Highlight-Aware Two-Stream Network for Single-Image SVBRDF Acquisition

JIE GUO, State Key Lab for Novel Software Technology, Nanjing University
SHUICHANG LAI, State Key Lab for Novel Software Technology, Nanjing University
CHENGZHI TAO, State Key Lab for Novel Software Technology, Nanjing University
YUELONG CAI, State Key Lab for Novel Software Technology, Nanjing University
LEI WANG, Guangdong OPPO Mobile Telecommunications Corp Ltd
YANWEN GUO*, State Key Lab for Novel Software Technology, Nanjing University
LING-QI YAN, University of California, Santa Barbara

Fig. 1. The proposed deep learning method is able to generate disentangled SVBRDF maps from a single, casually captured image. Thanks to a new highlight-aware convolution operation and a well-designed two-stream network, our method succeeds in recovering rich and detailed reflectance variation from the input image, and significantly outperforms the state-of-the-art solution [Deschaintre et al. 2018] that may be plagued with specular highlights. (Please use Adobe Acrobat and click the renderings to see the animation.)

This paper addresses the task of estimating spatially-varying reflectance (i.e., SVBRDF) from a single, casually captured image. Central to our method is a highlight-aware (HA) convolution operation and a two-stream neural network equipped with proper training losses. Our HA convolution, as a novel variant of standard (ST) convolution, directly modulates convolution kernels under the guidance of automatically learned masks representing potentially overexposed highlight regions. It helps to reduce the impact of strong specular highlights on diffuse components and at the same time, hallucinates plausible contents in saturated regions. Considering that variation of saturated pixels also contains important cues for inferring surface bumpiness and specular components, we design a two-stream network to extract features from two different branches stacked by HA convolutions and ST convolutions, respectively. These two groups of features are further fused in an attention-based manner to facilitate feature selection of each SVBRDF map. The whole network is trained end to end with a new perceptual adversarial loss which is particularly useful for enhancing the texture details. Such a design also allows the recovered material maps to be disentangled. We demonstrate through quantitative analysis and qualitative visualization that the proposed method is effective to recover clear SVBRDFs from a single casually captured image, and performs favorably against state-of-the-arts. Since we impose very few constraints on the capture process, even a non-expert user can create high-quality SVBRDFs that cater to many graphical applications.

CCS Concepts: • **Computing methodologies** → **Reflectance modeling**; *Neural networks*.

Additional Key Words and Phrases: Reflectance Modeling, SVBRDF, Deep Learning, Rendering

ACM Reference Format:
Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. 2021. Highlight-Aware Two-Stream Network for Single-Image SVBRDF Acquisition. *ACM Trans. Graph.* 40, 4, Article 123 (August 2021), 14 pages. https://doi.org/10.1145/3450626.3459854

*Corresponding author

Authors' addresses: Jie Guo, State Key Lab for Novel Software Technology, Nanjing University, guojie@nju.edu.cn; Shuichang Lai, State Key Lab for Novel Software Technology, Nanjing University, sclai@smail.nju.edu.cn; Chengzhi Tao, State Key Lab for Novel Software Technology, Nanjing University, tcz_tao@yeah.net; Yuelong Cai, State Key Lab for Novel Software Technology, Nanjing University, 171860689cyl@gmail.com; Lei Wang, Guangdong OPPO Mobile Telecommunications Corp Ltd, wanglei12@oppo.com; Yanwen Guo, State Key Lab for Novel Software Technology, Nanjing University, ywguo@nju.edu.cn; Ling-Qi Yan, University of California, Santa Barbara, lingqi@cs.ucsb.edu.

0730-0301/2021/8-ART123 $15.00
https://doi.org/10.1145/3450626.3459854

1 INTRODUCTION

Reconstructing reflectance properties of real-world materials from 2D images has been a long-standing problem in computer graphics and vision [Dong 2019; Guarnera et al. 2016; Weyrich et al. 2008], with applications ranging from product design and visual effects to virtual/mixed reality and cultural heritage. For opaque materials, the surface reflectance properties can be modeled by the 6D spatially-varying bidirectional reflectance distribution function (SVBRDF) [Nicodemus et al. 1977] of space and angles. Due to the high dimensionality of appearance data and the inherent ambiguity involved in the problem, reconstructing high-quality SVBRDFs favors multiple input images captured under different lighting and view directions. For instance, many early methods use expensive hardware setups to exhaustively sample the 6D space of SVBRDFs. This allows to capture high-frequency details and resolve the illumination-reflectance ambiguity [Dana et al. 1999; Holroyd et al. 2010; Lawrence et al. 2006], but leads to a tedious and arduous process that often limits the applicability.

To simplify the acquisition process, much effort has been devoted to reducing the number of input images to a minimum. This can be accomplished by exploring domain-specific priors as extra constraints, such as a sparse mixture of basis materials [Hui et al. 2017; Ren et al. 2011; Wu et al. 2016; Zhou et al. 2016] and stochastic texture-like behavior [Aittala et al. 2015]. However, strong prior knowledge usually restricts the family of materials that can be captured. Recent work try to estimate general SVBRDFs from a latent embedded space characterized by different deep neural networks [Gao et al. 2019; Guo et al. 2020]. These learned material priors significantly improve hand-crafted ones, but cumbersome optimization is required to support an arbitrary number of input images. Typically, multi-view methods require correct calibration of multiple cameras, which is often well beyond the capabilities of novice users and adds fragility of these methods.

For many appearance modeling scenarios, it is more convenient and appealing to reconstruct an SVBRDF from a single color image, neglecting any hardware assistance or user intervention. However, this problem is fundamentally underconstrained as many different combinations of material parameters can reach the same radiance observed in the image, resulting in a rather challenging task. Pioneered by the work of Li et al. [2017], deep learning has provided new opportunities for single-image SVBRDF acquisition [Deschaintre et al. 2018; Li et al. 2018a; Ye et al. 2018].

In this paper, we endeavor to develop a novel single-image SVBRDF reconstruction framework using deep neural networks. The input low dynamic range (LDR) image can be taken from a hand-held consumer-level camera with a built-in flash. Current state-of-the-art methods [Deschaintre et al. 2018; Li et al. 2018a] have concentrated on applying standard convolutions to learn important features from the image. However, we observe that standard convolutions will struggle to cope with saturated pixels stemming from specular highlights. Since all pixels are treated equally, overexposed highlight regions will provide ambiguous information that misleads the networks, incurring uncomfortable blotchy artifacts. To address this problem, we propose *highlight-aware (HA) convolution* which is designed to extract meaningful features from unsaturated contents, suppressing the influence of specular highlight pollution. Our HA convolution automatically learns a soft mask representing the potential overexposed regions from the input image, and uses it to weaken invalid features produced by saturated pixels.

Considering that the highlight corrupted regions could also provide useful cues for inferring specular components of SVBRDFs, we design a *two-stream network* to take full advantage of the input image. Our network contains two separate feature extracting streams: a HA convolution stream constructed to explicitly handle improperly exposed regions and to extract highlight-free features from multiple scales, while a standard convolution stream responsible for extracting additional knowledge. An attention-based fusion module is employed to selectively combine features from different streams. To train the network, we also design a generative adversarial loss [Goodfellow et al. 2014] to improve the predicted results with high-frequency details and consistent contents, in addition to the pixel-wise losses. With this two-stream architecture and the carefully designed losses, we are able to recover disentangled material maps even from a single image, eliminating potential specular pollution and blurry textures.

Overall, our contributions in this paper include:

- A new convolution variant, i.e., HA convolution, that significantly weakens the influence of saturated pixels in feature extraction.
- A well-designed two-stream network to fully exploit useful features from any casually captured image, facilitating disentangled learning of material properties.
- An attention-based feature selection (AFS) module to combine features from two different streams, improving the quality of reconstructed material maps.
- A novel training loss function incorporating a two-scale perceptual adversarial loss to preserve sharp edges and consistent contents.

We demonstrate the accuracy and flexibility of our method on both synthetic and real data. Comprehensive experiments show that our work significantly advances the state-of-the-art of SVBRDF acquisition from a single image.

2 RELATED WORK

Existing work on SVBRDF acquisition is generally grouped into two high-level categories according to the number of images used. Below we review some of the previous work and also briefly discuss image inpainting which is closely related to our work.

2.1 Multi-Image Surface Reflectance Acquisition

Due to the inherent complexity of real-world material appearances [Dong 2019; Guarnera et al. 2016; Weyrich et al. 2008], reconstructing a high-quality SVBRDF initially requires exhaustive spatial and angular sampling of each material, resulting in hundreds of images. Conventionally, these images are collected from specialized hardware systems that consist of multiple lights or cameras [Aittala et al. 2013; Asselin et al. 2020; Baek et al. 2018; Dana et al. 1999; Ghosh et al. 2010, 2008; Goldman et al. 2010; Holroyd et al. 2010; Kang et al. 2018, 2019; Lawrence et al. 2006; Nam et al. 2016; Tunwattanapong et al. 2013; Wang et al. 2008; Yu et al. 2016]. Despite the accuracy in

reconstruction, most previous devices are prohibitively expensive and the acquisition process is time-consuming. To reduce the acquisition cost, it is attractive to use off-the-shelf commodity mobile devices [Aittala et al. 2015; Nam et al. 2018; Riviere et al. 2016, 2017] or low-cost RGBD cameras [Ha et al. 2020; Hui et al. 2017; Wu et al. 2016]. To make the problem more tractable, domain-specific priors or assumptions are commonly imposed on the solution, such as a low-dimensional manifold in high-dimensional BRDF space [Dong et al. 2010], a linear combination over a limited number of basis materials [Hui et al. 2017; Ren et al. 2011; Wu et al. 2016; Zhou et al. 2016], repetitive or stochastic texture-like behavior [Aittala et al. 2015] and sparse incident lighting [Dong et al. 2014]. However, these assumptions usually limit the range of reconstructed SVBRDFs.

Many recent methods solve this problem by training deep neural networks using synthetic data for supervision. In the context of multiple input images, Deschaintre et al. [2019] introduced an order-independent pooling layer to fuse multiple feature maps produced by the single-image networks. This helps to combine appearance cues scattered across different inputs and improve their previous solution based on a single image [Deschaintre et al. 2018]. Gao et al. [2019] proposed a unified inverse rendering-based approach for reconstructing SVBRDFs from an arbitrary number of inputs. They trained an auto-encoder [Hinton and Salakhutdinov 2006] to extract a latent space for SVBRDFs and then optimized material maps in this space. The learned latent space serves as a material prior and ensures the plausibility of the reconstructed SVBRDFs. However, due to the fully convolutional natural, this method fails to capture global patterns in the material, and hence relies on previous methods for good initialization. In light of this, Guo et al.[2020] proposed MaterialGAN, a StyleGAN2-based network [Karras et al. 2019], to learn global correlation in material parameters. They show that optimization in such a globally robust latent space yields higher quality reconstruction. Currently, these methods are tailored for planar exemplar. To properly handle non-planar objects, Boss et al.[2020] designed a cascaded network for the estimation of SVBRDF, illumination and shape from two-shot images.

2.2 Single-Image Surface Reflectance Acquisition

SVBRDF acquisition methods (even deep learning-based methods [Gao et al. 2019; Guo et al. 2020]) relying on multiple input images hinder adoption by non-expert users since they require correct camera calibration and registration. For many appearance modeling scenarios, it is more convenient and accessible to use only one input image. However, without user intervention [Dong et al. 2011; Lin et al. 2019], reconstructing surface reflectance from a single image becomes a highly ill-posed problem. Currently, most successful methods resort to deep learning.

An early attempt leverages a neural Gram-matrix texture descriptor extracted from a CNN to estimate reflectance properties of a planar surface [Aittala et al. 2016]. As precise point-to-point correspondences are neglected, this method only works for stationary textured materials. Li et al. [2017] presented a learning-based solution to tackle with general textured materials. They introduced a new training strategy named self-augmentation to overcome the difficulty of insufficient labeled training data. This method is further extended to use just unlabeled data for training [Ye et al. 2018]. Considering the impact of appropriate training data on the accuracy of reconstructed SVBRDFs, Deschaintre et al. [2018] constructed a large dataset of artist-created, procedural SVBRDFs which cover a wide range of shading effects. With this dataset, they designed a deep neural network that combines a U-Net [Ronneberger et al. 2015] and a fully-connected global branch to extract both local and global features from a single image, achieving the state-of-the-art performance on single-image SVBRDF acquisition. Li et al. [2018a] also presented a synthetic SVBRDF dataset which is further manually classified into eight material types for better inference. Deschaintre et al. [2020] proposed a fine-tuning approach to capture SVBRDFs from large planar surfaces taken with ambient lighting. To remove the restriction of planar surfaces, Li et al. [2018b] proposed a cascaded CNN for recovering arbitrary shape and SVBRDF from a single mobile phone image lit by environment map and dominating collocated flashlight. This method is further extended to handle complex indoor scenes with a deep inverse rendering framework [Li et al. 2020]. A concurrent work [Zhou and Kalantari 2021] leveraged adversarial training and some real materials to improve the quality of reconstructed material properties. Since most methods do not explicitly handle improper exposure, they easily generate uncomfortable artifacts in highlight saturated regions. Our work addresses this problem by resorting to a learnable dynamic feature selection mechanism driven by specular highlights.

2.3 Image Inpainting

Recover missing contents from saturated pixels shares some similarity with image inpainting which aims at restoring missing regions of a corrupted image with plausible contents. Some early patch-based approaches attempt to fill missing regions by propagating uncorrupted contents to the holes [Barnes et al. 2009; Huang et al. 2014; Xu and Sun 2010]. Over the last few years, deep neural networks have advanced image inpainting by hallucinating missing pixels in a data-driven manner. One line of work seeks to leverage two stages to separately predict missing structures and textures in a step-by-step manner [Nazeri et al. 2019; Ren et al. 2019; Xiong et al. 2019; Yu et al. 2018]. Liu et al. [2020] recently pointed out that these two-stage methods may generate inconsistencies in appearances and suggested to use a mutual encoder-decoder to jointly learn features representing structures and textures. An alternative design is to construct networks with convolution variants to better handle irregular holes [Liu et al. 2018, 2020; Yi et al. 2020; Yu et al. 2019]. To avoid visual artifacts caused by standard convolutions, Liu et al. [2018] introduced the partial convolution and mask-update operation to force the network to use only valid pixels. Following this idea, Yu et al. [2019] proposed the gated convolution which provides a dynamic feature gating mechanism for each location across all layers, achieving better performance with free-form masks. Our proposed HA convolution is a new variant of the standard convolution that specially considers the impact of overexposed highlight regions. Unlike previous work, our HA convolution learns masks automatically from input images and performs proper normalization for faithful recovering of missing contents. Moreover, there have been studies showing that adversarial losses can enhance the realism of

inpainted contents [Iizuka et al. 2017; Liu et al. 2019; Pathak et al. 2016; Wang et al. 2020; Yi et al. 2020; Yu et al. 2018].

3 PROBLEM FORMULATION AND OVERVIEW

Our goal in this paper is to reconstruct a set of material properties from a single color image $\mathbf{I}$ taken from a nearly planar surface. This image is desired to be captured from roughly normal incidence to fully exploit important features on the surface. The surface is expected to be lit by an approximate point light source. Apart from these, our method has no other restrictions to the camera and the lighting. In particular, we do not need to know the internal and external camera intrinsics. We assume that the reflectance at any surface point is well represented by the Cook-Torrance BRDF model [1981] equipped with the GGX microfacet normal distribution function [Walter et al. 2007]. Under this circumstance, each SVBRDF comprises four parameters: diffuse albedo $\mathbf{k}_\mathrm{d}$, specular albedo $\mathbf{k}_\mathrm{s}$, specular roughness $\mathbf{r}$ and surface normal $\mathbf{n}$, corresponding to four material maps.

By designing and training a new CNN, we aim to learn a mapping function G that converts the input image $\mathbf{I}$ to its corresponding four material maps ($\mathbf{k}_\mathrm{d}$, $\mathbf{k}_\mathrm{s}$, $\mathbf{r}$ and $\mathbf{n}$). Due to the limited dynamic range of many consumer-level cameras, some image regions will be polluted by specular highlights, leading to saturated pixels. When trying to extract features from these improperly exposed images, standard (ST) convolutions will treat all input pixels, either saturated or unsaturated, as valid ones. However, saturated pixels contain less meaningful contents for diffuse components and will cause ambiguity during training as networks may regard them as additional features. This is demonstrated in the first row of Fig. 2. Improper feature maps tend to produce obvious blotchy artifacts which are easily observed in recent deep learning-based single-image SVBRDF recovery methods [Deschaintre et al. 2018].

To well solve the above problem and significantly improve the performance of SVBRDF acquisition from a single image, we introduce highlight-aware (HA) convolution to explicitly consider the impact of saturated pixels during feature extraction. The key idea is to automatically learn a series of hierarchical soft masks from an input image. These masks reflect potential influence of overexposed specular highlight regions at multiple scales. When incorporating these masks into CNNs and replacing the standard convolution with a highlight-aware variant, the networks will suppress the highlights and hallucinate plausible contents in overexposed regions, as shown in the second row of Fig. 2. The details of HA convolution will be exposed in Sec. 4.

Although saturated pixels pose challenges for recovering diffuse maps, they could provide important cues for inferring surface bumpiness and specular components [Chen et al. 2006; Wang et al. 2011]. In fact, bumpiness observed in an image is mostly due to specular reflection. This indicates that specular highlights and normal variation are highly correlated, as evidenced in Fig. 3. Even saturated regions could contain much information about the statistics of a surface. In this regard, we design a two-stream network comprising two different branches (HA-Branch and ST-Branch) to respectively extract highlight-aware (HA) features and standard (ST) features. Since the four material maps have unequal dependencies on these

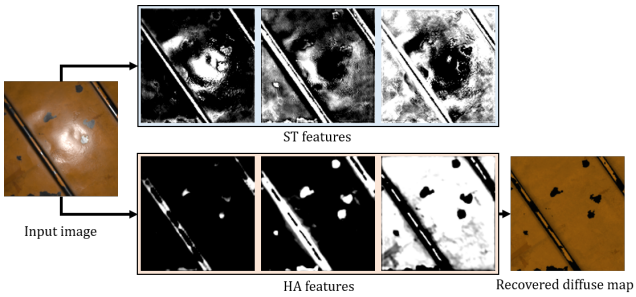

Fig. 2. Comparison between ST features and HA features. HA features extracted by stacked HA convolutions can well handle overexposed highlight regions and are beneficial for the recovery of diffuse maps.

Fig. 3. Strong correlations exist between specular highlights and normal variation. In each group, we show the input image (Input), ground-truth normal (GT) and our recovered normal (Recovered).

two groups of features, we employ addition FU-Branches to fuse features maps in an attention-based manner. Currently, we construct independent FU-Branch for each material map, and these four branches share the same architecture. Mathematically, we formulate our pipeline as

$$\{\mathbf{k}_\mathrm{d}, \mathbf{k}_\mathrm{s}, \mathbf{r}, \mathbf{n}\} = G(\mathbf{I}; \boldsymbol{\theta}) = \mathrm{FU}(\mathrm{HA}(\mathbf{I}), \mathrm{ST}(\mathbf{I})) \tag{1}$$

in which G represents our network parameterized by a set of weights $\boldsymbol{\theta}$. FU, HA and ST denote the corresponding branches, respectively. Sec. 5 dives into these steps in greater detail.

To train our network, we utilize the large-scale dataset provided by Deschaintre et al. [2018] which contains N example pairs of input LDR images and their corresponding ground-truth material maps, i.e., $\mathcal{D}_N = \{(\mathbf{I}^1, \mathbf{k}_\mathrm{d}^1, \mathbf{k}_\mathrm{s}^1, \mathbf{r}^1, \mathbf{n}^1), ..., (\mathbf{I}^N, \mathbf{k}_\mathrm{d}^N, \mathbf{k}_\mathrm{s}^N, \mathbf{r}^N, \mathbf{n}^N)\}$. The learning aims to find an optimal solution $\hat{\boldsymbol{\theta}}$ via a loss function $\mathcal{L}$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(G(\mathbf{I}^i; \boldsymbol{\theta}), \mathbf{k}_\mathrm{d}^i, \mathbf{k}_\mathrm{s}^i, \mathbf{r}^i, \mathbf{n}^i). \tag{2}$$

4 HIGHLIGHT-AWARE CONVOLUTION

In this section we provide details of our HA convolution, which is the fundamental building block of HA-Branch in our two-stream network. The proposed HA convolution is illustrated in Fig. 4 and described as follows. Given a feature map $\mathbf{X}^l \in \mathbb{R}^{H \times W \times C}$ at layer l where C is the number of channels, and H, W are respectively the height and width of the feature map, we first obtain its per-channel mean and standard deviation as

$$\mu_c^l = \frac{\sum_{h,w} \mathbf{X}_{c,h,w}^l}{HW}, \quad \sigma_c^l = \sqrt{\frac{\sum_{h,w} (\mathbf{X}_{c,h,w}^l - \mu_c^l)^2}{HW}}. \tag{3}$$

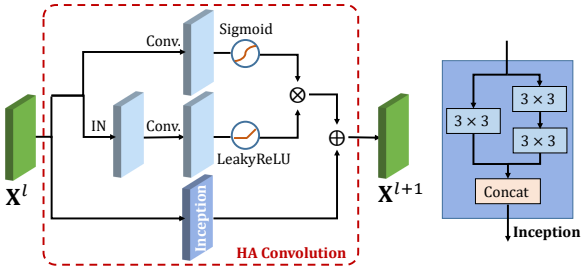

Fig. 4. Illustration of the proposed HA convolution and the inception block. Here, ⊕ and ⊗ denote element-wise addition and element-wise multiplication, respectively. IN denotes instance normalization.

Then, we normalize $\mathbf{X}^l$ with $\bar{\mathbf{X}}^l = (\mathbf{X}^l - \mu^l)/(\sigma^l + \epsilon)$ where ϵ is a small value to avoid division by zero. This is a typical instance normalization (IN) operation [Ulyanov et al. 2016] which makes the input features approach independent and identical distribution by a shared mean and variance. It is widely recognized that normalization improves performance and speeds up training. In our HA convolution, we also find that IN helps to remove shading from material maps.

Subsequently, the original feature $\mathbf{X}^l$ and its normalized counterpart $\bar{\mathbf{X}}^l$ are fed into two separate convolution layers, yielding

$$\mathbf{H}^l = \mathbf{X}^l * \mathbf{W}_h \quad (4)$$

$$\mathbf{F}^l = \bar{\mathbf{X}}^l * \mathbf{W}_f \quad (5)$$

where $\mathbf{W}_h$ and $\mathbf{W}_f$ are two trainable filters. The first convolution aims to identify potential overexposed highlight regions in $\mathbf{X}^l$. When activated by a sigmoid function $s(\cdot)$, we expect it (i.e., $s(\mathbf{H}^l)$) to provide free-form soft masks to reduce the contribution of features generated from the masked content. Fig. 5 visualizes automatically learned masks (averaged across channels) at three selected layers of the network. As seen, the masks tend to be blurrier and more uniform as we step deeper into the network. The second convolution can adopt any activation function $\phi(\cdot)$ (e.g., ReLU and LeakyReLU [Maas et al. 2013]) and is encouraged to extract features from valid contents with the help of the masks, i.e., $\phi(\mathbf{F}_l) \otimes s(\mathbf{H}_l)$.

Although IN stabilizes network training, it fails to maintain non-local information about the input image. To tackle this issue, previous work of Deschaintre et al. [2018] developed a global branch to inject global information about the input image back into the local track after every convolution. However, we observe that this strategy will cause over blurriness in our pipeline, since only global mean values are used. To preserve both global information and local details, we augment HA convolution with an inception block as shown in Fig. 6. This inception block has two parallel tracks. One track has a simple 3×3 convolution, and the other has two consecutive 3×3 convolutions which are similar to a 5×5 convolution but have a lower computational cost. The channel of the feature map is halved along each track and then restored to the origin by concatenation. Let p be the mapping learned by the inception block, we formulate our HA convolution as

$$\mathbf{X}^{l+1} = \phi(\mathbf{F}^l) \otimes s(\mathbf{H}^l) \oplus p(\mathbf{X}^l) \quad (6)$$

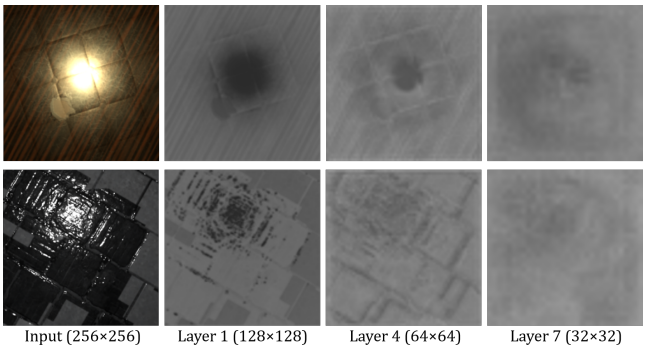

Fig. 5. Evolution of learned masks of two input images in the first column. The right three columns show the average masks (across channels) of selected layers. As we step deeper through the network, the masks from HA convolutions become blurrier and more uniform.

where $\mathbf{X}^{l+1}$ is the output feature map.

4.1 Discussion

Our HA convolution bears some similarity with gated convolution [Yu et al. 2019], a popular convolution variant that has proven fruitful in image inpainting. However, two key factors distinguishing our HA convolution from gated convolution: the automatically learned masks and the inception block. Unlike HA convolution, gated convolution requires a predetermined mask as input. However, in our application, such a mask is not easy to obtain. Instead, we let the network learn the masks automatically. Some examples are provided in Fig. 5. If we replace all HA convolutions in our network with gated convolution, it will fail to recover missing contents in saturated pixels, leading to objectionable artifacts as shown in the first row of Fig. 6. In the bottom row of Fig. 6, we also compare the evolution of the root-mean-square errors (RMSEs) of both material maps and renderings with respect to training epochs. We test on 60 SVBRDFs, all have obvious specular highlights, from the dataset of [Deschaintre et al. 2018]. The masks for gated convolutions are generated by clipping highlight regions. The renderings are conducted over 9 random view and lighting directions for each SVBRDF. As seen, the errors of both reconstructed material maps and novel-view renderings decrease as the training epochs increase. However, a network trained with gated convolution converges much slower than our network.

5 TWO-STREAM NETWORK FOR SVBRDF RECOVERY

Built upon HA convolutions, we design our two-stream network for single-image SVBRDF recovery. The overall architecture of our network contains an HA-Branch, an ST-Branch and four FU-Branches, which is illustrated in Fig. 7. The HA-Branch and ST-Branch perform the main task of feature learning, while four FU-Branches leverage these learned features to recover four material maps. Thanks to the fully convolutional neural network, images of arbitrary resolutions can be fed into the network as input.

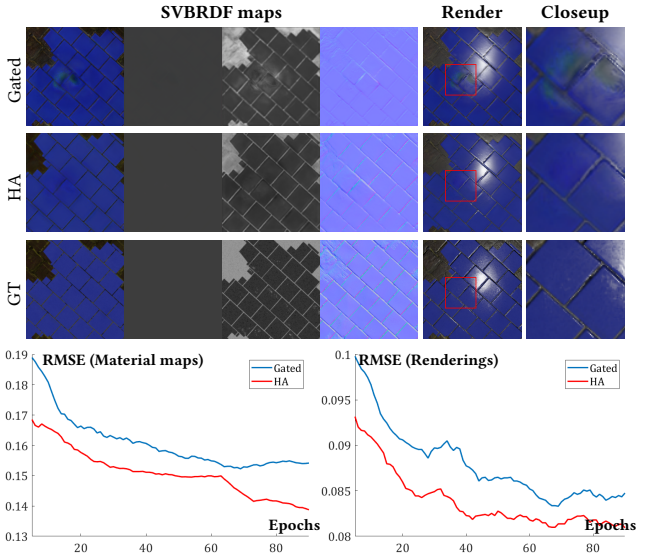

Fig. 6. Comparison between HA convolution and gated convolution [Yu et al. 2019] in SVBRDF recovery. The first three rows illustrate the visual effects of both material maps ($\mathbf{k}_d$, $\mathbf{k}_s$, $\mathbf{r}$ and $\mathbf{n}$) and novel-view renderings on a selected material. The last row shows average error plots (computed on 60 SVBRDs) of reconstructed material maps (left) and renderings (right) over the number of epochs.

5.1 Feature Extraction with HA-Branch and ST-Branch

By replacing ST convolution with HA convolution in traditional networks and training on proper datasets, we are able to hallucinate missing contents from saturated pixels. This is to be expected and validated by our experiments. However, we have observed that simply stacking ST convolutions tends to generate overly blurred normal ($\mathbf{n}$) and biased specular components ($\mathbf{k}_s$ and $\mathbf{r}$) since ST convolutions ignore gloss evident in the reflection off the samples if the reflected regions are overexposed. To remedy this situation, we design a two-stream network which contains two separated branches, namely HA-Branch and ST-Branch, for feature extraction.

Our network starts with a ST convolution which converts an input image $\mathbf{I}$ to a low-level feature map of the same resolution as $\mathbf{I}$. Then, this feature map is simultaneously fed into HA-Branch and ST-Branch for further processing. In our current design, HA-Branch comprises six HA convolutional blocks for feature map inpainting and downsampling, followed by a dilated HA convolution block and two upsampling blocks. Each block contains a LeakyReLU activation. Downsampling is implemented by using a stride of 2 in the convolutions while upsampling is implemented by bilinear interpolation and ST convolution. In the seventh layer of HA-Branch, dilated HA convolution with a factor 2 is used to increase the size of the receptive field. This has an equal effect on the receptive field with respect to the input features, but allows the network to keep additional information in the bottleneck. We do not use HA convolutions in the upsampling stage of HA-Branch because feature maps have already been repaired after passing the bottleneck.

Two skip connections [Ronneberger et al. 2015] are introduced between those same-sized layers of the encoder and decoder to help the decoder retain details as much as possible. We only add skip connections to the high-level layers since low-level layers at the encoder side may still contain saturated pixels that pollute subsequent features.

For ST-Branch, it has the same architecture with HA-Branch but replaces all HA convolutions with ST convolutions. Similarly, each convolutional block contains a LeakyReLU activation. In Fig. 9, we show that this branch is beneficial for retaining details from the input image. Without this branch (the first row of Fig. 9), the recovered material maps, in particular the normal map, become quite blurry. This results in large differences between novel-view renderings of the output material maps and the ground-truth maps, as highlighted in the closeups.

5.2 Feature Fusion and Selection with FU-Branch

The extracted features from HA-Branch and ST-Branch devote unequally to recover different maps of an SVBRDF. Hence, we propose an attention-based mechanism for feature selection, i.e., AFS. Attention mechanisms have been widely adopted in recent CNNs which bias the allocation of the most informative feature expressions and simultaneously suppress less useful ones. In our pipeline, we design FU-Branch to self-recalibrate the feature map via channel-wise importances.

Let $\mathbf{X}^{\mathrm{HA}} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{X}^{\mathrm{ST}} \in \mathbb{R}^{H\times W\times C}$ be the two groups of features extracted by the final convolutional blocks of HA-Branch and ST-Branch, respectively, we first fuse these features via concatenation: $\mathbf{X}^{\mathrm{HA}}©\mathbf{X}^{\mathrm{ST}}$ with © being the concatenation operation. Then, we embed the global information by using global average pooling (GAP) to generate channel-wise statistics as $\mathbf{d} \in \mathbb{R}^{1\times 1\times 2C}$. Specifically, $\mathbf{d}$ is computed as

$$\mathbf{d} = \frac{\sum_{h,w}(\mathbf{X}^{\mathrm{HA}}©\mathbf{X}^{\mathrm{ST}})_{h,w}}{HW}. \tag{7}$$

After passing through a multi-layer perceptron (MLP) with a single hidden layer and a sigmoid activation, $\mathbf{d}$ is further mapped to a soft attention vector $\mathbf{a}$. The final output of this module $\tilde{\mathbf{X}}$ is obtained by rescaling the input feature map with the soft attention vector $\mathbf{a} \in \mathbb{R}^{1\times 1\times 2C}$ with channel-wise multiplication, i.e.,

$$\tilde{\mathbf{X}}_c = \mathbf{a}_c \cdot (\mathbf{X}^{\mathrm{HA}}©\mathbf{X}^{\mathrm{ST}})_c \tag{8}$$

where c represents the c-th channel. This channel-wise attention is expected to select HA features and ST features adapted to different material maps. In Fig. 10, we validate that AFS plays an important role in our two-stream network. Without it, our network may fail to recover missing contents in saturated regions since the FU-Branch treats $\mathbf{X}^{\mathrm{HA}}$ and $\mathbf{X}^{\mathrm{ST}}$ equally. Moreover, the generated textures are blurry mostly due to the mutual interference between different features.

Another important design choice for FU-Branch is that we adopt four separate branches to recover per pixel diffuse albedo, surface normal, specular roughness and specular albedo, respectively. This means the learned attention vector $\mathbf{a}$ and the subsequent convolutional kernels in FU-Branch vary across material maps. This allows each material map to select features benefiting itself and facilitates

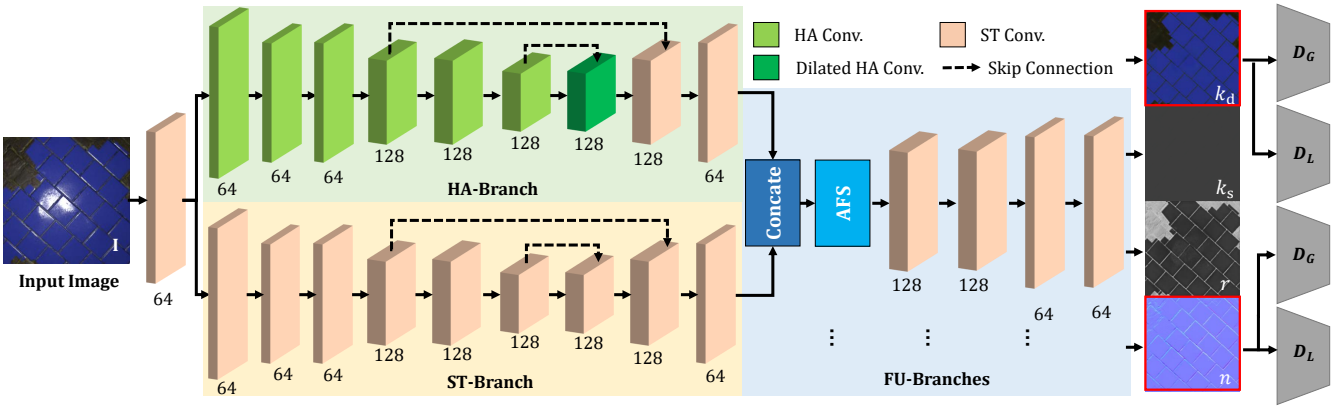

Fig. 7. The architecture of our proposed two-stream network. It consists of two separate branches (HA-Branch and ST-Branch) for feature extraction and four feature fusion branches (FU-Branches) for final prediction. D_G and D_L represent global context discriminator and local context discriminator, respectively. The channel counts are shown below the convolutional layers. Note that four FU-Branches (one for each material map) have the same architecture and only one is shown here as an example.

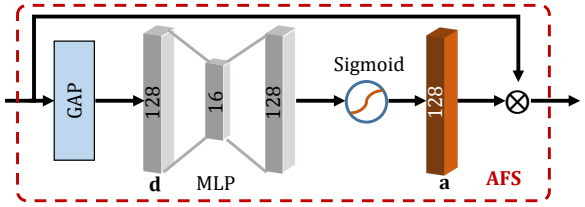

Fig. 8. Attention-based feature selection (AFS) module. Here, GAP refers to global average pooling, and MLP refers to multi-layer perceptron.

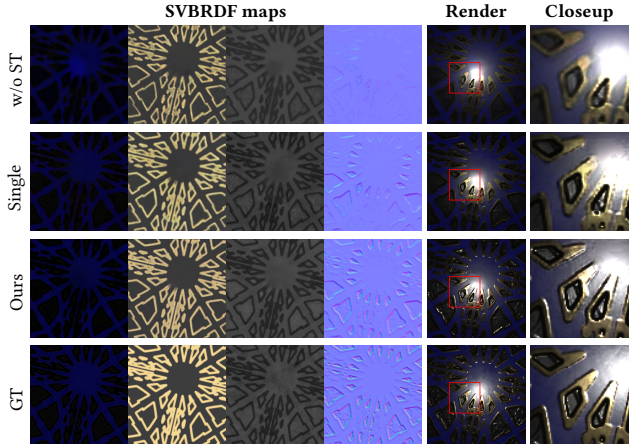

Fig. 9. Influence of branches. We compare our model with two variants: one without ST-Branch (w/o ST) and the other with a single decoder (Single).

disentangled learning of material properties. As compared in Fig. 9 (the second row vs. the third row), network trained with a single decoder achieves sub-optimal solutions in SVBRDF recovery.

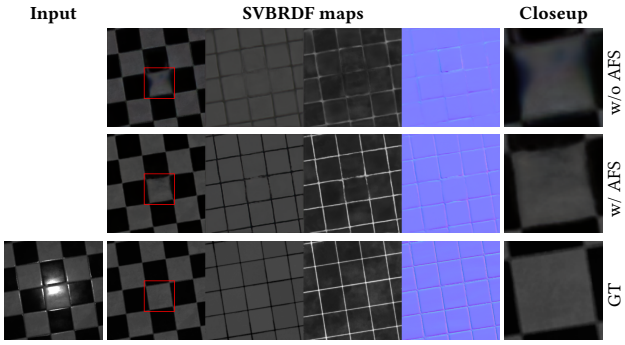

Fig. 10. Influence of AFS. Without AFS (the first row), our network may fail to recover missing contents in saturated regions and tends to produce blurry textures.

5.3 Loss Function

The proposed network is trained over a joint loss function that consists of a map loss $\mathcal{L}_{\text{map}}$ computed on the reconstructed material maps with l_1 norm, a rendering loss $\mathcal{L}_{\text{render}}$ computed on 9 novel-view renderings with l_1 norm and an adversarial loss $\mathcal{L}_{\text{adv}}$. Specifically,

$$\mathcal{L} = \lambda_{\text{map}}\mathcal{L}_{\text{map}} + \lambda_{\text{render}}\mathcal{L}_{\text{render}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \tag{9}$$

where λ_{map}, λ_{render} and λ_{adv} are weights to balance the influence of each term. For our experiments, we choose $\lambda_{\text{map}} = 100$, $\lambda_{\text{render}} = 10$ and $\lambda_{\text{adv}} = 1$. Previous networks mostly leverage pixel-wise losses ($\mathcal{L}_{\text{map}}$ and $\mathcal{L}_{\text{render}}$) [Deschaintre et al. 2018; Li et al. 2017, 2018a; Ye et al. 2018] to penalize the discrepancy occurred in the pixel space. However, this often produces blurry textures that lack high-frequency details. To alleviate this problem, we design a two-scale perceptual adversarial loss to make the network generate more vivid results. Different from conventional perceptual loss, our perceptual adversarial loss undergoes an adversarial training process.

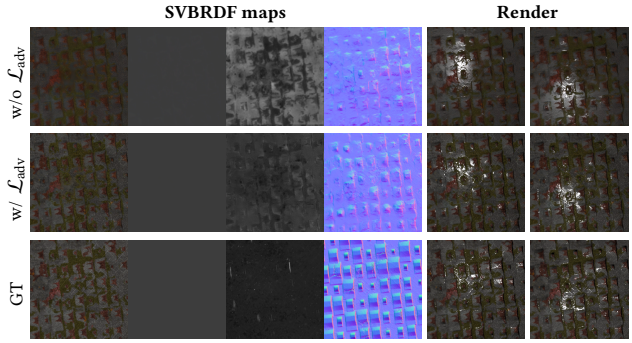

Fig. 11. Comparison between without and with the adversarial loss $\mathcal{L}_{adv}$ in training.

To evaluate the adversarial loss, we employ four discriminators operating on two different scales: two global context discriminators and two local context discriminators, as shown in the right part of Fig. 7. This design is inspired by [Iizuka et al. 2017] which also leverages global and local context discriminators to perform consistent image inpainting. We currently impose discriminators only on the recovered diffuse map and normal map considering their contributions to final renderings and the training time.

For each input image ($\mathbf{k}_d$ or $\mathbf{n}$), we first map it from the pixel space to the high-level feature space using VGG-16 [Simonyan and Zisserman 2015] (pretrained on the ImageNet dataset [Deng et al. 2009]). Then, we select the layer of `conv3_pool` and connect it to two additional convolutional layers, yielding a feature map of size $32 \times 32 \times 1$ for adversarial loss evaluation. The major difference between global context discriminator and local context discriminator is on the input. For the local context discriminator, the input image is modulated by a pre-determined mask M to focus on highlight regions. The mask is generated by thresholding the average pixel values with 0.92. This ensures that most saturated pixels are included in the mask. Specifically, for each material map ($\mathbf{k}_d$ or $\mathbf{n}$), the adversarial loss is evaluated by

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D_G(G(\mathbf{I})))] + \mathbb{E}[\log(1 - D_L(M \otimes G(\mathbf{I})))] \quad (10)$$

in which D_G and D_L represent the global context discriminator and local context discriminator, respectively. The expectation value is the average over the training images $\mathbf{I}$. The influence of the adversarial loss is shown in Fig. 11. As seen, without the adversarial loss, the generated SVBRDF may differ from the ground truth. In particular, the diffuse map and the normal map will be blurred, leading to overly smooth renderings. Quantitative analysis of the adversarial training is provided in Table 1. To show the benefit of discriminators, we evaluate different variants of our model. We respectively remove D_G, D_L, ST-Branch, and HA-Branch from our complete model and observe performance degradation in each case.

5.4 Training Details

We implement our pipeline using Tensorflow [Abadi et al. 2015]. For training, we use the dataset provided by Deschaintre et al. [2018], which contains around 200,000 training examples. Similar to MaterialGAN [Guo et al. 2020], we leave one case for images that are synthesized from the same SVBRDF but with slightly different viewing or lighting directions, and finally collect 96,294 training examples. During training, all input images are randomly cropped to 256×256 and augmented by horizontally random flipping. The training is split into two phases. First, our two-stream network is trained with pixel-wise losses for 60 epochs. Afterwards, both the two-stream network and discriminators are trained jointly until the end of training. For loss optimization, we use the Adam optimizer [Kingma and Ba 2015]. The learning rate is initialized to 0.0002, adjusted with the powered of 0.9 every three epochs until it reaches 0.00002. All other hyperparameters are set by Tensorflow's default. We train the network with a batch size of 12 for 90 epochs, and it takes about 4 days on two Tesla V100 graphics cards.

6 EXPERIMENTS

We conduct qualitative and quantitative experiments to validate our method on a wide variety of SVBRDFs from different publicly available datasets, as well as our captured images. In particular, we show our results on images that are typically hard to handle with one input image, e.g., those with obvious highlights. Please refer to the supplemental material for more results.

6.1 Comparison on Synthetic Data

We first compare our network against Rendering-Aware Deep Network (RADN) of Deschaintre et al. [2018] on synthetic data that provide us with ground-truth material maps. We also make comparison against Deep Inverse Rendering (DIR) method of Gao et al. [2019] and MaterialGAN of Guo et al. [2020] [1]. Fig. 12 provides four examples, two (left) from the dataset of [Deschaintre et al. 2018] and the other two (right) from the Adobe Stock dataset [Li et al. 2018a]. These testing examples are never used during training. RADN uses a classical U-Net architecture enriched by a global branch to recover global information from input images. Since this method does not explicitly handle improperly exposed images, it suffers from objectionable artifacts seen in the highlight regions. DIR method of Gao et al. [2019] can also accept a single-view image as input, but it is significantly dependent on initialization. In Fig. 12, we show the results with initialization provided by RADN of Deschaintre et al. [2018]. Obviously, if the initialized maps differ greatly from the reference maps, the optimized results of Gao et al. [2019] will also have large errors, as shown in the right two examples. See the strong correlations between the results of DIR and RADN. Objectionable artifacts still exist after optimization. Similarly, MaterialGAN also achieves sub-optimal results if only one image is used.

In comparison, the proposed HA convolutions suppress these artifacts by reweighting the convolutional kernels with learned masks. This allows our network to inpaint highlight saturated regions with plausible contents. In addition, our method is able to recover both globally and locally consistent features from input images thanks to the adversarial losses. This is particularly evident in the two examples from the Adobe Stock dataset. Although this dataset has a quite different data distribution with the dataset of [Deschaintre et al.

[1] Both DIR and MaterialGAN were not designed for just a single input.

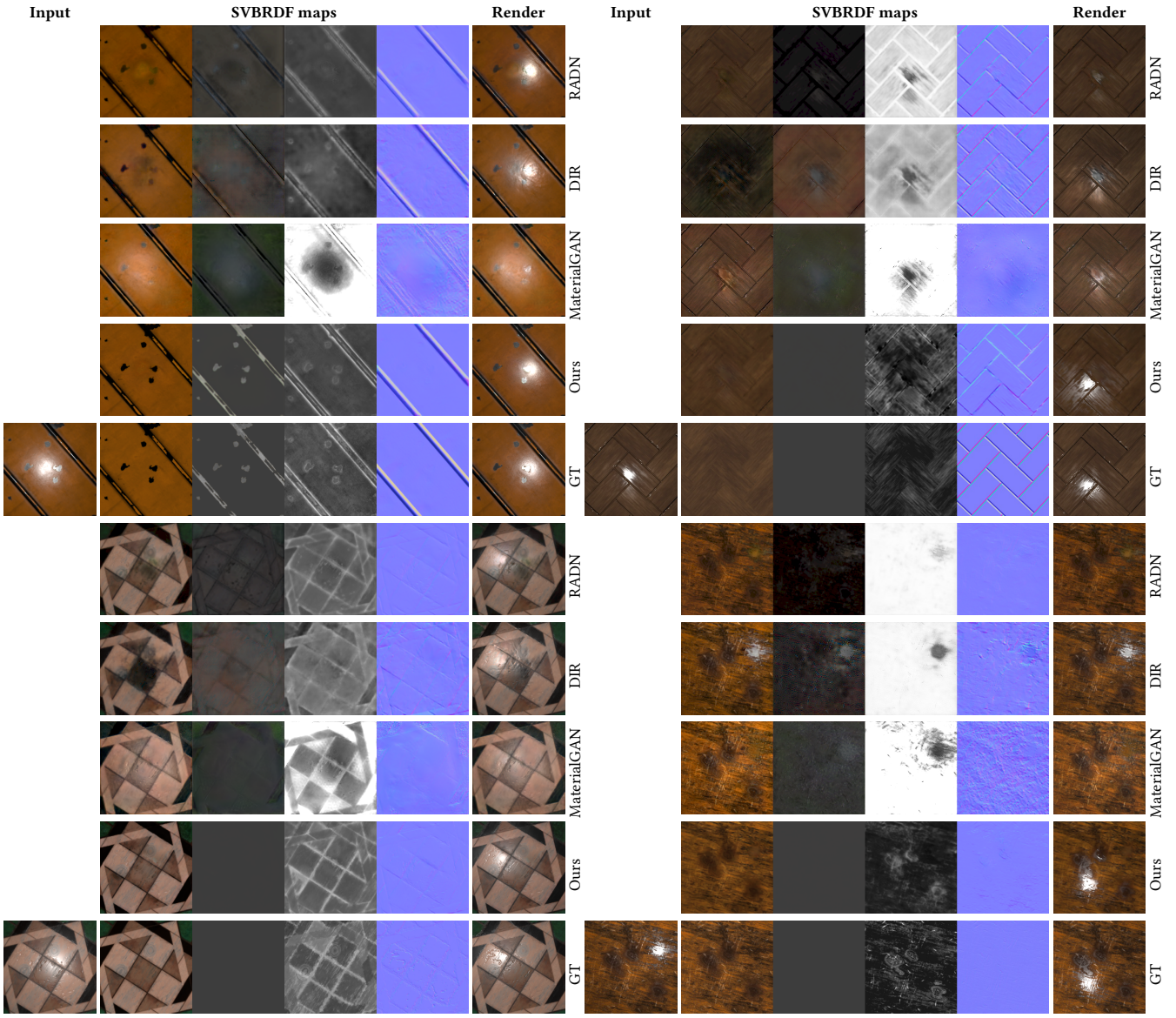

Fig. 12. Comparison against RADN of Deschaintre et al. [2018], DIR of Gao et al. [2019] and MaterialGAN of Guo et al.[2020] on synthetic data. The left two examples are from the dataset of [Deschaintre et al. 2018], while the right two examples are from the Adobe Stock dataset [Li et al. 2018a].

2018] which is used for training our model, we still faithfully recover vivid material maps. However, previous methods yield much large errors in reconstructing these two SVBRDFs. It should be noted that our method is able to disentangle different material properties due to two feature extraction streams and separated decoders, leading to less correlated material maps.

To further validate the effectiveness of our method, we conduct a quantitative analysis on both reconstructed material maps and novel renderings. The results are listed in Table 1. We randomly select 500 SVBRDFs from the dataset of [Deschaintre et al. 2018]. These SVBRDFs are not involved in training. The renderings are performed on 9 random view and lighting directions. The RMSE values clearly reveal that our method significantly outperforms previous solutions on both reconstructed material maps and renderings.

6.2 Comparison on Real Data

In Fig. 13, we validate our method on four real samples and make comparison with RADN of Deschaintre et al. [2018] and DIR of Gao et al. [2019]. Here, we provide the input images and recovered material maps, together with novel renderings under point lighting and environment lighting, respectively. The input images are taken from a hand-held consumer-level camera with the flash enabled

Table 1. Reconstruction and rendering error (in terms of RMSE) comparison between previous work (RADN [Deschaintre et al. 2018], DIR [Gao et al. 2019] and MaterialGAN [Guo et al. 2020]) and our method with different variants. The values are averaged over 500 randomly selected SVBRDFs from the dataset of [Deschaintre et al. 2018]. These SVBRDFs are never used in training. Here, we report the RMSE of each map, as well as the renderings which are performed on 9 random view and lighting directions. Best scores are highlighted in bold.

Method	Diffuse	Specular	Roughness	Normal	Render
RADN	0.056	0.109	0.113	0.035	0.055
DIR	0.083	0.102	0.127	0.043	0.065
MaterialGAN	0.064	0.083	0.148	0.037	0.066
Ours Complete	**0.029**	0.034	**0.051**	**0.032**	**0.042**
Ours w/o ST-Branch	0.033	**0.033**	0.062	0.042	0.048
Ours w/o HA-Branch	0.045	0.047	0.086	0.049	0.064
Ours w/o D_G	0.030	0.034	0.064	0.041	0.047
Ours w/o D_L	0.031	0.034	0.062	0.041	0.047

and are stored in LDR format. They all contain strong highlights which pose challenges for single-image based methods, since these highlights will generate ambiguities and saturated pixels. For previous solutions, e.g., RADN of Deschaintre et al. [2018], the recovered material properties are entangled, specifically the diffuse and specular reflectance behaviors (e.g., in the third example of Fig. 13). In contrast, our method has shown considerable success in disentangling different material properties from a single input image, thanks to the four separate decoders and the adversarial training strategy. With disentangled material maps, we are able to generate novel renderings that closely match the input. DIR of Gao et al. [2019] relies on accurate camera parameters in final optimization. On casually captured photographs without accurate cameral parameters, our method performs better than DIR. Moreover, the special design of HA convolution allows us to hallucinate missing pixels in the saturated regions. The adversarial losses also make the inpainted contents consistent with other parts. Another two examples tested on real data are provided in Fig. 1. We can see that the rendering results fit the input images quite well, with fine details and consistent specular reflection.

Although our method only uses one input image, it sometimes achieves comparable performance as those multi-image based methods, e.g., MaterialGAN [Guo et al. 2020]. Two examples are illustrated in Fig. 14. Despite some mismatches due to insufficient information contained in one image, our method still recovers major features that are close to those generated by MaterialGAN. Again, our method disentangles the material properties quite well, as evidenced in the left example where the snowflakes are expected to be glossy. In addition, we can observe that saturated pixels in the right example are recovered properly, outperforming RADN of Deschaintre et al. [2018].

6.3 Test on High-Resolution Images

Since our two-stream network is fully convolutional, images of arbitrary resolutions can be directly fed into the network, without the need of any retraining. To demonstrate this, we show two cases in Fig. 15. Here, we provide the novel-view rendered results of high-resolution (1024×1024) SVBRDFs acquired by our method. Clearly, they closely match the input. This indicates that our reconstructed high-resolution material maps are of high enough quality to render realistic specular effects under novel lighting and view directions, if the highlights are of relatively small size. The details and missing contents in saturated regions are also faithfully recovered. However, for images containing large highlights, our method may fail to hallucinate missing contents in the saturated regions since the lowest-resolution feature maps in our current network are 1/8 of the resolution of the input.

6.4 Performance

We evaluate the runtime performance on a PC with a 3.6 GHz Intel Core i7 processor and an NVIDIA GeForce RTX 2080Ti GPU. Since we put the computational burden at training stage, the SVBRDF predicting process at inference stage runs very fast. For an input image of the resolution 256×256, the prediction only takes roughly 0.17s. This is close to the runtime performance of RADN [Deschaintre et al. 2018] which also uses a feed-forward network for inference. In comparison, DIR of Gao et al. [2019] requires 58s per-image on the same platform, due to a lengthy optimization process.

7 LIMITATIONS AND FUTURE WORK

Despite high-quality SVBRDFs acquired by our single-image based pipeline, our method still has some limitations.

First, similar to some previous methods [Deschaintre et al. 2018; Li et al. 2017, 2018a; Ye et al. 2018], our method is limited to near-planar samples with little depth variation. Recovering both material properties and geometric information from a single view is an extremely challenging and ill-posed problem as strong ambiguities exist. It would be handled by more complex networks which could infer the geometry with the help of large-scale datasets containing both geometry and material.

Second, images containing large portions of highlight regions may not be well treated by our method since the HA-Branch fails to capture sufficient information from the inputs. As illustrated in Fig. 16, when the saturated pixels dominate the whole image, the recovered diffuse maps become overly dark and the normal maps lack details in overexposed regions. To address this issue, probably more inputs with different view and lighting directions should be provided to compensate large portions of missing contents. Hence, it will be interesting to extend our method to support more than one input image in the future.

Last but not least, our network is currently trained on synthetic data which will inevitably introduce biases to our network. To significantly improve the performance of our network and previous ones, a large-scale dataset with real-world examples and correct material maps is desired. However, constructing such a dataset is not an easy task considering the complexity in capturing, but we believe it will be an important future direction in this field.

8 CONCLUSION

We have presented a new framework for single-image SVBRDF acquistion. The core of our framework is a two-stream neural network

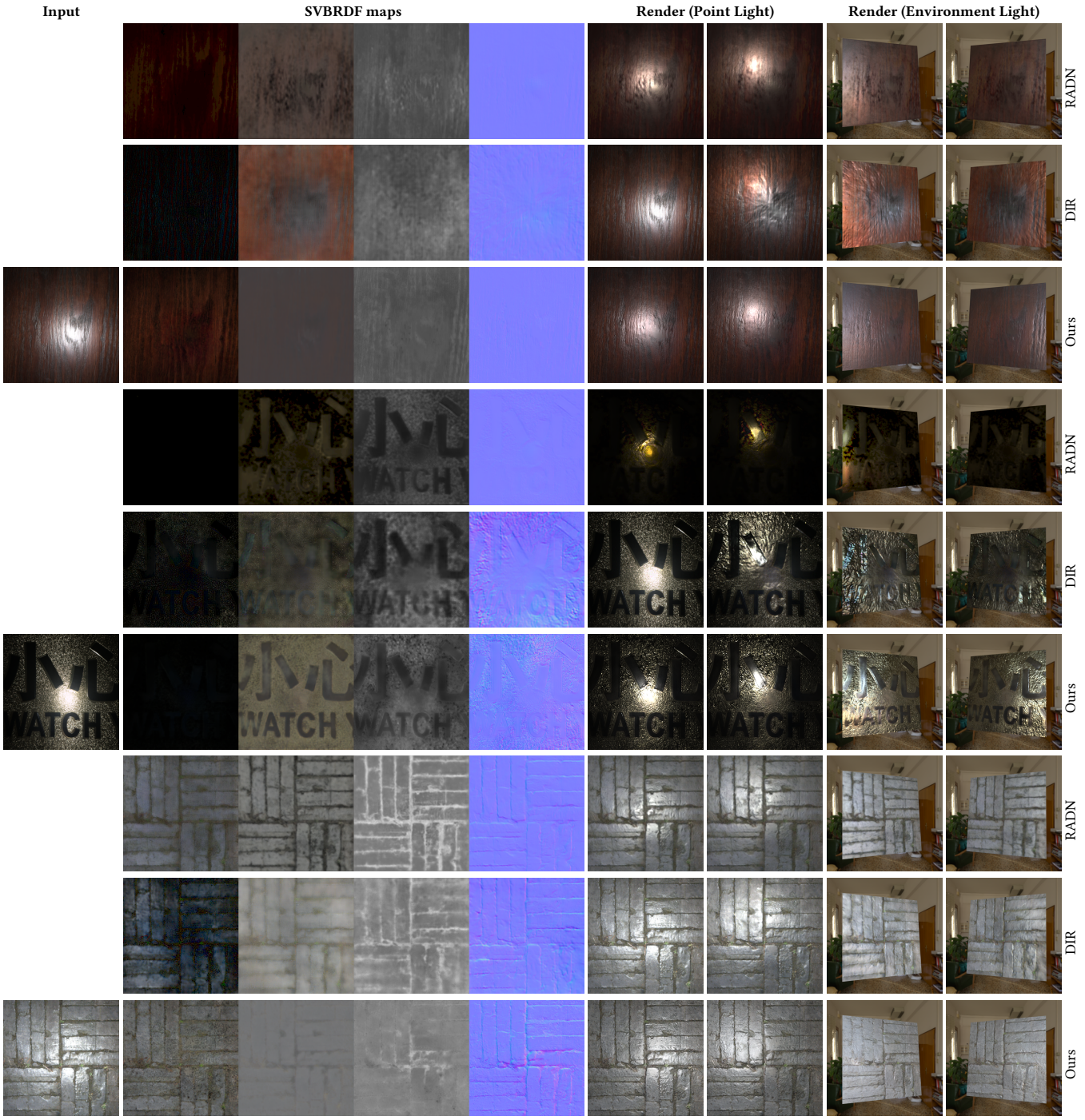

Fig. 13. SVBRDF reconstruction on real data. All images are taken from a hand-held consumer-level camera with the flash enabled. Here, we compare our method with RADN of Deschaintre et al. [2018] and DIR of Gao et al. [2019]. Note the disentangled representations obtained by our single-image method. For fair comparison, DIR only accepts one image as the input and uses the default camera parameters.

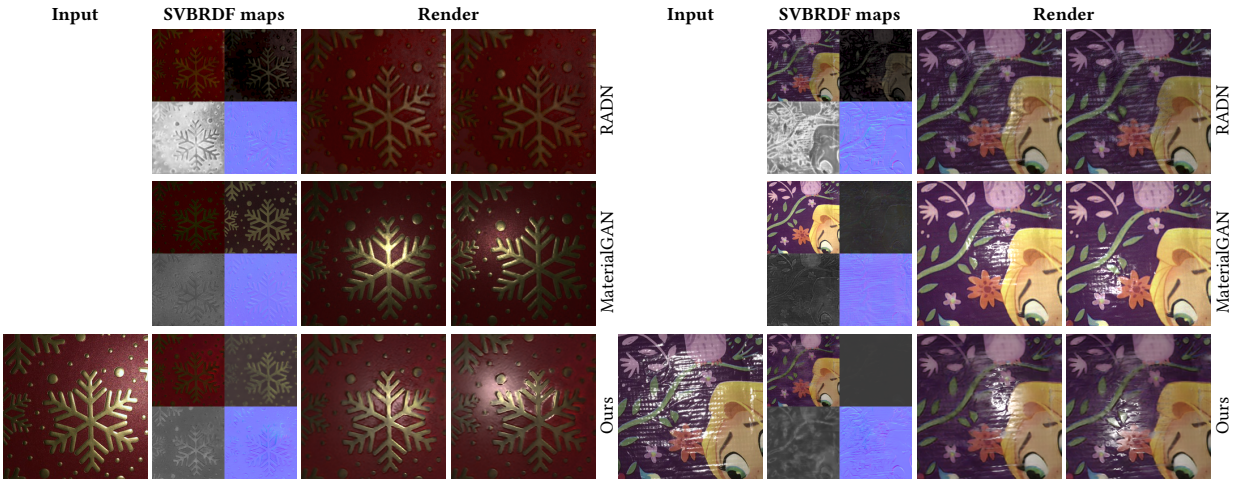

Fig. 14. Test on real data from [Guo et al. 2020]. Even with a single input image, our method can achieve plausible reconstructions that are close to the multi-view method of Guo et al. [2020] for some cases, and outperforms RADN of Deschaintre et al. [2018].

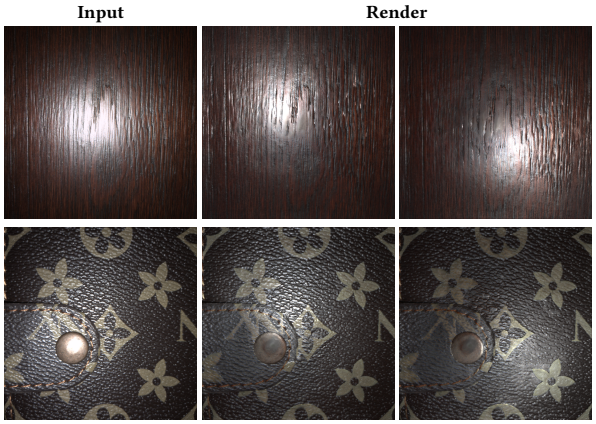

Fig. 15. Test on input images with a high-resolution (1024 × 1024).

containing several key elements: (a) a well-designed HA convolution layer for hallucinating missing contents in overexposed image regions, (b) two parallel branches for learning a complementary set of features from an input image, (c) an attention-based fusion and selection mechanism for improved feature learning and plausible SVBRDF reconstruction, and (d) a two-scale perceptual adversarial loss for enhancing texture details. Extensive experiments on both synthetic and real image datasets demonstrate consistently superior performance of the proposed method.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable comments. We also would like to thank Valentin Deschaintre from Adobe Research, Duan Gao from Tsinghua University and Yu Guo from University of California, Irvine for their kindly help. This work was supported by NSFC (62032011 and 61972194).

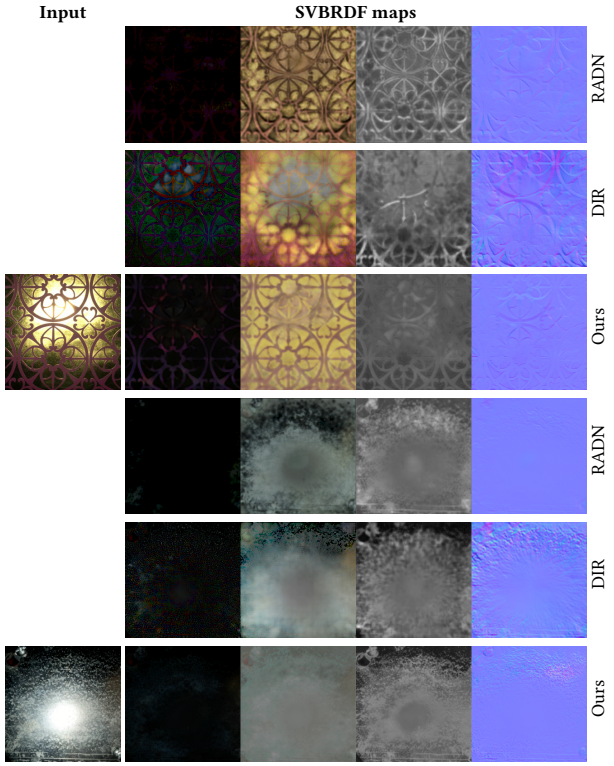

Fig. 16. Failure cases. Images containing large highlight regions may not be well treated by our method since too little information can be exploited by HA-Branch. For comparison, we also provide the results generated by RADN of Deschaintre et al. [2018] and DIR of Gao et al. [2019], respectively.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://tensorflow.org/ Software available from tensorflow.org.

Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2016. Reflectance Modeling by Neural Texture Synthesis. *ACM Trans. Graph.* 35, 4, Article 65 (July 2016), 13 pages.

Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. 2013. Practical SVBRDF Capture in the Frequency Domain. *ACM Trans. Graph.* 32, 4, Article 110 (July 2013), 12 pages.

Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. 2015. Two-Shot SVBRDF Capture for Stationary Materials. *ACM Trans. Graph.* 34, 4, Article 110 (July 2015), 13 pages.

Louis-Philippe Asselin, Denis Laurendeau, and Jean-François Lalonde. 2020. Deep SVBRDF Estimation on Real Materials. *arXiv e-prints*, Article arXiv:2010.04143 (Oct. 2020), arXiv:2010.04143 pages.

Seung-Hwan Baek, Daniel S. Jeon, Xin Tong, and Min H. Kim. 2018. Simultaneous Acquisition of Polarimetric SVBRDF and Normals. *ACM Trans. Graph.* 37, 6, Article 268 (Dec. 2018), 15 pages.

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 28, 3 (Aug. 2009).

Mark Boss, Varun Jampani, Kihwan Kim, Hendrik P.A. Lensch, and Jan Kautz. 2020. Two-Shot Spatially-Varying BRDF and Shape Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tongbo Chen, Michael Goesele, and Hans-Peter Seidel. 2006. Mesostructure from Specularity. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. 1825–1832.

Robert L. Cook and Kenneth E. Torrance. 1981. A Reflectance Model for Computer Graphics. *SIGGRAPH Comput. Graph.* 15, 3 (Aug. 1981), 307–316.

Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. 1999. Reflectance and Texture of Real-World Surfaces. *ACM Trans. Graph.* 18, 1 (Jan. 1999), 1–34.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-Image SVBRDF Capture with a Rendering-Aware Deep Network. *ACM Trans. Graph.* 37, 4, Article 128 (July 2018), 15 pages.

Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. 2019. Flexible SVBRDF Capture with a Multi-Image Deep Network. *Computer Graphics Forum* 38, 4 (2019), 1–13.

Valentin Deschaintre, George Drettakis, and Adrien Bousseau. 2020. Guided Fine-Tuning for Large-Scale Material Transfer. *Computer Graphics Forum* (2020).

Yue Dong. 2019. Deep appearance modeling: A survey. *Visual Informatics* 3, 2 (2019), 59 – 68.

Yue Dong, Guojun Chen, Pieter Peers, Jiawan Zhang, and Xin Tong. 2014. Appearance-from-Motion: Recovering Spatially Varying Surface Reflectance under Unknown Lighting. *ACM Trans. Graph.* 33, 6, Article 193 (Nov. 2014), 12 pages.

Yue Dong, Xin Tong, Fabio Pellacini, and Baining Guo. 2011. AppGen: Interactive Material Modeling from a Single Image. *ACM Trans. Graph.* 30, 6 (Dec. 2011), 1–10.

Yue Dong, Jiaping Wang, Xin Tong, John Snyder, Yanxiang Lan, Moshe Ben-Ezra, and Baining Guo. 2010. Manifold Bootstrapping for SVBRDF Capture. In *ACM SIGGRAPH 2010 Papers* (Los Angeles, California) *(SIGGRAPH '10)*. Association for Computing Machinery, New York, NY, USA, Article 98, 10 pages.

Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2019. Deep Inverse Rendering for High-Resolution SVBRDF Estimation from an Arbitrary Number of Images. *ACM Trans. Graph.* 38, 4, Article 134 (July 2019), 15 pages.

Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A. Wilson, and Paul Debevec. 2010. Circularly Polarized Spherical Illumination Reflectometry. *ACM Trans. Graph.* 29, 6, Article 162 (Dec. 2010), 12 pages.

Abhijeet Ghosh, Tim Hawkins, Pieter Peers, Sune Frederiksen, and Paul Debevec. 2008. Practical Modeling and Acquisition of Layered Facial Reflectance. *ACM Trans. Graph.* 27, 5, Article 139 (Dec. 2008), 10 pages.

D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz. 2010. Shape and Spatially-Varying BRDFs from Photometric Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 6 (2010), 1060–1071. https://doi.org/10.1109/TPAMI.2009.102

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc., 2672–2680.

D. Guarnera, G.C. Guarnera, A. Ghosh, C. Denk, and M. Glencross. 2016. BRDF Representation and Acquisition. *Computer Graphics Forum* 35, 2 (2016), 625–650.

Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020. MaterialGAN: Reflectance Capture Using a Generative SVBRDF Model. *ACM Trans. Graph.* 39, 6, Article 254 (Nov. 2020), 13 pages.

Hyunho Ha, Seung-Hwan Baek, Giljoo Nam, and Min H. Kim. 2020. Progressive Acquisition of SVBRDF and Shape in Motion. *Computer Graphics Forum* 39, 6 (2020), 480–495.

G. E. Hinton and R. R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (2006), 504–507.

Michael Holroyd, Jason Lawrence, and Todd Zickler. 2010. A Coaxial Optical Scanner for Synchronous Acquisition of 3D Geometry and Surface Reflectance. In *ACM SIGGRAPH 2010 Papers* (Los Angeles, California) *(SIGGRAPH '10)*. Association for Computing Machinery, New York, NY, USA, Article 99, 12 pages.

Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. 2014. Image Completion Using Planar Structure Guidance. *ACM Trans. Graph.* 33, 4, Article 129 (July 2014), 10 pages.

Zhuo Hui, Kalyan Sunkavalli, Joon-Young Lee, Sunil Hadap, Jian Wang, and Aswin C. Sankaranarayanan. 2017. Reflectance Capture Using Univariate Sampling of BRDFs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and Locally Consistent Image Completion. *ACM Trans. Graph.* 36, 4, Article 107 (July 2017), 14 pages.

Kaizhang Kang, Zimin Chen, Jiaping Wang, Kun Zhou, and Hongzhi Wu. 2018. Efficient Reflectance Capture Using an Autoencoder. *ACM Trans. Graph.* 37, 4, Article 127 (July 2018), 10 pages.

Kaizhang Kang, Cihui Xie, Chengan He, Mingqi Yi, Minyi Gu, Zimin Chen, Kun Zhou, and Hongzhi Wu. 2019. Learning Efficient Illumination Multiplexing for Joint Capture of Reflectance and Shape. *ACM Trans. Graph.* 38, 6, Article 165 (Nov. 2019), 12 pages.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019. Analyzing and Improving the Image Quality of StyleGAN. *CoRR* abs/1912.04958 (2019). http://arxiv.org/abs/1912.04958

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2015).

Jason Lawrence, Aner Ben-Artzi, Christopher DeCoro, Wojciech Matusik, Hanspeter Pfister, Ravi Ramamoorthi, and Szymon Rusinkiewicz. 2006. Inverse Shade Trees for Non-Parametric Material Representation and Editing. In *ACM SIGGRAPH 2006 Papers* (Boston, Massachusetts) *(SIGGRAPH '06)*. Association for Computing Machinery, New York, NY, USA, 735–745.

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling Surface Appearance from a Single Photograph Using Self-Augmented Convolutional Neural Networks. *ACM Trans. Graph.* 36, 4, Article 45 (July 2017), 11 pages.

Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2475–2484.

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. 2018a. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *Computer Vision – ECCV 2018*. Springer International Publishing, Cham, 74–90.

Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2018b. Learning to Reconstruct Shape and Spatially-Varying Reflectance from a Single Image. *ACM Trans. Graph.* 37, 6, Article 269 (Dec. 2018), 11 pages.

Y. Lin, P. Peers, and A. Ghosh. 2019. On-Site Example-Based Material Appearance Acquisition. *Computer Graphics Forum* 38, 4 (2019), 15–25.

Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. 2020. Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations. In *Computer Vision – ECCV 2020*. Springer International Publishing, Cham, 725–741.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019. Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H. Kim. 2018. Practical SVBRDF Acquisition of 3D Objects with Unstructured Flash Photography. *ACM Trans. Graph.* 37, 6, Article 267 (Dec. 2018), 12 pages.

Giljoo Nam, Joo Ho Lee, Hongzhi Wu, Diego Gutierrez, and Min H. Kim. 2016. Simultaneous Acquisition of Microscale Reflectance and Normals. *ACM Trans. Graph.* 35, 6, Article 185 (Nov. 2016), 11 pages.

Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. 2019. EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. *CoRR* abs/1901.00212 (2019).

F.E. Nicodemus, J.C. Richmond, J.J. Hsia, I.W. Ginsberg, and T. Limperis. 1977. *Geometrical considerations and nomenclature for reflectance.* Technical Report. NBS Monograph 160, U.S. Dept. of Commerce.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. 2016. Context Encoders: Feature Learning by Inpainting.

Peiran Ren, Jiaping Wang, John Snyder, Xin Tong, and Baining Guo. 2011. Pocket Reflectometry. In *ACM SIGGRAPH 2011 Papers* (Vancouver, British Columbia, Canada) *(SIGGRAPH '11)*. Association for Computing Machinery, New York, NY, USA, Article 45, 10 pages.

Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, and Ge Li. 2019. StructureFlow: Image Inpainting via Structure-Aware Appearance Flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

J. Riviere, P. Peers, and A. Ghosh. 2016. Mobile Surface Reflectometry. *Computer Graphics Forum* 35, 1 (2016), 191–202.

Jérémy Riviere, Ilya Reshetouski, Luka Filipi, and Abhijeet Ghosh. 2017. Polarization Imaging Reflectometry in the Wild. *ACM Trans. Graph.* 36, 6, Article 206 (Nov. 2017), 14 pages.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, Cham, 234–241.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR 2015*.

Borom Tunwattanapong, Graham Fyffe, Paul Graham, Jay Busch, Xueming Yu, Abhijeet Ghosh, and Paul Debevec. 2013. Acquiring Reflectance and Shape from Continuous Spherical Harmonic Illumination. *ACM Trans. Graph.* 32, 4, Article 109 (July 2013), 12 pages.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR* abs/1607.08022 (2016).

Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. 2007. Microfacet Models for Refraction through Rough Surfaces. In *Rendering Techniques*, Jan Kautz and Sumanta Pattanaik (Eds.). The Eurographics Association.

Chun-Po Wang, Noah Snavely, and Steve Marschner. 2011. Estimating Dual-Scale Properties of Glossy Surfaces from Step-Edge Lighting. *ACM Trans. Graph.* 30, 6 (Dec. 2011), 1–12.

Jiaping Wang, Shuang Zhao, Xin Tong, John Snyder, and Baining Guo. 2008. Modeling Anisotropic Surface Reflectance with Example-Based Microfacet Synthesis. In *ACM SIGGRAPH 2008 Papers* (Los Angeles, California) *(SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, Article 41, 9 pages.

Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. 2020. VCNet: A Robust Approach to Blind Image Inpainting. *arXiv preprint arXiv:2003.06816* (2020).

Tim Weyrich, Jason Lawrence, Hendrik Lensch, Szymon Rusinkiewicz, and Todd Zickler. 2008. Principles of Appearance Acquisition and Representation. In *ACM SIGGRAPH 2008 Classes* (Los Angeles, California) *(SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, Article 80, 119 pages.

Hongzhi Wu, Zhaotian Wang, and Kun Zhou. 2016. Simultaneous Localization and Appearance Estimation with a Consumer RGB-D Camera. *IEEE Transactions on Visualization and Computer Graphics* 22, 8 (2016), 2012–2023. https://doi.org/10.1109/TVCG.2015.2498617

Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. 2019. Foreground-Aware Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Z. Xu and J. Sun. 2010. Image Inpainting by Patch Propagation Using Patch Sparsity. *IEEE Transactions on Image Processing* 19, 5 (2010), 1153–1165.

Wenjie Ye, Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2018. Single Image Surface Appearance Modeling with Self-augmented CNNs and Inexact Supervision. *Computer Graphics Forum* 37, 7 (2018), 201–211.

Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. 2020. Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting With Contextual Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2019. Free-Form Image Inpainting With Gated Convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jiyang Yu, Zexiang Xu, Matteo Mannino, Henrik Wann Jensen, and Ravi Ramamoorthi. 2016. Sparse Sampling for Image-Based SVBRDF Acquisition. In *Workshop on Material Appearance Modeling*, Reinhard Klein and Holly Rushmeier (Eds.). The Eurographics Association.

Xilong Zhou and Nima Khademi Kalantari. 2021. Adversarial Single-Image SVBRDF Estimation with Hybrid Training. *Computer Graphics Forum* (2021).

Zhiming Zhou, Guojun Chen, Yue Dong, David Wipf, Yong Yu, John Snyder, and Xin Tong. 2016. Sparse-as-Possible SVBRDF Acquisition. *ACM Trans. Graph.* 35, 6, Article 189 (Nov. 2016), 12 pages.