Workshop on Material Appearance Modeling (2020)
H. Rushmeier and R. Klein (Editors)

The Problem of Entangled Material Properties in SVBRDF Recovery

S. Saryazdi , C. Murphy and S. Mudur

Concordia University, Montreal, Canada

Abstract

SVBRDF (spatially varying bidirectional reflectance distribution function) recovery is concerned with deriving the material properties of an object from one or more images. This problem is particularly challenging when the images are casual rather than calibrated captures. It makes the problem highly under specified, since an object can look quite different from different angles and from different light directions. Yet many solutions have been attempted under varying assumptions, and the most promising solutions to date are those which use supervised deep learning techniques. The network is first trained with a large number of synthetically created images of surfaces, usually planar, with known values for material properties and then asked to predict the properties for image(s) of a new object. While the results obtained are impressive as shown through renders of the input object using recovered material properties, there is a problem in the accuracy of the recovered properties. Material properties get entangled, specifically the diffuse and specular reflectance behaviors. Such inaccuracies would hinder various down stream applications which use these properties. In this position paper we present this property entanglement problem. First, we demonstrate the problem through various property map outputs obtained by running a state of the deep learning solution. Next we analyse the present solutions, and argue that the main reason for this entanglement is the way the loss function is defined when training the network. Lastly, we propose potential directions that could be pursued to alleviate this problem.

Categories and Subject Descriptors (according to ACM CCS): I.4.1 [Image Procesing and Computer Vision]: Digitization and Image Capture—Reflectance

1. Introduction

The appearance of an object depends on the view (eye or camera), light source and the way in which light interacts with (gets scattered by) the surface and material of the object. For an opaque surface, this light interaction is modelled by a four-dimensional function called as the bi-directional reflectance distribution function, BRDF for short, which models the output light in any direction as a function of the incoming light in any direction. For heterogeneous materials, we often use the spatially varying BRDF, or SVBRDF for short, which has a location on the surface as additional parameters. A number of mathematical models, such as Phong, or the more physically-based Cook-Torrance, and other variants have been proposed for compact BRDF representation [KE09]. These models have a fixed number of parameters (properties of the surface and material), which take values to correctly model light reflectance behavior at a surface location. BRDF recovery essentially amounts to deriving/estimating the values of these parameters from captured images. For example, if we wish to recover the BRDF for a planar surface using, say the GGX micro-facet distribution model (based on Cook-Torrance) [WMLT07] for isotropic reflectance behavior, then we would need diffuse albedo and specular albedo for each of the colour channels, and specular roughness. In addition, local surface normals are often recovered to account for fine variations in surface geometry. For an image, each pixel is taken to represent a surface location. Hence, if we assume the RGB colour model, we would need 3 values each for diffuse and specular albedo, 1 value for specular roughness and 2 values for the normal direction. If every pixel location has to be assigned these values, then we would get 4 property maps to be recovered. Increasingly, deep learning networks are being trained to learn prediction of SVBRDFs from one or more casually captured images of the object, and with reasonable success.

The major problem in recovering BRDF properties from an image of the object arises from the following. Since the colour output seen at a pixel is a complex function of the incident light, the view and the different property values of the object's surface area which this pixel is imaging, one requires an inversion method which can disentangle these individual property values. Otherwise the same colour value may be obtained by a completely different set of property values, possibly incorrectly representing the underlying physical material and surface of the object. Given just one or a few

DOI: 10.2312/mam.20201138

images in the wild, this inversion is an ill-posed problem. However, recent efforts in training deep learning networks to learn this inversion function have shown considerable success in SVBRDF recovery as discussed next in the related work section. Usually, in attempting SVBRDF recovery, it is also assumed that the underlying surface is planar. In spite of these results, careful observation reveals that in most of these methods, even if the predicted maps when used in rendering do yield a near identical version of the ground truth image, the individual properties are often not the same as the ground truth maps, particularly the diffuse albedo and specular albedo property maps.

Given how ill-posed this problem is, we believe that it would be near impossible to predict property maps exactly identical to ground truth maps. However, accurate maps are essential when the BRDFs are used in various downstream applications, such as the following:

1. Material type classification - this requires matching/clustering of BRDFs and has vital applications in remote sensing, paint industry, food inspection, material science, recycling, etc [GGPL18].
2. Artist editable - artists in the entertainment industry often rework/change the BRDF by editing the property maps [BOR06]. Inaccurate property maps would cause significant overheads and pain.
3. Virtual object insertion in mixed reality environments - one often introduces virtual object(s) into virtual/augmented scenes. Accurate BRDFs are essential if the virtual object(s) have to appear realistic and natural in their environment, which would only be possible if light interaction between the virtual object(s) and the environment is realistically modelled [KKT11].

2. State of the Art in SVBRDF Recovery

Classical BRDF measurement approaches rely on capturing a large number of images under different calibrated viewing and lighting conditions using dedicated acquisition setups [MWL*99]. While these approaches can recover nice BRDFs, they require specialized hardware for image acquisition in addition to a large number of input images. Since then, there has been work on identifying an optimal subsample of views and lighting conditions for image acquisition [NJR15]. However, these methods are restricted to homogeneous materials and the requirement of pre-calibrated cameras for image acquisition severely restricts their use. For this paper, we focus on methods of spatially-varying BRDF recovery which use light-weight image captures of materials in the wild.

As mentioned earlier, deep learning models have shown a lot of promise in reflectance modeling from images in the wild [LDPT17,DAD*18,LSC18,DAD*19]. In [LDPT17] the traditional L2 loss over the predicted maps is used to train their deep network which generates maps that do not reproduce well the appearance of ground truth renders. [DAD*18] have shown that this loss function does not lead to predicting very accurate BRDFs nor ground truth render reproductions. Instead, in their work on recovering SVBRDF from a single flash-lit image, both [DAD*18] and [LSC18] found *r*endering loss to be a better alternative if one's goal is to recover accurate renders from the predicted maps. Rendering loss is computed by using the L1 or L2 loss between the

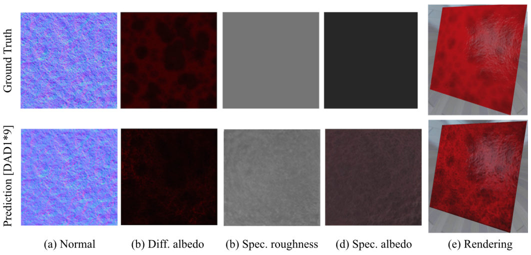

Figure 1: *Model trained with the rendering loss and L1 loss on the individual maps can generate a realistic looking render despite incorrectly assigning the red color of the material to its specular albedo to compensate for the incorrect diffuse albedo.*

rendered image using all 4 of the predicted maps and the input image obtained by a rendering which uses all 4 of the ground truth maps under the same lighting and viewing conditions. However, in their approach the specular and diffuse albedo maps recovered have errors when compared to ground truth maps, despite the fact that the final rendered image(s) looks similar to the input image(s). The success of these deep networks on predicting maps from single images is due to the strong priors that they learn on the material property, and thus it is not possible for these models to generalize to images rendered from arbitrary SVBRDF maps.

Currently, the best results are obtained using multi-image deep networks [GLD*19, DAD*19]. These are networks that use multiple images of the same material under different light and view conditions as their input. In the multiple image setup, the different views can now provide the network with more cues on what the BRDF should be, and ideally we would like the network to rely less on the learned priors about the material properties and more on the visual cues in the different images as the number of views increases. Very recent is the work by [DAD*19] which can handle an arbitrary number of input views. Similar to previous work by [GLD*19], [DAD*19] found that using a combination of L1 loss on the predicted maps and rendering loss during training helped stabilize the training procedure. However, the individual recovered SVBRDF maps still have inaccuracies, and there are often instances where the network predicts incorrect maps that render to a similar looking image, for example, by incorrectly assigning the color from the diffuse albedo map to the specular albedo map. Figure 1 shows an example of this case.

3. The Causes for Errors in Property Recovery

It should be noted that a single or a few images by themselves may not contain enough cues for one to be able to infer material properties precisely. Thus recovering material properties from a single image, or even a few images, is an ill-posed problem. Arbitrarily increasing the number of input images with different view and light directions leads to larger data collection requirements but not necessarily better quality results. Hence, one of the major goals in new research would be to recover more accurate property maps by training with a few casually captured images. As per our analysis of current deep learning solutions, there are a few causes for these

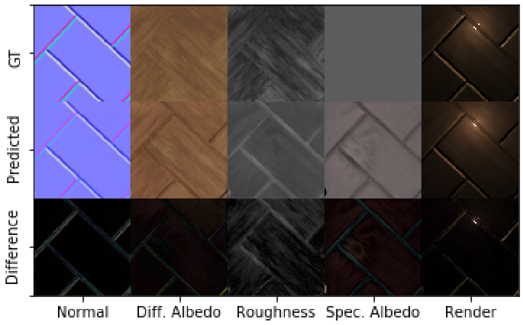

Figure 2: *Using render loss leads to the recovery of maps which create similar renders to the ground truth, despite incorrect property maps.*

inaccuracies, mainly arising from the way the loss function is defined.

- Emphasis in training is on rendered image similarity rather than material properties.
- No effort at disentangled learning of properties.
- Dependence on a few views for render comparison.

As mentioned earlier, rendering loss was shown to be more effective and hence gets used in all recent work. [DAD*18, LSC18, LXR*18, GLD*19, DAD*19]. Using this loss as opposed to the traditional L1 or L2 loss on predicted maps lets the physical meanings of each map and the interplay between them to be relegated to the update steps. Formally, the rendering loss is given by:

$$L_R(\vec{l},\vec{v}) = |R_{N,D,R,S}(\vec{l},\vec{v}) - R_{\hat{N},\hat{D},\hat{R},\hat{S}}(\vec{l},\vec{v})| \qquad (1)$$

Where $L_R(\vec{l},\vec{v})$ is the rendering loss under some light direction $\vec{l}$ and view direction $\vec{v}$, $R_{N,D,R,S}(\vec{l},\vec{v})$ is the rendering function parameterized by the 4 material maps N, D, R and S which are the predicted normal, diffuse albedo, specular roughness and specular albedo maps respectively, and $\hat{N}$, $\hat{D}$, $\hat{R}$ and $\hat{S}$ are the ground truths for those maps respecively. Since the rendering loss is light and view dependent, in practice the average of the rendering loss over multiple randomly sampled light and view directions is used for training. We note that this is the Monte Carlo method for approximating $\mathbb{E}_{\vec{l},\vec{v}}[L_R(\vec{l},\vec{v})]$. This definition of the rendering loss has several major drawbacks.

Firstly, the rendering loss under limited light and view directions has multiple global minima. This is because two very different combinations of SVBRDF maps can generate the same rendering under limited light and view directions. As a direct implication of this, models trained with rendering loss tend to compensate for the incorrectness in one of the predicted maps by modifying another map in a way that would give a similar render. An example of this is shown in Figure 2, where a model trained with rendering loss predicts a pinkish color as part of the specular map to compensate for the incorrect diffuse albedo and roughness map predictions.

Secondly, the many-to-one nature of the rendering function implies that the gradient is either zero or non-zero with respect to all 4 property maps. E.g. If during training of the network has already learned to predict three of the four maps correctly and has a mistake in one of them which causes the render to look different, the rendering loss will have non-zero gradients with respect to all 4 maps, thus making the network forget about maps it has already learned.

Thirdly, the number of light and view directions is a heuristic that needs to be selected empirically. Sampling more light and view directions would make the approximation of $\mathbb{E}_{l,v}[L_R(l,v)]$ more accurate, albeit at the cost of more computation. Using a single render to compute loss with presents many loss minima possible, shown in Figure 3, so most recent works use 9 (a heuristic) renders to compute the loss with as they find it has the best computation to test render accuracy trade-off.

4. Alleviation

1. **Disentanglement**: Various fields of research have shown that disentangling parameters in complex tasks helps to train the network to better understand the problem, which then leads to the network learning more accurate solutions for unseen data. Some examples of disentangled tasks includes learning from videos [D*17] and face image editing [SYH*17]. We believe that corresponding strategies should be attempted to disentangle material maps.
2. **Loss and training strategy**: One could look at ways to define the loss function differently, so that the network is trained to learn each property separately while at the same time using all the properties for rendering. The dependence on a discrete number of views could be eliminated if one could define the loss integrated over all light and view directions over the hemisphere.
3. **Continual learning**: Continual learning attempts to address the network problem of forgetfulness [LAM*19]. One could consider applying this strategy by treating learning of different material properties as a sequence of tasks.

References

[BOR06] Ben-Artzi A., Overbeck R. S., Ramamoorthi R.: Real-time BRDF editing in complex lighting. *ACM Trans. Graph. 25*, 3 (2006), 945–954. URL: https://doi.org/10.1145/1141911.1141979, doi:10.1145/1141911.1141979. 6

[D*17] Denton E. L., et al.: Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems* (2017), pp. 4414–4423. 7

[DAD*18] Deschaintre V., Aittala M., Durand F., Drettakis G., Bousseau A.: Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 128. 6, 7

[DAD*19] Deschaintre V., Aittala M., Durand F., Drettakis G., Bousseau A.: Flexible svbrdf capture with a multi-image deep network. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 1–13. 6, 7

[GGPL18] Guo J., Guo Y., Pan J., Lu W.: Brdf analysis with directional statistics and its applications. *IEEE transactions on visualization and computer graphics PP* (10 2018). doi:10.1109/TVCG.2018.2872709. 6

[GLD*19] Gao D., Li X., Dong Y., Peers P., Xu K., Tong X.: Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 134. 6, 7

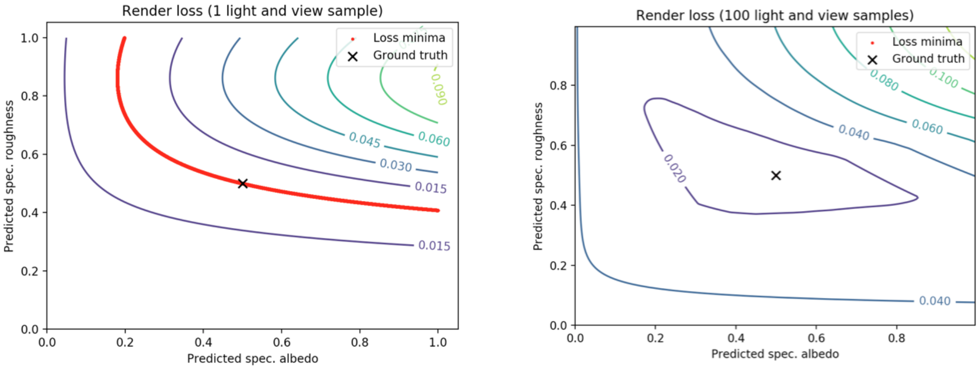

Figure 3: ***Contour plot of the rendering loss landscape*** *with respect to the specular albedo and roughness. Multiple local minima exist when using one light and view sample (left). As we increase light and view samples, the problem of multiple local minima is reduced (right).*

[KE09] Kurt M., Edwards D.: A survey of brdf models for computer graphics. *10.1145/1629216.1629222 ACM SIGGRAPH Computer Graphics* (05 2009). doi:43. 5

[KKT11] Kühtreiber P., Knecht M., Traxler C.: Brdf approximation and estimation for augmented reality. In *15th International Conference on System Theory, Control and Computing* (Oct 2011), pp. 1–6. 6

[LAM*19] Lange M. D., Aljundi R., Masana M., Parisot S., Jia X., Leonardis A., Slabaugh G., Tuytelaars T.: A continual learning survey: Defying forgetting in classification tasks, 2019. arXiv:1909.08383. 7

[LDPT17] Li X., Dong Y., Peers P., Tong X.: Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Trans. Graph. 36*, 4 (July 2017). URL: https://doi.org/10.1145/3072959.3073641, doi:10.1145/3072959.3073641. 6

[LSC18] Li Z., Sunkavalli K., Chandraker M.: Materials for masses: Svbrdf acquisition with a single mobile phone image. In *Computer Vision – ECCV 2018* (Cham, 2018), Ferrari V., Hebert M., Sminchisescu C., Weiss Y., (Eds.), Springer International Publishing, pp. 74–90. 6, 7

[LXR*18] Li Z., Xu Z., Ramamoorthi R., Sunkavalli K., Chandraker M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Trans. Graph. 37*, 6 (Dec. 2018). URL: https://doi.org/10.1145/3272127.3275055, doi:10.1145/3272127.3275055. 7

[MWL*99] Marschner S. R., Westin S. H., Lafortune E. P., Torrance K. E., Greenberg D. P.: Image-based brdf measurement including human skin. In *Rendering Techniques' 99*. Springer, 1999, pp. 131–144. 6

[NJR15] Nielsen J. B., Jensen H. W., Ramamoorthi R.: On optimal, minimal brdf sampling for reflectance acquisition. *ACM Transactions on Graphics (TOG) 34*, 6 (2015), 186. 6

[SYH*17] Shu Z., Yumer E., Hadap S., Sunkavalli K., Shechtman E., Samaras D.: Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5541–5550. 7

[WMLT07] Walter B., Marschner S., Li H., Torrance K.: Microfacet models for refraction through rough surfaces. pp. 195–206. doi:10.2312/EGWR/EGSR07/195-206. 5