Pacific Graphics 2018
H. Fu, A. Ghosh, and J. Kopf
(Guest Editors)

Volume 37 (2018), Number 7

Single Image Surface Appearance Modeling with Self-augmented CNNs and Inexact Supervision

Wenjie Ye[1,3] Xiao Li[2,3] Yue Dong[3] Pieter Peers[4] Xin Tong[3]

[1] Tsinghua University
[2] University of Science and Technology of China
[3] Microsoft Research Asia
[4] College of William & Mary

Abstract

This paper presents a deep learning based method for estimating the spatially varying surface reflectance properties from a single image of a planar surface under unknown natural lighting trained using only photographs of exemplar materials without referencing any artist generated or densely measured spatially varying surface reflectance training data. Our method is based on an empirical study of Li et al.'s [LDPT17] self-augmentation *training strategy that shows that the main role of the initial approximative network is to provide guidance on the inherent ambiguities in single image appearance estimation. Furthermore, our study indicates that this initial network can be inexact (i.e., trained from other data sources) as long as it resolves the inherent ambiguities. We show that the single image estimation network trained without manually labeled data outperforms prior work in terms of accuracy as well as generality.*

CCS Concepts

•***Computing methodologies*** → ***Reflectance modeling;***

1. Introduction

Modeling the appearance of a spatially varying material is a challenging problem that has spurred significant research interest over the past decade. Recent solutions have endeavored to estimate spatially varying surface appearance from a single image of a planar surface under a variety of lighting conditions, ranging from flash lights [AAL16, DAD*18] to uncontrolled natural illumination [LDPT17]. Appearance estimation from a single image is especially difficult and a highly ill-posed problem, resulting in a variety of ambiguities that are resolved by exploiting prior knowledge on either the lighting or material properties. For example without making any assumptions, one trivial (and undesired) solution would be to bake in the lighting, normal variations, and specular properties in the diffuse albedo texture.

In this work, we focus on estimating the spatially varying appearance from a planar surface under unknown and uncontrolled natural lighting. Recently, Li et al. [LDPT17] proposed to use deep learning to leverage prior knowledge of natural materials. A key challenge for any deep learning solution is to gather a sufficiently large labeled training dataset. Li et al. introduce a novel training strategy named *self-augmentation* that leverages the information embedded in a large set (> 1000) of unlabeled photographs of spatially varying materials to augment a rough approximative convolutional neural network trained on a very small set (< 100) of labeled training data consisting of a diffuse albedo map, a normal map, and specular reflectance properties for each training material.

In this paper we explore two fundamental questions via a series of carefully crafted experiments regarding Li et al.'s self-augmentation training strategy in the context of single image appearance estimation under uncontrolled lighting. First, can self-augmentation train a network from only unlabeled data? Second, what role does the initial approximative network play in the training process? Is it possible to relax the accuracy of this initial network? To answer these questions, we design a synthetic test dataset and perform a series of experiments. From these experiments we conclude that self-augmentation cannot resolve the inherent ambiguities in single image appearance estimation without proper guidance from the labeled training data. Furthermore, we also show that the initial network does not need to be trained on labeled data from the same distribution as the unlabeled training data, as long as it resolves the ambiguities inherent to spatially varying appearance estimation of the target material distribution under unknown natural lighting from a single image. We denote such a training set as *inexact* in contrast to the commonly used *exact* training data drawn from the exact same distribution as the target distribution. We leverage this knowledge, and propose a method for synthesizing *inexact* labeled data (i.e., spatially varying appearance maps drawn) directly from the unlabeled dataset that yields a network

that not only outperforms Li et al.'s [LDPT17] solution, but also produces a single network suited for multiple material categories.

In contrast to prior and concurrent work that either uses multiple images [AWL15, HSL*17], requires a specific lighting condition [DAD*18,LSC18], or requires labeled training data [LDPT17, DAD*18, LSC18], our method has the ability to predict spatially varying reflectance properties using only information gathered from photographs mined from uncontrolled online photo-repositories, and without the need for artist generated or densely measured spatially varying appearance training data.

In summary, our contributions are:

- a synthetic training dataset for qualitative and quantitative evaluation of deep learning training strategies on spatially varying appearance;
- novel insights regarding self-augmentation and the introduction of *inexact* but ambiguity free supervision; and
- a novel method for training a general network for estimating real-world spatially varying appearance from a single image under unknown natural lighting using only unlabeled training data.

2. Related Work

There exists an extensive body of prior research on reflectance modeling. For brevity, we focus our discussion of related work on methods that estimate the surface reflectance from a single RGB image. We refer to the comprehensive surveys of Dorsey et al. [DRS08] and of Weinmann and Klein [WK15] for an in-depth overview of general reflectance modeling. Modeling surface reflectance from a single image is an ill-posed problem, and several strategies have been employed to better constrain the system by either restricting the material properties, or by constraining the lighting.

Restricted Material Properties Early work in reflectance estimation from a single image assumes a homogeneous object with known geometry and exploits either prior distributions of (unknown) natural lighting and/or natural materials [RZ10, LN12, LN16]. Rematas et al. [RRF*16] leverage a convolutional neural network to estimate the reflectance map from a single photograph of a homogeneous object of unknown shape and under unknown illumination. This work was further extended [GRR*17] to estimate the reflectance properties. Unlike the above methods that are limited to homogeneous materials, Barron et al. [BM15] allow for spatial albedo variations for purely Lambertian materials, and solve for shape, (low frequency) lighting, and albedo and normal maps. Our method is not limited to diffuse and/or homogeneous materials.

Constraints on Lighting Conditions An alternative strategy to better constrain the estimation of reflectance properties relies on active illumination. Wang et al. [WSM11] use a step-edge lighting condition to infer spatially varying surface normals, as well as the reflectance properties, of a homogeneous material. Xu et al. [XNY*16] recover piecewise constant surface reflectance from an optimized near-field observation lit by a directional light source. Aittala et al. [AAL16] exploit the self-similarity of stationary materials to estimate spatially varying normals and reflectance properties using a deep texture synthesis framework from a single photograph of a planar material sample under flash lighting. Instead of relying on active illumination, Oxholm and Nishino [ON12,ON16] rely on passive but known illumination to recover the shape and homogeneous surface reflectance from a single photograph of an object under natural illumination. All these methods require either active illumination or full knowledge on the lighting, which excludes their applicability to input photographs acquired under uncontrolled capture conditions, such as those mined from internet image repositories. Our method does not make any assumptions on the incident lighting, while retaining the capability to estimate plausible spatially varying surface reflectance properties from a single image.

Deep Learning based Methods In concurrent work, Deschaintre et al. [DAD*18] use a flash-lit photograph as input to a convolutional neural network designed to consider global information provided by local lighting and texture detail to estimate spatially varying surface reflectance. In other concurrent work, Li et al. [LSC18] propose a new network architecture and a differentiable densely connected conditional random field post-processing step trained on a large synthetic training data set for single image reflectance recovery under arbitrary environment lighting augmented with additional flash lighting to further regularize the process. Li et al. also demonstrate that their architecture is suited for reflectance recovery from environment lighting only. Both the concurrent works of Deschaintre et al. [DAD*18] and Li et al. [LSC18] introduce improved deep learning architectures for reflectance estimation, while relying on large synthetic training data sets. In this paper, we focus on the complementary problem of reducing the required amount of labeled training data.

Closest related to our work is Li et al.'s [LDPT17] convolutional neural network based solution for estimating spatially varying surface reflectance for a particular material class (e.g., wood, plastic, and metal) from a single photograph under an unknown natural lighting condition. A key issue in training such a network is to gather a sufficiently large training dataset of densely measured spatially varying surface reflectance. Li et al. propose a novel training strategy called *"self-augmentation"* that only requires a small labeled training set augmented by a larger collection of unlabeled photographs of the target class of materials. In self-augmentation, an initial approximative convolutional neural network is first trained from a small set of labeled training data consisting of spatially varying reflectance parameters and corresponding images of the material under natural lighting. Next, the network is refined in an iterative fashion using the unlabeled training data and knowledge of the exact inverse process of the target network (i.e., rendering of reflectance parameters under natural lighting). The current approximative network is used to generate provisional reflectance parameters from the unlabeled photographs, and for each set of provisional reflectance parameters, a corresponding image is rendered under a randomly chosen natural lighting condition. This rendered image, in conjunction with the provisional reflectance parameters, forms a valid labeled training pair and is used to further train the estimation network. To avoid drift, self-augmentation alternates between training from labeled training data and generated training data (from the unlabeled images).

In this paper, we build on, and significantly expand on Li et al.'s self-augmentation training strategy by removing the need of

labeled data, enabling us to train a network from just unlabeled photographs. Furthermore, the resulting network is not geared towards a particular material class, but general.

3. Empirical Study of Self-augmentation

Before introducing our novel training method for convolutional neural networks for estimating the spatially varying surface reflectance of a planar exemplar from a single image under unknown lighting in Section 4, we first conduct a series of experiments to answer two fundamental questions regarding Li et al.'s [LDPT17] *self-augmentation* training strategy, namely: (1) what is the role of the initial approximative network, and (2) how exact does this initial network need to be. The answers to these questions will inform the design of our novel training strategy that allows us to train such networks from only a collection of unconstrained photographs of spatially varying materials, i.e., *without* relying on any labeled training data.

To conduct our experiments and answer the above two questions, we first introduce a novel synthetic training dataset (Section 3.1) for spatially varying appearance modeling. Modeling the appearance distribution of real-world materials is a challenging and interesting research question. However, we note that in order to explore the fundamental questions we do not need to exactly model the space of physical appearance, but only need a synthetic dataset that shares the typical characteristics of real world appearance. Given this synthetic dataset, we perform two carefully crafted empirical experiments (Section 3.3 and 3.4) to better understand the limits of self-augmentation as well as to gain a better intuition on the role of the initial approximative network in self-augmentation. The outcome of this empirical study leads us to conclude that (1) self-augmentation requires an approximative initial network to resolve the ambiguities inherent to reflectance estimation from a single image, and (2) the labeled data used to train the initial network does not need to be exact, and can be sampled from a different material distribution as the unlabeled images (and thus the target distribution), as long as it resolves the ambiguities embedded in the target distribution.

3.1. Synthetic Training Dataset

Ideally, the study of the properties of self-augmentation in the context of surface reflectance estimation from a single image under unknown lighting should be performed on real measured datasets that perfectly reflect the properties of the target materials. However, one of the key motivations of Li et al. [LDPT17] for introducing self-augmentation was to address the lack of sufficiently large measured spatially varying appearance datasets. The lack of such datasets also impedes our study of the properties of self-augmentation as we also require a large enough validation and training dataset. Instead we opt to create a synthetic training and validation dataset. However, such a dataset needs to be carefully designed to mimic the relevant properties of real-world materials. Accurate procedural modeling of physical materials is still an open research problem. Fortunately, we note that for the purpose of the empirical study, we do not require an exact replica of the distribution of real-world materials, but only require a dataset that covers a wide range of texture distributions that include low frequency and high frequency features, and supports easy generation of samples.

As in Li et al. we will focus on spatially varying materials with a homogeneous specular component and spatially varying diffuse texture and surface normals. We will rely on a random mixture of two procedural texture primitives based on Perlin noise [Per02] for the spatially varying diffuse albedo and surface normals, and a Ward BRDF [War92] with random parameters for the homogeneous specular component.

Procedural Texture Primitives We define the spatially varying diffuse albedo map and normal map through two procedural texture primitives: a smooth map f_p and a map that exhibits sharp texture features f_e.

We directly use the Perlin noise function [Per02] as the smooth texture primitive: $f_p(x,y,s)$, where x and y are the 2D surface coordinates, and s is the seed vector of random variables. The Perlin noise grid is defined by a spacing of distance d where $1/d$ is uniformly sampled (per texture) in $[10.0, 20.0]$. The random sampling of d ensures that our synthetic dataset contains multiple scales of smooth features.

Natural textures exhibit not only smooth features, but also sharp edges. We therefore define a second texture primitive to model this:

$$f_e(x,y,s) = \begin{cases} 0, & 0 \leq f_p(x,y,s) < 0.499 \\ 0.5(f_p(x,y,s) - 0.499), & 0.499 \leq f_p(x,y,s) \leq 0.501 \\ 1, & 0.501 < f_p(x,y,s) \leq 1. \end{cases} \quad (1)$$

Diffuse Albedo Map We procedurally generate diffuse albedo maps as a mixture of smooth and sharp texture features:

$$A(x,y) = c_0 + \sum_{i=1}^{m} c_i f_p(x,y,s_i) + \sum_{i=1}^{n} c_{(m+i)} f_e(x,y,s_{(m+i)}), \quad (2)$$

where c_i are random *RGB* color triplets, and c_0 is the texture's base color. Thus each diffuse texture is characterized by the set: $(c_0, ..., c_{(m+n)}, s)$. We set m and n, the number of smooth and sharp components respectively, such that $m+n$ is either equal to 1 or 2. We generate the same number of diffuse albedo textures for each of the 5 combinations of m and n.

To ensure a uniform color distribution over *all* diffuse albedo textures, we perform a histogram regularization over all pixels and textures.

Normal Map The normal maps are generated from the generated diffuse albedo maps by converting each diffuse albedo map to a height field based on each pixel's intensity. To ensure a good distribution of height variations, we normalize each height map to the $[0,1]$ range, and subsequently scale it by a global scale factor uniformly chosen from $[-0.1, 0.1]$. Finally, we convert to the height map to a normal map via discrete differentiation. Note however, that we do not *"link"* the generated normal and albedo maps; during training we will consider all possible combinations (Section 3.2).

Specular Surface Reflectance Modeling We assume a homogeneous specular component over the surface modeled by a Ward BRDF [War92] with the logarithm of the specular albedo uniformly

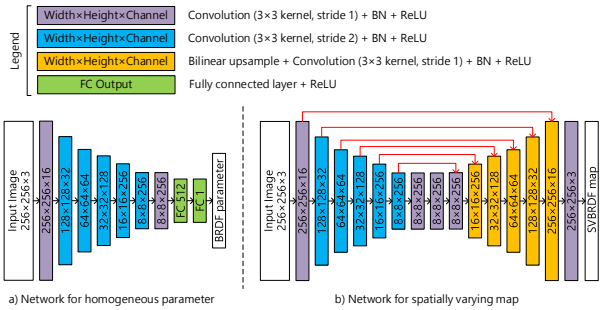

Figure 1: *Overview of the network structure. The* left *network structure is used for estimating the homogeneous specular albedo and roughness parameters. No activation layer was used for the last fully connected layer, and we predict the logarithm of the respective parameters. The* right *network is similar to the network structure by Li et al. [LDPT17] and it estimates the spatially varying diffuse albedo and normal maps. The last convolution layer foregoes batch normalization and uses a Sigmoid activation function instead of the common ReLU function.*

sampled in $[\log(0.01), \log(0.4)]$ and the logarithm of the roughness uniformly sampled in $[\log(0.01), \log(0.6)]$.

3.2. Network Structure and Training Settings

Convolution Neural Network Structure Our network structure follows Li et al.'s [LDPT17] network structure, with a slightly modified specular subnetwork as we observed some overfitting in Li et al.'s network. Figure 1 summarizes our network structure. Identical to Li et al., our network for estimating the diffuse albedo map and normal map follows a U-net structure which consists of a downsampling encoder stage followed by a upsampling stage with corresponding layers linked through "jump links". The network structure for estimating the specular albedo and roughness uses an identical downsampling encoder stage as the diffuse albedo and normal map networks. The decoder stage consists of two fully connected layers: the first reduces the output to a 512 length vector, and the second fully connected layer reduces this further to a single output value, i.e., the logarithm of the specular albedo and roughness respectively. We use four fully separated networks to estimate each of the four components (i.e., spatially varying diffuse albedo map, spatially varying normal map, (log) specular albedo, and (log) specular roughness). The four networks are trained using a $L1$ loss on each of the components as opposed to the $L2$ loss used by Li et al. [LDPT17]. We use this network structure for all experiments in this paper. To provide a fair comparison, we also retrain Li et al.'s SA-SVBRDF networks for wood, metal, and plastic using this network structure and the $L1$ loss functions.

Decorrelated Training As we generate normal maps directly from the diffuse albedo maps, the variations between both will be correlated. To cover a broader range of materials, we decorrelate diffuse albedo textures and surface normals maps using a novel outer-product training strategy. At the beginning of each epoch, we place all diffuse albedo maps, normal maps, specular albedos and roughness in four separate sets (i.e., one for each reflectance component). Then, we generate new training samples by randomly selecting and combining a component from each of the four sets. This process corresponds to taking a random sample from the outer-product space of the four reflectance component sets. Over a large number of epochs this will result in the network being trained on each possible combination from the outer-product over each of the components. This training strategy does not only decorrelate diffuse albedo textures and normal maps, it also increases the size of the learned appearance space. We apply this outer-product training strategy to all experiments in this paper.

Training For self-augmentation we require both labeled and unlabeled training exemplars. We generate 5,000 labeled exemplars (i.e., we synthesize 5,000 diffuse albedo maps, normal maps, specular roughness values, and specular albedo values), and apply a random reordering of the components to decorrelate diffuse albedo and normals. Similar to Li et al. [LDPT17] we also generate, during training, a corresponding input image under a randomly chosen and rotated environment map. We also generate 10,000 unlabeled exemplars. We follow a similar process as for the labeled exemplars, except that we render (and retain) the corresponding image before training, and discard the reflectance components. We also generate 1,000 validation exemplars similar to the labeled data, for which we pre-render and fix the images before training, and we retain all reflectance components. We use the same validation set (and corresponding rendered images) for all experiments.

We train the reflectance estimation network using the Adam optimizer. We first train the initial approximative network for 50,000 iterations on labeled training data, followed by 50,000 self-augmentation training steps. For each self-augmentation step, we train on a mini-batch of 16 labeled data exemplars, and a mini-batch of the same size of unlabeled data. We set the initial learning rate to 0.001 and apply an inverse time decay (rate = 0.0001): $learningrate = 0.001/(1 + 0.0001 * step)$.

3.3. Experiment 1: Ambiguity

Estimating the surface reflectance from a single image under unknown lighting is a highly ill-posed problem with many ambiguities. For example, a trivial solution would be to assign all reflectance to the diffuse albedo texture, with zero specular, and set the normals to the equivalent of a flat surface. To resolve such ambiguities, classic data-driven methods bias the solution to a preferred solution embedded in the training data. Self-augmentation has a similar requirement to ensure correct resolution of ambiguities. We posit that the self-augmentation training strategy requires complete (i.e., ambiguity free) labeled training data in order to train a correct network, and that no matter how much unlabeled training data is used, ambiguities unaddressed by the labeled data cannot be corrected. We validate this conjecture empirically, and generate three different target spaces using the texture primitives described in Section 3.1:

- **Regular Target:** is the full target space as described in Section 3.1.
- **Trivial Target:** is the target space where all reflectance is assigned to the diffuse albedo texture. To construct this space, we

Model	Diffuse Loss (e-2)	Normal Loss (e-2)	Specular Albedo Loss (e-1)	Specular Roughness Loss (e-1)
Regular Target	4.843	3.179	3.470	3.844
Trivial Target	11.25	6.768	41.49	18.25
Mixture Target	8.605	6.748	7.335	7.258

Table 1: *Average loss for each of the reflectance components for the models trained on the* Regular*,* Trivial*, and* Mixture *targets.*

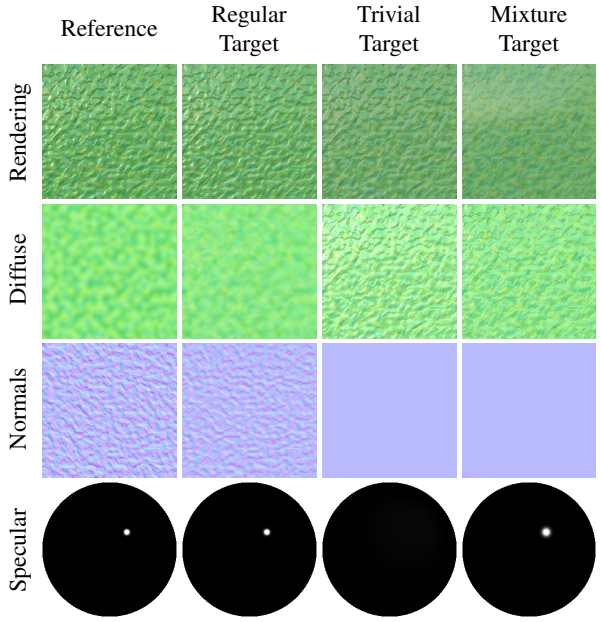

Figure 2: *A comparison of the estimated reflectance component results from the networks trained on the* regular *(2nd column),* trivial *(3rd column), and* mixture *(last column) targets. From top to bottom shows the rerendered image, recovered diffuse albedo map, normal map, and the specular BRDF applied to a sphere under a directional light source. The "rendering" row shows the input image for the "reference" column, and the rendering of estimated components under the input lighting for the other columns.*

gather all the rendered images from the training dataset, and change their reflectance labels such that the diffuse albedo texture equals the rendered image, the normal map is uniformly aimed up, and the specular albedo and roughness are set to fixed values indicating very weak specular (0.001 and 0.5 respectively).

- **Mixture Target:** consists of 50% from the *regular* target and 50% from the *trivial* target.

Both the *regular* and *trivial* target space are complete, but mutually ambiguous, and the *mixture* target is highly self-ambiguous. For each set we generate a labeled, an unlabeled, and a validation set as described before, and train a reflectance estimation network with self-augmentation. Note that due to the construction, the unlabeled training data is the same for all three cases.

Table 1 and Figure 2 show the results for each of the three cases. The network trained for the *regular* target space produces plausible results for the validation test set, and the model is able to successfully separate the diffuse and specular components for a given image. Similarly, the network trained on the *trivial* target space, produces the expected output where all the specular reflectance is assigned to the diffuse albedo. Finally, the network trained on the *mixture* target produces a result that is neither fully separated or fully diffuse, defaulting to a mix of both. Hence, this network is unable to solve the ambiguities inherent to single image reflectance estimation when the labeled training data does not adequately resolve these ambiguities.

From this experiment we conclude that although self-augmentation uses a perfect inverse mapping (i.e., rendering) to refine the network from the unlabeled photographs, it ultimately relies on the initial network trained on the labeled data to disambiguate the unlabeled training data. If the labeled training data is ambiguous (e.g., as in the case of the *mixture* target), then self-augmentation cannot correct the behavior of the network. Furthermore, even if all labeled training data are complete with respect to a specific space (e.g., all reflectance is due to the diffuse albedo), then self-augmentation cannot alter the networks disambiguation preference to another space (e.g., correctly separated diffuse and specular reflectance).

Conclusion The role of the initial approximative network, and thus labeled training data, is to provide cues on how to disambiguate the input in the target domain.

3.4. Experiment 2: Exactness of Initial Network

A key unwritten constraint in most data-driven methods is that the labeled training data is drawn from the same distribution as the target space. However, in practice it is not always possible to collect such labeled data, for example because the target distribution is not known, or because it is very expensive to generate labeled training samples. In many cases, using inexact labeled training samples drawn from a different distribution incurs a performance penalty. The goal of this second experiment is to validate the robustness of the self-augmentation training strategy, and in particular of the initial approximative network, with respect to inexact training data. The experiment detailed in this subsection will show that self-augmentation does not require exact labeled training data, and that any inexact set of labeled training data can be used as long as it correctly addresses the ambiguities in the desired *target space*.

For this experiment we again create three different training sets based on the synthetic texture primitives introduced in Section 3.1:

- **Regular Target** is the full target space as before and as described in Section 3.1.
- **Smooth Target** only uses the *smooth* texture primitive f_p.

Test set	Model	Iterations	Diffuse Loss (e-2)	Normal Loss (e-2)	Specular Albedo Loss (e-1)	Specular Roughness Loss (e-1)
Sharp	Regular	50k	5.710	2.514	3.695	4.856
	Regular-to-Sharp	100k	5.258	2.370	3.492	3.831
	Regular-to-Smooth	100k	6.874	2.684	3.711	3.906
	Sharp-to-Sharp	100k	4.237	1.750	3.157	4.038
	Sharp-to-Smooth	100k	6.094	3.500	3.854	5.344
	Smooth-to-Sharp	100k	11.91	6.210	10.48	16.50
	Sharp	50k	4.714	1.852	3.467	5.091
	Smooth	50k	15.87	10.96	12.30	13.62
Smooth	Regular	50k	5.724	4.252	4.298	4.694
	Regular-to-Sharp	100k	5.842	4.942	4.102	3.910
	Regular-to-Smooth	100k	5.342	4.244	4.107	3.637
	Smooth-to-Smooth	100k	4.718	3.285	3.132	3.042
	Sharp-to-Smooth	100k	10.33	8.957	6.739	7.222
	Smooth-to-Sharp	100k	6.761	6.086	3.733	3.678
	Sharp	50k	9.808	8.072	6.534	6.793
	Smooth	50k	4.895	3.532	3.412	3.433

Table 2: *Average loss for each of the reflectance components on networks trained on different combinations of the* Regular, Sharp, *and* Smooth *targets for the labeled and unlabeled datasets. The loss is computed for the* Sharp *and* Smooth *validation test sets.*

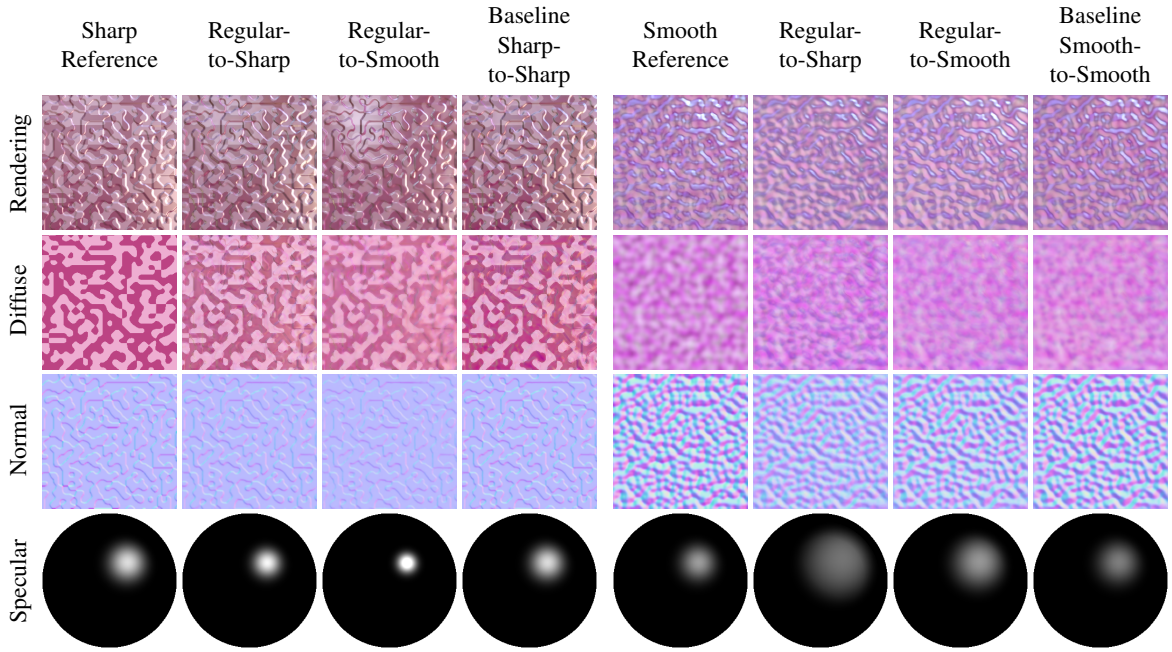

Figure 3: *Two selected examples of reflectance estimation with a network trained on inexact labeled data. The input of the first example (left) is drawn from the* Sharp *validation set, and from the* Smooth *validation set for the second example (right). A baseline comparison is shown for each example (*Sharp-to-Sharp *and* Smooth-to-Smooth *respectively).*

- **Sharp Target** only uses the *sharp* texture primitive f_e.

We also generate two additional validation test sets in the *smooth* and *sharp* target space. To demonstrate the exactness of the training dataset, we consider different combination of labeled training exemplars and unlabeled exemplars drawn from the different target spaces. We will denote by *"x-to-y"* the network with the initial network trained with labeled data drawn from *"x"* and self-augmented with unlabeled training data drawn from *"y"*. Table 2 shows the average loss for each network when validated on a test set drawn from the *smooth* or *sharp* target spaces. Figure 3 shows a visualization of two selected examples from this experiment.

From these results we can see that:

- *regular-to-sharp* outperforms *regular-to-smooth* on the *sharp*

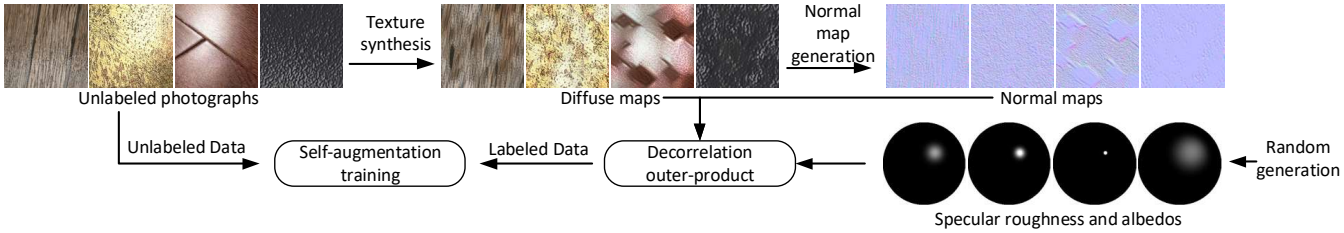

Figure 4: *Summary of our self-augmentation method with inexact supervision for surface reflectance estimation from a single image under unknown natural lighting.*

validation set. We see a similar result on the *smooth* validation set.

- *sharp-to-sharp* outperforms the *sharp* (without self-augmentation) on the *sharp* validation set. Again, a similar result holds for the *smooth* validation set.
- *regular-to-sharp* outperforms *regular* (without self-augmentation) on the *sharp* validation set. *regular-to-sharp* does not quite achieve the same accuracy as *sharp* or *sharp-to-sharp*. Note that the *sharp* target space is a subset of the *regular* target space and thus the ambiguities inherent to the *sharp* target space are resolved in the *regular* target space.
- *smooth-to-sharp* does not univocally perform better than only *smooth*. The *smooth* target space does not overlap the *sharp* target space, and it is therefore unlikely to completely address the ambiguities in the *sharp* target space. Again, we can draw similar conclusions for *sharp-to-smooth* on the *smooth* target space.

Conclusion Self-augmentation does not require that the labeled training data is drawn from the same distribution as the target distribution. However, it does require that the ambiguities inherent to the problem are clearly disambiguated. In most cases, self-augmentation will succeed from inexact labeled training data if the distribution of the labeled training data has significant overlap with the target distribution.

4. Self-augmentation with Inexact Supervision

Our experiments in Section 3 indicate that self-augmentation cannot succeed without labeled training data to resolve the ambiguities inherent to appearance estimation from a single image. Furthermore, our empirical study also indicates that the labeled data does not need to be drawn from the target distribution, as long as it addresses the ambiguities in the target distribution. In this section we will leverage this knowledge and design a pipeline to train a real-world single image surface appearance estimation network from unlabeled photographs only. While this goal seems to contradict our first observation regarding the need for labeled training data, we will exploit the second observation by synthesizing an inexact labeled training set directly from the unlabeled images without the need for manual labeling.

4.1. Labeled Data Synthesis

A key ambiguity in single image reflectance estimation is to correctly prorate the observed reflectance to the diffuse and specular reflectance. The labeled data, thus, provides guidance on how to decide which observed image structures are due to diffuse albedo variations and which are due to the specular reflections. Ideally, we would like to synthesize labeled training data that retains the diffuse albedo texture structures without specular "pollution". Removing the specular reflections directly from the unlabeled photographs is exactly what we would like the final network to do. To overcome this conundrum, we exploit the prior result that the labeled data does not need to be sampled from the same distribution as the target distribution. Hence, any set of labeled training data that comes from a similar, but different, distribution and which addresses the ambiguities suffices as a starting point for self-augmentation. Our solution is to synthesize the diffuse albedo maps, using neural texture synthesis, from the unlabeled photograph. The key idea is that the texture synthesis will retain the essence of the diffuse albedo structure, while destroying the undesired spatial structure of the specular reflections.

4.2. Practical Implementation

To synthesize the diffuse albedo maps, we largely follow the neural texture synthesis method of Gatys et al. [GEB15] with minor modifications. Instead of using 5 layers from the VGG feature map to compute the Gram matrix, we instead only use the 3 first layers (i.e., *conv1_1*, *pool1*, and *pool2*) to place more emphasis on the low level texture features. We synthesize slightly larger textures (10 additional pixels on all sides) and crop the center to avoid artifacts at the boundaries.

Normal maps and specular reflectance properties are synthesized following a similar process as in Section 3.1; we generate the normal map by converting the synthesized diffuse albedo maps to height maps, and randomly generate specular roughness and albedos. Since we have no knowledge on the presence or absence of correlations between the different reflectance components, we default to the least restrictive, and assume there is no correlation. We therefore, adopt the same decorrelation outer-product training strategy as in Section 3.2. Our labeled data generation process is summarized in Figure 4. Note how the synthesized textures shown in this pipeline overview differ significantly from the input texture; the specular structure is destroyed, while the essence of the diffuse texture is maintained.

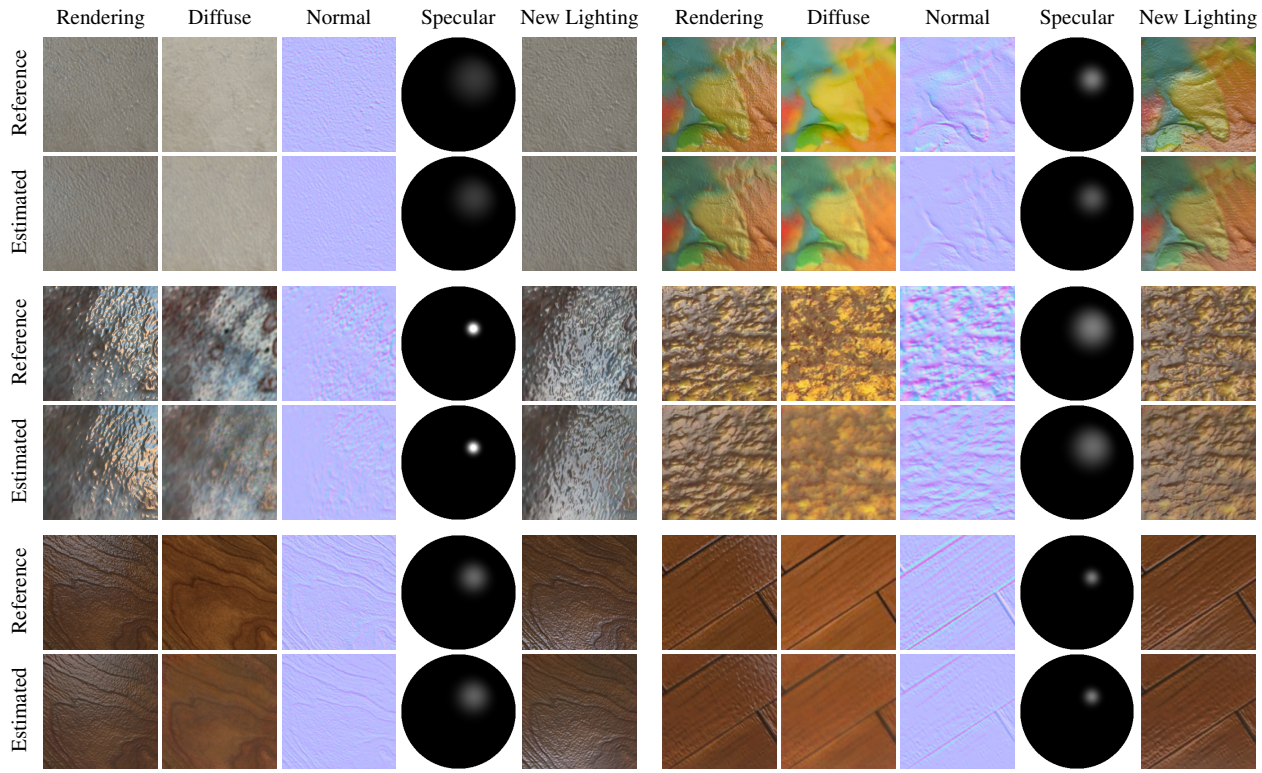

Figure 5: *Estimated reflectance components obtained with a network trained with self-augment and inexact supervision. The "rendering" column depicts the input image for the reference, and a re-rendered image under the same lighting of the recovered reflectance components. We also show a rendering under "new lighting".*

4.3. Results and Discussion

Training We collect an unlabeled dataset of 4827 images from the OpenSurfaces dataset [BUSB13] for three types of materials: wood, metal, and plastics. Unlike Li et al. [LDPT17], we train a network on all three types of materials simultaneously. From this unlabeled dataset, we generate a labeled dataset of also 4827 exemplars. With exception of the training data, we follow the same training settings and procedures as in Section 3.2.

Results We validate the quality of the surface reflectance estimated by our network on the set of artist modeled spatially varying surface reflectance for wood, metal, and plastics from Li et al. [LDPT17], and render an image for each under a randomly chosen environment map. Figure 5 compares recovered surface reflectance with reference reflectance components. As can be seen, the network trained from just unlabeled images is able to infer visually plausible diffuse albedo maps, normal maps, and homogeneous specular reflectance, and the re-rendered results exhibit the same visual qualities as the input image.

In addition, we also compare to a network trained using self-augmentation only (cf. [LDPT17]) with 60 labeled training exemplars and the same set of 4827 unlabeled images (Figure 6). Note: we employ our improved network structure for both estimation networks. Overall the network trained from unlabeled images produces better results compared to classic self-augmentation [LDPT17]. More specifically, the diffuse albedo map exhibits less artifacts (examples: $1,2,3,4,5,7,9$). The normal maps produced also appear sharper (examples: $1,2,3,6,7,8$). Finally, the predicted specular BRDF is also more accurate (examples: $1,5,6,7,8$). We suspect that the synthesized inexact "labeled" training data exhibits a larger variety. Even though the synthesized labeled training data does not exactly match the target distribution, it contains a denser sampling compared to sparse sampling of materials in the 60 labeled training exemplars [LDPT17]. A similar result can be observed in Table 3 which compares the average losses over 12 random crops from 25 artist labeled materials each rendered under 5 randomly selected lighting conditions; a total of 1,500 samples.

While Li et al. [LDPT17] suggest to train a separate network per material class, we found that this only provides minor quality improvements. As shown in Figure 6, our network trained with self-augmentation and inexact supervision outperforms both the general self-augmented network as well as a per-material class trained self-augmented network. The latter is trained according to the settings in [LDPT17] with a labeled dataset size of 40, 30, and 30 for *wood*, *metal*, and *plastic* respectively, and we use the subset from the 4827 image that matches the material class as the unlabeled training data.

Figure 7 shows results of our single image reflectance estimation network applied to selected photographs mined from online photo repositories captured under unknown conditions. In this case

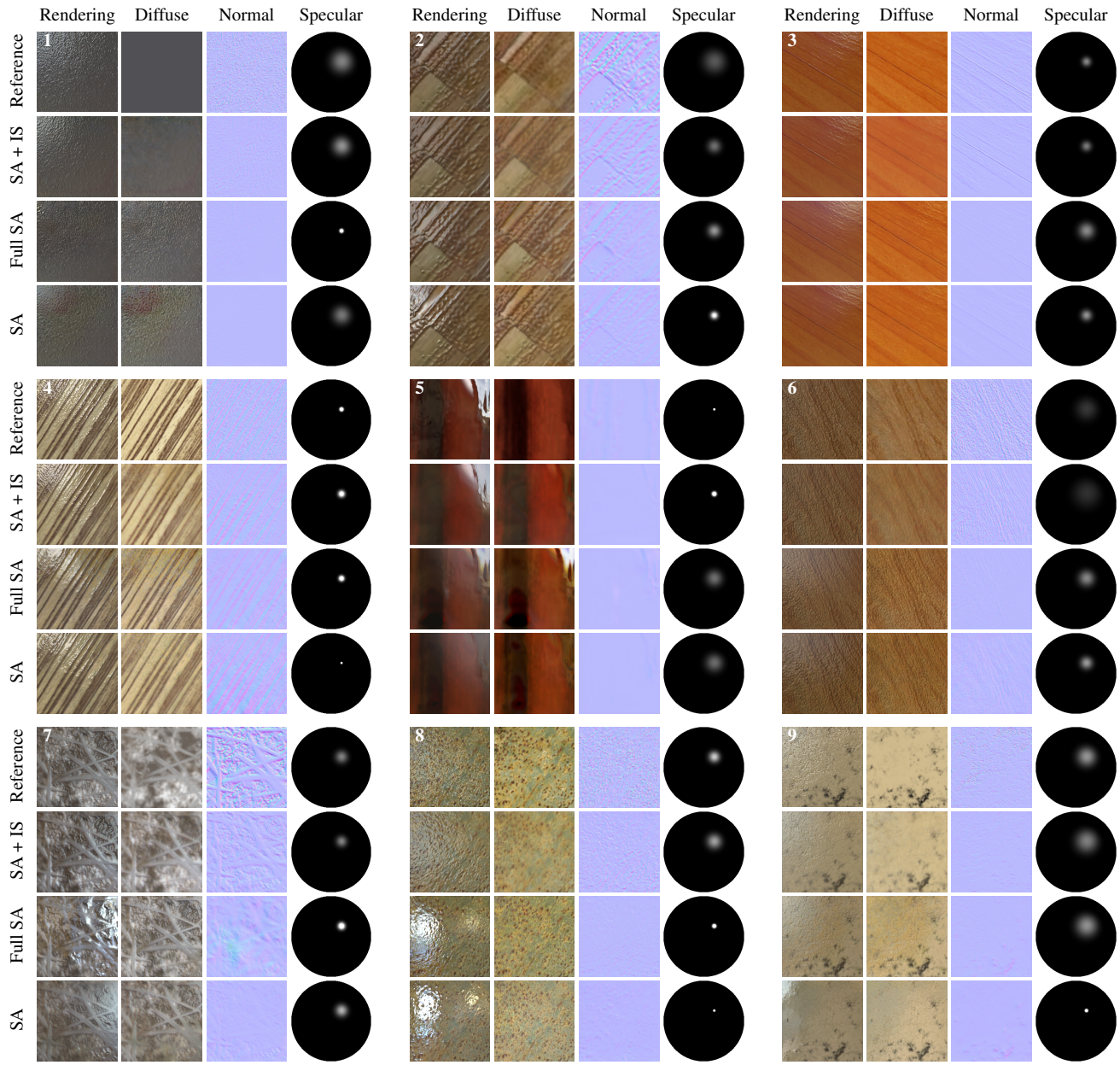

Figure 6: *Comparison of reflectance components estimated with our model trained with self-augmentation and inexact supervision (SA + IS), Li et al.s' [LDPT17] network trained with only self-augmentation on all materials classes (Full SA), and on Li et al.'s per material class trained self-augmentation network (SA). The "rendering" column shows the input image for the "reference" row, and the rendering of estimated components under the input lighting for other rows.*

Model	Diffuse Loss (e-2)	Normal Loss (e-2)	Specular Albedo Loss (e-1)	Specular Roughness Loss (e-1)	Rerendering input lighting (e-2)	Rerendering novel lighting (e-2)
Self-augmentation with Inexact Supervision	4.469	3.374	6.553	4.473	0.556	1.148
Self-augmentation [LDPT17]	4.625	3.403	6.001	5.715	0.882	1.193

Table 3: *Average loss over a validation set of* 1,500 *images (obtained from* 12 *random crops from a set of* 25 *reference materials rendered under* 5 *random lighting conditions) for each of the components and rendered images on networks trained with our self-augmentation with inexact supervision training strategy and Li et al.'s original self-augmentation strategy [LDPT17] jointly trained on all material classes.*

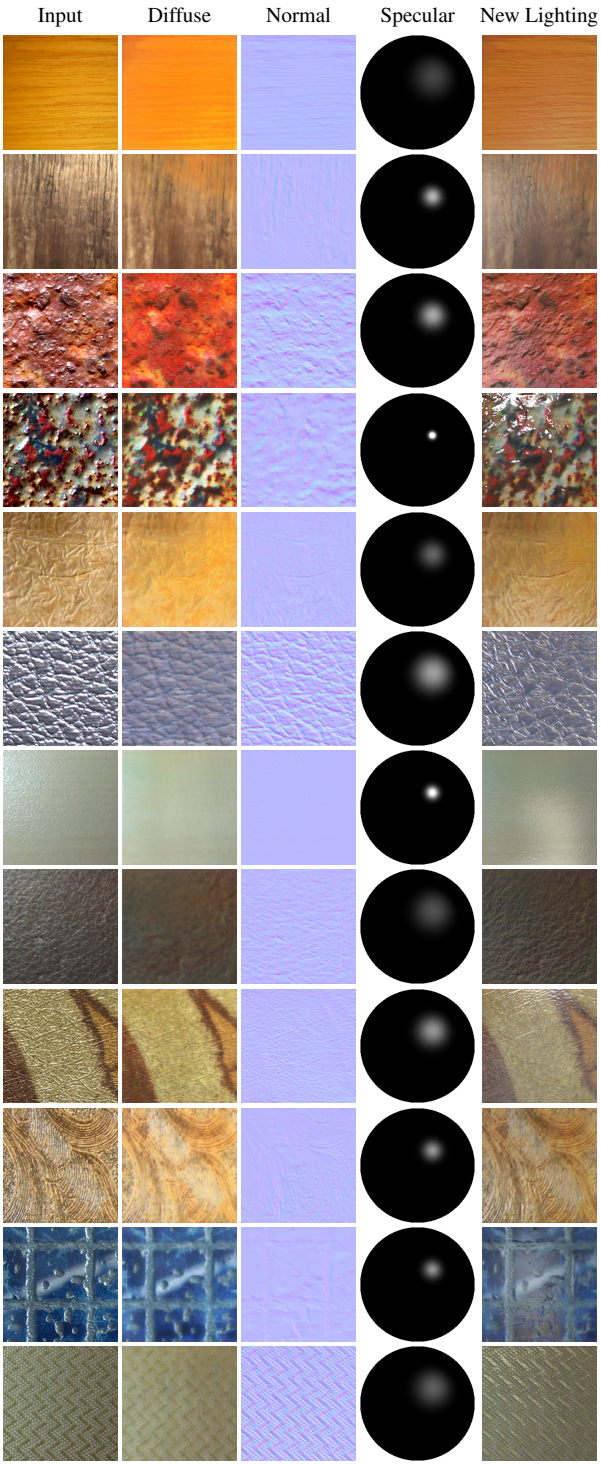

Figure 7: *Reflectance components estimated from photographs mined from (crops from) online photo repositories (rows 1 to 6; photographs courtesy of fireloopcreative, Chris Pond, Vincent Tcheng Chang, Barta IV, Jimmy Coupe, Mr Thinktank distributed under CC BY) and from captured photographs (rows 7 to 12).*

we do not have access to the ground truth reflectance components for comparisons. Nevertheless, visualizations under novel lighting conditions indicate that the recovered reflectance components are plausible when the photographs and materials match the conditions for which the network is trained such as a homogeneous specular layer, absence of shadows, normal view, and natural lighting conditions.

Discussion A key issue in training neural networks for reflectance estimation is to gather a sufficiently large training set. Self-augmentation attempts to resolve this issue by only using a small labeled dataset augmented by a large unlabeled dataset. Li et al. [LDPT17] only use tens of labeled dataset. While this greatly reduces the need for acquiring labeled data, it also affects the quality. First, self-augmentation does not offer any guarantees outside the space spanned by the labeled data. In higher dimensions, "spanned space" is difficult to quantify, and thus essentially one needs to stay close to the labeled data. From the perspective of resolving ambiguities, the "spanned space" refers to the space in which the labeled exemplar provides the necessary guidance to resolve the ambiguities. Second, during self-augmentation, half of the training batches consists of labeled data. Hence, the model is biased towards the relatively small space "spanned" by the limited amount of labeled data.

The proposed method relies on synthesis to overcome the difficulty of gathering labeled data. The property that the initial approximative network can be trained from a different distribution allows us to generate a large collection of "labeled" training exemplars. However, the synthesis process needs to be carefully designed and embed prior knowledge in how to avoid the inherent domain specific ambiguities. We have designed explicit strategies for this purpose:

1. We use neural texture synthesis to break the specular reflection patterns (as opposed to the much harder problem of completely removing the patterns).
2. We have included a decorrelation outer-product training strategy to avoid baking in the correlations between the diffuse albedo maps and normals (as a consequence of the fact that normal maps are generated from the diffuse albedo maps).
3. We exploit robustness with respect to the exactness of the labeled data, and use a labeled training set with a different distribution as the unlabeled data.

A key contribution of our empirical analysis is the observation that the labeled training dataset does not need to be sampled from the exact same distribution as the target space. This allows us to use a straightforward texture synthesis technique for generating the labeled training data. Note, however, that such a synthesized texture dataset is unsuited for directly training with labeled data only. While it is possible to manually craft and fine-tune a procedural texture generator, it will require significantly more user input to ensure that sufficient properties suitable for training of real-world textures are maintained.

While our method greatly eases training of convolutional neural networks for reflectance estimation from a single photograph, it is not without limitations. We assume that the lighting is distant, and our method will fail for photographs illuminated by strong local lights or when the whole image is covered by a highlight. Our

method assumes a homogeneous specular component, and will fail in the presence of strong spatial variations. Our network is trained for normal view directions and without taking in account shadows from self-occlusion; deviations from these conditions will adversely affect the plausibility of the results. We also assume that the normal maps can be defined by a height map. This excludes non-integrable surface normals (e.g., when the surface features are smaller than a pixel). Finally, we assume that the images are acquired with a fixed field of view (i.e., 60 degrees). While the method is robust to some variation, when the field of view deviates significantly, so will the structures in the reflections, and potentially the network will be unable to disambiguate diffuse albedo texture and reflections.

5. Conclusion

We presented an empirical study on the properties of Li et al.'s [LDPT17] self-augmentation training strategy in the context of spatially varying reflectance estimation from a single image under unknown natural lighting. Our experiments lead us to conclude that (1) self-augmentation requires an initial approximative network that informs the network on how to resolve the inherent ambiguities of reflectance estimations from a single image, and (2) the labeled data required to train this initial network does not need to be sampled from the same distribution as the target distribution, as long as it provides guidance on how to resolve the ambiguities present in the target space.

Based on these two conclusions, we design a novel self-augmentation training strategy for spatially varying reflectance estimation that only requires unlabeled training data. The required labeled data for training the initial network is synthesized directly from the unlabeled training data using a neural texture synthesis method.

For future work we would like to investigate the properties and constraints on the unlabeled training data, and design automated gathering strategies that maximize the accuracy of the method. Finally, our self-augmentation training strategy with inexact supervision is specifically geared towards reflectance modeling. Our synthesis strategy incorporates domain knowledge to obtain complete labeled training data. An interesting avenue for future research would be to generalize this strategy to other problem domains.

Acknowledgments We would like to thank the reviewers for their constructive feedback. Pieter Peers was partially supported by NSF grant IIS-1350323 and gifts from Google, Activision, and Nvidia.

References

[AAL16] AITTALA M., AILA T., LEHTINEN J.: Reflectance modeling by neural texture synthesis. *ACM Trans. Graph. 35*, 4 (July 2016), 65:1–65:13. 1, 2

[AWL15] AITTALA M., WEYRICH T., LEHTINEN J.: Two-shot svbrdf capture for stationary materials. *ACM Trans. Graph. 34*, 4 (July 2015), 110:1–110:13. 2

[BM15] BARRON J. T., MALIK J.: Shape, illumination, and reflectance from shading. *PAMI 37*, 8 (Aug. 2015), 1670–1687. 2

[BUSB13] BELL S., UPCHURCH P., SNAVELY N., BALA K.: OpenSurfaces: a richly annotated catalog of surface appearance. *ACM Trans. Graph. 32*, 4 (July 2013), 111:1–111:17. 8

[DAD*18] DESCHAINTRE V., AITTALA M., DURAND F., DRETTAKIS G., BOUSSEAU A.: Single-image svbrdf capture with a rendering-aware deep network. *ACM Trans. Graph. 37*, 128 (Aug. 2018), 15. 1, 2

[DRS08] DORSEY J., RUSHMEIER H., SILLION F.: *Digital Modeling of Material Appearance*. Morgan Kaufmann Publishers Inc., 2008. 2

[GEB15] GATYS L., ECKER A. S., BETHGE M.: Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems* (2015), pp. 262–270. 7

[GRR*17] GEORGOULIS S., REMATAS K., RITSCHEL T., GAVVES E., FRITZ M., GOOL L. V., TUYTELAARS T.: Reflectance and natural illumination from single-material specular objects using deep learning. *PAMI* (2017). 2

[HSL*17] HUI Z., SUNKAVALLI K., LEE J.-Y., HADAP S., WANG J., C. SANKARANARAYANAN A.: Reflectance capture using univariate sampling of brdfs. In *ICCV* (10 2017), pp. 5372–5380. 2

[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), 45:1–45:11. 1, 2, 3, 4, 8, 9, 10, 11

[LN12] LOMBARDI S., NISHINO K.: Reflectance and natural illumination from a single image. In *ECCV* (2012), pp. 582–595. 2

[LN16] LOMBARDI S., NISHINO K.: Reflectance and illumination recovery in the wild. *PAMI 38*, 1 (Jan. 2016), 129–141. 2

[LSC18] LI Z., SUNKAVALLI K., CHANDRAKER M. K.: Materials for masses: Svbrdf acquisition with a single mobile phone image. 2

[ON12] OXHOLM G., NISHINO K.: Shape and reflectance from natural illumination. In *ECCV* (2012), pp. 528–541. 2

[ON16] OXHOLM G., NISHINO K.: Shape and reflectance estimation in the wild. *PAMI 38*, 2 (Feb. 2016), 376–389. 2

[Per02] PERLIN K.: Improving noise. In *ACM Transactions on Graphics (TOG)* (2002), vol. 21, ACM, pp. 681–682. 3

[RRF*16] REMATAS K., RITSCHEL T., FRITZ M., GAVVES E., TUYTELAARS T.: Deep reflectance maps. In *CVPR* (2016), pp. 4508–4516. 2

[RZ10] ROMEIRO F., ZICKLER T.: Blind reflectometry. In *ECCV* (2010), pp. 45–58. 2

[War92] WARD G. J.: Measuring and modeling anisotropic reflection. *SIGGRAPH Comput. Graph. 26*, 2 (1992), 265–272. 3

[WK15] WEINMANN M., KLEIN R.: Advances in geometry and reflectance acquisition. In *ACM SIGGRAPH Asia, Course Notes* (2015). 2

[WSM11] WANG C.-P., SNAVELY N., MARSCHNER S.: Estimating dual-scale properties of glossy surfaces from step-edge lighting. *ACM Trans. Graph. 30*, 6 (2011), 172:1–172:12. 2

[XNY*16] XU Z., NIELSEN J. B., YU J., JENSEN H. W., RAMAMOORTHI R.: Minimal brdf sampling for two-shot near-field reflectance acquisition. *ACM Trans. Graph. 35*, 6 (Nov. 2016), 188:1–188:12. 2