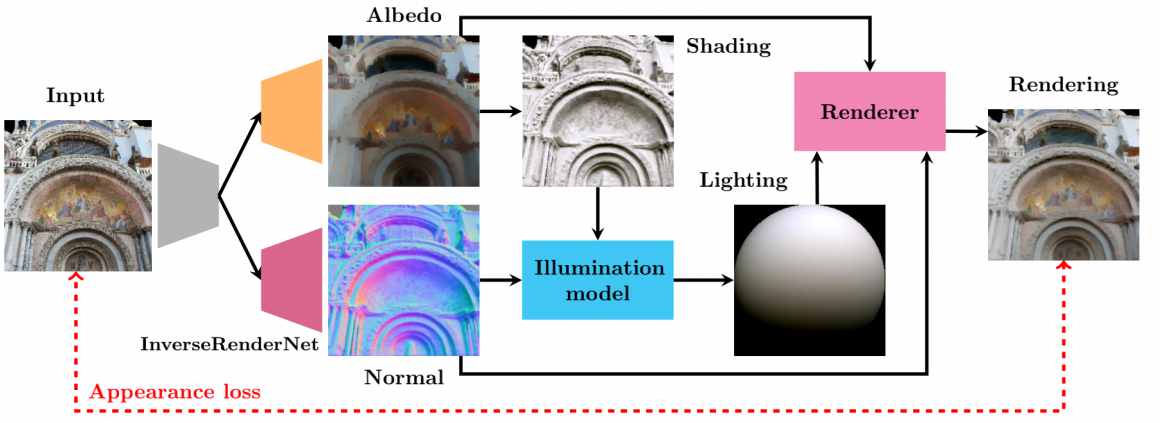

Figure 1: The Architecture of Network.

2.2 Self-supervision via Differentiable Rendering

As shown in Fig. 1, we use a data term (the error between predicted and observed appearance) for self-supervision.However, inverse rendering using only a data term is illposed (an infinite set of solutions can yield zero data error) and so we use additional sources of supervision, all of which are essential for good performance.

Given estimated normal and albedo maps and spherical harmonic illumination coefficients, we compute a predicted image. This local illumination model is straight forward to differentiate. Self-supervision is provided by the error between the predicted, ipred, and observed, iobs, intensities. We compute this error in LAB space as this provides perceptually more convincing results:

$$l_{appearance} = \|\mathbf{LAB}(\mathbf{i_{pred}}) - \mathbf{LAB}(\mathbf{i_{obs}})\| \tag{1}$$

2.3 Training Strategy

We train our network to minimise:

$$l = \omega_1 l_{appearance} + \omega_2 l_{NM} + \omega_3 l_{albedo} + \omega_4 l_{cross-rend} + \omega_5 l_{albedo-smooth} + \omega_6 l_{albedo-pseudoSup} \tag{2}$$

The preprocessed images have arbitrary shapes and orientations. For ease of training, we crop square images and resize to a fixed size. We choose our crops to maximise the number of pixels with defined depth values. Where possible, we crop multiple images from each image.We create mini-batches with overlap between all pairs of images in the mini-batch and sufficient illumination variation. Finally, before inputting an image to our network, we detect and mask the sky region using PSPNet [1]. This is because the albedo map and normal map in sky area are meaingless and it severely influences illumination estimation.

3 Conclusion

Our results show that "shape-from-shading" in the wild is possible and are far superior to classical methods. We believe the reason this is possible is because of the large range of cues that the deep network can exploit, for example shading, texture, ambient occlusion. For example, once a region is recognised as a "window", the possible shape and configuration is much restricted. There are many promising ways in which this work can be extended. First, our modelling assumptions could be relaxed, for example using more general reflectance models and estimating global illumination effects such as shadowing. Second, our network could be combined with a depth prediction network. Either the two networks could be applied in dependently and then the depth and normal maps merged, or a unified network could be trained in which the normals computed from the depth map are used to compute the losses we use in this paper. Third, our network could benefit from losses used in training intrinsic image decomposition networks. For example, if we added the timelapse dataset of [2] to our training, we could incorporate their reflectance consistency loss to improve our albedo map estimates.

References

[1] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[2] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9039–9048, 2018.