SVBRDF Recovery from a Single Image with Highlights Using a Pre-trained Generative Adversarial Network

Tao Wen, Beibei Wang, Lei Zhang, Jie Guo, Nicolas Holzschuch

Volume xx (200y), Number z, pp. 1–13

SVBRDF Recovery From a Single Image With Highlights using a Pretrained Generative Adversarial Network

Tao Wen[1], Beibei Wang[†1,2], Lei Zhang[2], Jie Guo[3] and Nicolas Holzschuch[4]

[1]Nanjing University of Science and Technology
[2]The Hong Kong Polytechnic University
[3]Nanjing University
[4]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK

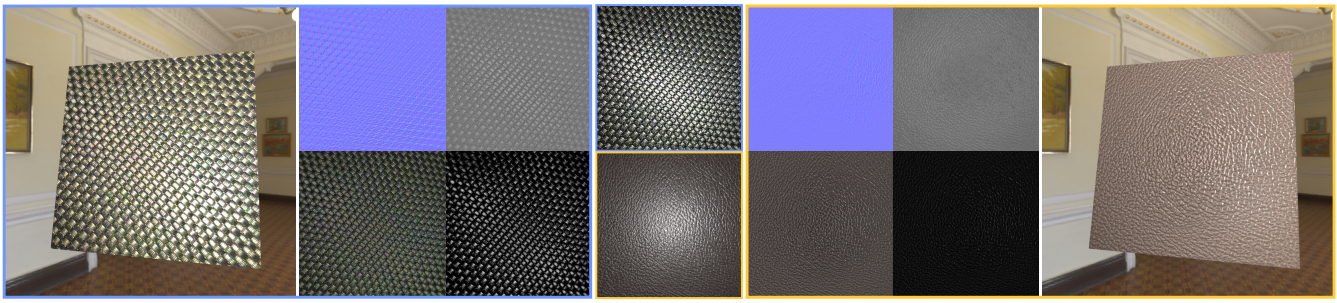

Figure 1: *Our method generates high-quality and natural SVBRDF maps from a single input photograph with overexposed highlight regions, and provides vivid rendering. Our Fourier-based loss function separates the effects due to the specular highlight from those due to the material.*

Abstract

Spatially-varying bi-directional reflectance distribution functions (SVBRDFs) are crucial for designers to incorporate new materials in virtual scenes, making them look more realistic. Reconstruction of SVBRDFs is a long-standing problem. Existing methods either rely on extensive acquisition system or require huge datasets which are nontrivial to acquire. We aim to recover SVBRDFs from a single image, without any datasets. A single image contains incomplete information about the SVBRDF, making the reconstruction task highly ill-posed. It is also difficult to separate between the changes in color that are caused by the material and those caused by the illumination, without the prior knowledge learned from the dataset. In this paper, we use an unsupervised generative adversarial neural network (GAN) to recover SVBRDFs maps with a single image as input. To better separate the effects due to illumination from the effects due to the material, we add the hypothesis that the material is stationary and introduce a new loss function based on Fourier coefficients to enforce this stationarity. For efficiency, we train the network in two stages: reusing a trained model to initialize the SVBRDFs and fine-tune it based on the input image. Our method generates high-quality SVBRDFs maps from a single input photograph, and provides more vivid rendering results compared to the previous work. The two-stage training boosts runtime performance, making it 8 times faster than previous work.

Keywords: Reflectance modeling, SVBRDF.

CCS Concepts

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

1. Introduction

The reconstruction of real world material appearance is a long standing problem in computer graphics and vision. The reflectance

† Corresponding author
Email: beibei.wang@njust.edu.cn, wentao96@njust.edu.cn

submitted to COMPUTER GRAPHICS *Forum* (1/2023).

parameters of opaque materials can be modeled by the 6D spatially-varying bi-directional reflectance distribution function (SVBRDF). It is difficult to recover the SVBRDFs of a real world material because of its high dimensionality and the inherent ambiguity of the unknown parameters: color variations could be caused by changes in any material parameter: albedo, roughness or normal. Several previous methods required complex acquisition equipments to densely sample materials in different light and view directions. These can faithfully capture the appearance parameters of a material, but are also very expensive and time-consuming, limiting the accessibility.

Recent works have shown that it is possible to recover the SVBRDFs from a few photos, or even a single image, of the material [DAD*18, LSC18, GLD*19, GSH*20, GLT*21]. These lightweight methods used deep neural networks to capture four SVBRDF maps (diffuse and specular albedo, normal map and roughness parameters) from photographs of a material. They usually rely on convolutional neural networks (CNNs), trained on synthetic images and corresponding SVBRDF maps, to model the appearance of real world materials.

These deep learning-based methods are *supervised* and require large training datasets. These datasets are difficult to acquire [LDPT17, DAD*18, AWL*15]. Existing methods either need professional designers to generate procedural models, or rely on numerous samples of real world materials. Zhao et al. [ZWX*20] proposed the first approach to exploit GAN architecture for unsupervised SVBRDF maps recovery. They rely on a two-stream generator to train the SVBRDF maps (diffuse, specular, normal, roughness) and calculate the adversarial loss. Their network is able to predict plausible SVBRDF maps from a single input image, and does not require any dataset. They also provide high quality texture synthesis through a well-designed encoder-decoder structure.

When the input image includes an intense specular highlight, it is difficult for acquisition methods to separate between albedo and illumination. As the specular highlight has a large intensity, it gets a strong priority in the learning process, often resulting in bright spots at the center of the albedo maps. To solve this issue, we need to introduce a specific constraint: we make the hypothesis that the material we acquire is stationary (its features repeat themselves after a certain period). We enforce stationarity in the reconstructed SVBRDF maps using a new loss function based on the Fourier transforms of the SVBRDF maps. Also, the rendering loss function based on pixel-wise comparisons does not work well when the exact positions of the camera and the light source are not well known. We introduce a new loss function, based on the perceptual difference between the input image and the reconstructed image [JAFF16]. Combined together, these new loss functions generate high-quality SVBRDF maps from a single input image. In particular, when there are overexposed highlights in the input photograph, our method generates more reasonable SVBRDF maps compared to previous works, leading to more realistic results when re-rendering with different view and illumination.

To speed-up the reconstruction process, we also introduce a two-stage training strategy: we pretrain our network on a single material, which provides the starting parameters for the training on each input image. This pretraining strategy makes the treatment of an image 8 times faster.

In summary, our contributions include:

- a Fourier loss function to enforce stationarity in the SVBRDFs, which produces more plausible results when the input image has intense highlights,
- a perceptual loss function that measures the semantic similarity between the input image and re-rendering result, and
- a two-stage training strategy to speedup the train process without quality degradation.

The rest of the paper is organized as follows. In Sec. 2, we review previous works on SVBRDFs recovery. As our method builds on Zhao et al. [ZWX*20], we present this method in depth in Sec. 3. We present our method and implementation in Sec. 4. We show and discuss our results in Sec. 5, and conclude in Sec. 6.

2. Related work

The problem of appearance capture has been extensively researched. Please refer to Guarnera et al. [GGG*16] and Gao et al. [GLD*19] for a more comprehensive introduction. In this paper, we focus on light-weight appearance capture, which can be grouped into multi-image methods and single-image methods according to the number of input images.

2.1. Multi-image appearance modeling

Non-learning based methods. With multiple images as input, previous works can capture the SVBRDF based on optimization usually with some domain-specific priors or assumptions, e.g. the known illumination [Cha14, HS15, RPG16], or sparsity in some domain [HSL*17, DCP*14, XDPT16]. Aittala et al. [AWL*15] used two photographs (one with flash and one without) to recover the reflectance, assuming the maps are stationary. Xu et al. [XNY*16] used two images from a near-field perspective camera, and assume spatial relation for reflectance recovery.

Learning based methods. Recently, deep learning has been widely used in appearance modeling. Deschaintre et al. [DAD*19] extracted the feature from each input image via a single-image appearance modeling network (similar to [DAD*18]) and then fused the features for SVBRDF recovery, to support arbitrary number of input images. Gao et al. [GLD*19] proposed an auto-encoder to extract the latent space from SVBRDFs as a material "prior", and then optimized material maps in this latent space to better leverage the inherent connections between maps. However, their method needs an initial value of SVBRDF maps. Guo et al. [GSH*20] trained a MaterialGAN to produce plausible material maps from a small number (3-7) of images. They used three optimizing strategies for the intermediate vector and noise vector in the latent space to learn the correlations in SVBRDF parameters. In order to tackle the shape/SVBRDF ambiguity, Boss et al. [BJK*20] designed a cascaded network for shape, illumination and SVBRDF estimation, using two images captured by a cellphone with flash both on and off.

2.2. Single-image appearance modeling

Another group of works only use a single image as input. Aittala et al. [AAL16] proposed a convolutional neural network (CNN) to extract a neural Gram-matrix texture descriptor from a single image to estimate the reflectance properties of a stationary textured materials. Under the same assumption, Zhao et al. [ZWX*20] proposed an unsupervised generative adversarial network for joint SVBRDF recovery and synthesis. When the input image has intense highlights, their method confuses material properties and tends to produce maps with a bright spot for the specular albedo. It also takes a long time to process each input image. Our method addresses both issues.

Li et al. [LDPT17] trained a CNN with a novel self-augmentation training strategy, which requires only a small number of labeled SVBRDF training pairs, to learn a large number of unlabeled photos of spatially varying materials. Ye et al. [YLD*18] improved this method and completely eliminated the need for labeled training dataset. Deschaintre et al. [DAD*18] proposed a secondary network to extract global features from each stage of an U-net architecture. They also introduce a rendering loss to enhance the estimated reflectance parameters by comparing the appearance rendered of predicted SVBRDF maps with the input image. Li et al. [LSC18] designed an in-network rendering layer to regress SVBRDF maps from a single image and a material classifier to constrain the latent representation of a CNN. They also utilized a densely-connected conditional random fields module to further refine the results. A current work by Guo et al. [GLT*21] proposed a new convolution variant called highlight-aware convolution (HA-convolution). They train the HA-convolution to "guess" the saturated pixels (specular highlight area) by the unsaturated area surrounded, making the extracted features more uniform. Their work achieves state-of-the-art performance on single-image SVBRDF acquisition and can well handle images with intense highlights. Compared to all these works, our method avoids the large dataset and learns the maps individually, under the stationary assumption.

To remove the limitation of planar materials, Li et al. [LXR*18] proposed a cascaded network architecture to recover shape and SVBRDF simultaneously from a single image. This method is further extended to handle complex indoor scenes [LSR*20].

In terms of predicting procedural texture parameters, Hu et al. [HDR19] introduced a novel framework for inverse procedural texture modeling: they trained an unsupervised clustering model to select a most appropriate procedural model and then used a CNN pool to map images to material parameters.

3. Background and Motivation

In this section, we review the network by Zhao et al. [ZWX*20], since our method relies on it. Then, we discuss the motivation of our method.

3.1. Background

We assume the light and view directions to be identical, as it happens when the flash light is close to the camera lens. We generate four maps to represent the SVBRDF: the diffuse albedo $\rho_d \in \mathbb{R}^3$, specular albedo $\rho_s \in \mathbb{R}$, roughness $\alpha \in \mathbb{R}$, and surface normal $\mathbf{n} \in \mathbb{R}^3$. Our goal is to compute $\mathbf{u} = [\rho_d, \rho_s, \alpha, \mathbf{n}]$. We use the Cook-Torrance reflectance model [CT82] for rendering.

Network architecture. Zhao et al. [ZWX*20] proposed a generative adversarial network for SVBRDF recovery and synthesis. The network is unsupervised, thus no dataset is required for training. With a stationary image as an input, the network produces four SVBRDF maps, by training on different cropped tiles from the input image. The network consists of a *two-stream generator* and a *patch discriminator*.

The generator includes an encoder and two decoders, where two groups of maps of the tiles are generated separately: normal and roughness, diffuse and specular. The maps are then used to render an image. Both this rendered image and the tile from the input image are fed to the discriminator to determine the correctness of the generated SVBRDF maps. Regarding the network structures, please see Zhao et al. [ZWX*20] for more details.

Loss Function. The loss function in Zhao et al. [ZWX*20] consists of a guessed diffuse map loss and the adversarial loss:

$$\mathcal{L}_{\text{Zhao}} = \lambda \mathcal{L}_{\text{GAN}}(G,D) + \mathcal{L}_d(G), \tag{1}$$

$$\mathcal{L}_{\text{GAN}}(G,D) = \mathbb{E}[\log Dis(\mathbf{x})] + \mathbb{E}[\log(1 - Dis(\mathbf{y})], \tag{2}$$

$$\mathcal{L}_d(G) = \mathbb{E}\left[\|\tilde{\rho_{\mathbf{d}}} - \rho_{\mathbf{d}}\|_1\right]. \tag{3}$$

The guessed diffuse map $\tilde{\rho_{\mathbf{d}}}$ is obtained via normalizing the input image, and considering the statistical distribution, which is used as the ground truth of the diffuse map, since the ground truth maps are absent.

3.2. Analysis

When highlights exist in the input image, Zhao et al. [ZWX*20] fail to recover satisfactory SVBRDF maps due to the ambiguous highlight spot. The region with high intensity is often classified as part of the specular albedo, resulting in wrong results for the roughness and specular maps (see Figure 5). This is an issue for all existing SVBRDF acquisition methods, due to the ambiguity between illumination and material.

In this paper, we introduce extra information to resolve the ambiguity: the material maps we wish to recover are stationary, that is they repeat themselves after a certain period. As a consequence, the recovered maps should also be stationary. We enforce the stationarity of the recovered maps with a new loss function, based on their Fourier transform. We focus on the acquisition part of Zhao et al. [ZWX*20] and ignore the texture synthesis part, although it can be easily included.

Zhao et al. [ZWX*20] also has an issue with computation time: as the network is trained from scratch on each individual image, processing can take up to 4 hours for a single image. We solve this issue with a two-stage training strategy, using a pretrained model for initialization and a fine-tuning stage. The computation time is down to 30 min for each image.

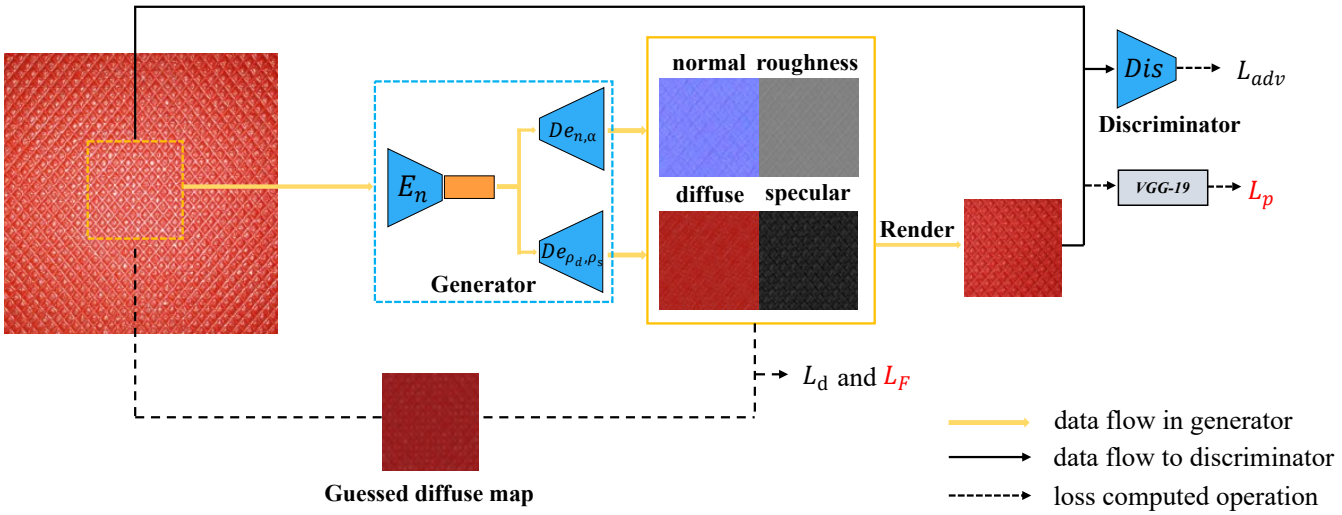

Figure 2: *Our method is similar to Zhao et al. [ZWX*20], with several key differences: we reduce the number of de-conv layers in the decoder to let the Generator output SVBRDF maps of same resolution as input. The predicted maps are used to calculate the diffuse loss L_d and Fourier loss L_F. The input tile and re-rendered image are feed into discriminator to get the adversial loss L_{adv} and a pretrained VGG-19 network to get the perceptual loss L_p. Our new loss terms are shown in red in the architecture.*

4. Our method

In this section, we propose a novel loss for the SVBRDF GAN [ZWX*20] to enforce the stationarity of SVBRDF maps and relax the pixel-wise connection between the guessed diffuse map and the input image. Then we present a two-stage training strategy to reduce the training time cost. Lastly, we show the implementation details.

4.1. Stationarity-aware loss function

We propose a joint loss, including a Fourier loss and a perceptual loss, where Fourier loss enforces the stationarity in SVBRDF maps and perceptual loss makes the rerendering result more plausible. In Figure 2, we show the difference between our method and Zhao et al. [ZWX*20].

Fourier loss. With a single image, it is difficult to separate between the color changes due to the material and those due to illumination. Without guidance, the network tends to place the highlights as part of the albedo or normal map. We introduce an extra constraint: the material should be stationary. In a stationary texture, variations are high-frequency and illumination effects are low-frequency. We introduce a new loss function based on Fourier analysis to enforce this stationarity: we compute the Fourier transform of the guessed diffuse map $\tilde{\rho_d}$ and of the predicted SVBRDF maps $\mathbf{u}([\rho_d, \rho_s, \alpha, \mathbf{n}])$. We compute the Fourier loss function as the L_1 loss in the logarithmic domain:

$$\mathcal{L}_F(\mathbf{u}, \tilde{\rho_d}) = \log \mathbb{E}\left[\|\mathrm{FFT}(\mathbf{u}) - \mathrm{FFT}(\tilde{\rho_d})\|_1\right]. \tag{4}$$

The Fourier L_1 loss is to ignore outliers in the frequency domain while the logarithmic compression is to reduce the impact of large errors comparing to other losses.

We use the fact that, after normalization, the guessed diffuse map has a stationary distribution of gray scale. With Fourier loss as a guidance on the frequency domain, the predicted SVBRDF maps will be less affected by the illumination and have similar variations as guessed diffuse map.

Perceptual loss. The exact light and view directions are sometimes unknown, especially for captured photographs. Using loss functions based on pixel-wise difference between input images and rendered images produces poor results, because we cannot guarantee the consistency of our rendering parameters: the rendered image is not rendered with exactly the same view and light condictions as the input image. To solve this issue, we use a perceptual loss function, to measure the semantic similarity between the input I and the re-rendered result R, via a pretrained VGG-19 network [SZ14]:

$$\mathcal{L}_P(I, R) = \mathbb{E}\left[\|\mathrm{VGG}(I) - \mathrm{VGG}(R)\|_1\right]. \tag{5}$$

The perceptual loss is able to improve the visual quality, since it does not require pixel-to-pixel alignment. With this perceptual rendering loss, the re-rendering result of predicted SVBRDF maps is more realistic and reliable.

Summary. We present a joint loss function combining these three losses:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{Zhao}} + \lambda_1 \mathcal{L}_F(\mathbf{u}, \tilde{\rho_d}) + \lambda_2 \mathcal{L}_P(I, R) \tag{6}$$

Trained with this joint loss function, the network can achieve better recovery result of SVBRDFs, leading to more reasonable rendering results with novel view and light directions, especially when handling images with intense highlights. We will show our recovery results in Sec. 5.1.

Discussion. Our loss function assumes that all maps being computed for the current SVBRDF have the similarly structured patterns with the same frequency, and that non-stationarity comes from illumination. This is a reasonable expectation for a large class

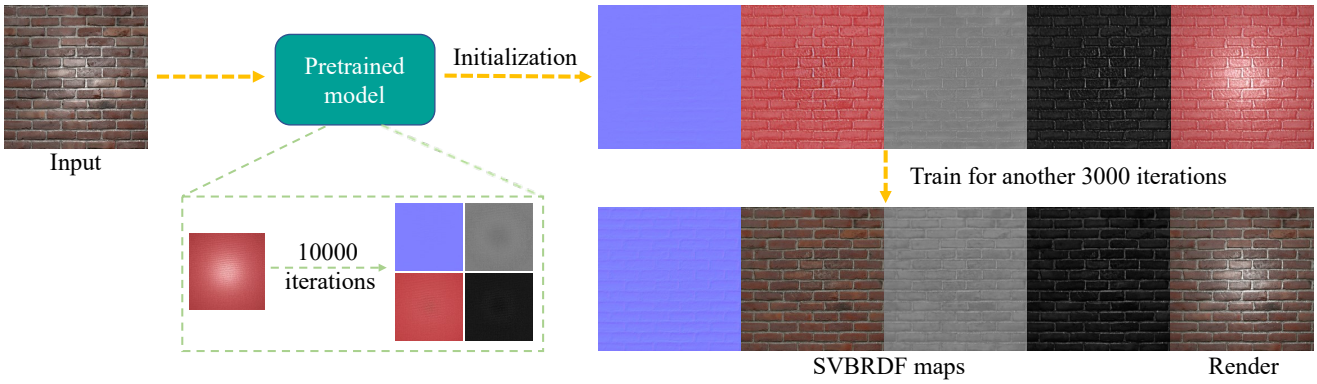

Figure 3: *Our two-stage training strategy. At the first stage, we train the network on one image to get the pretrained model. For any new input images, we use this pretrained model as an initialization to start our second stage training. Note that at second stage, texture information already exists in the initialized SVBRDF maps, without extra training.*

of materials (leather, fabrics, wallpapers), but can be wrong for other materials, with local patterns. We also assume that all maps have the same frequency in their patterns, and that we can use the guessed diffuse map as a guide for learning in the other maps.

4.2. Two-stage training strategy

Training the network from scratch for each input image is time consuming: all parameters in the network have to be initialized as random value and trained over and over again for each new input. It can take several hours for the network to converge.

To improve this process, we propose a two-stage training strategy. We first train our network on an image (e.g. red book) for 10,000 iterations to get a pretrained model. For a new input (e.g. brick), we use this pretrained model as an initialization of network parameters, and then train on this model for another 3,000 iterations to get the "plausible" SVBRDF maps. The key insight is that the generator acts as a prior knowledge about how the four maps look like in general after training for 10,000 iterations:

- the RGB value of normal map is close to light blue, due to the planar property of material,
- the roughness map looks "grey",
- the specular map is "dark", and
- the color of diffuse map mostly depends on the color of input image.

Besides, as shown in Figure 3, without any extra training, the generator could recover texture information of the new input image in SVBRDF maps (although the color is not "correct"). Apparently, with the pretrained parameters as a good initialization, it becomes relatively easier for the network to get a plausible recovery result of SVBRDF maps, comparing to training from scratch.

We have tried different pretrained models with different input images and find little difference in the final results. We provide some results in Figure 11.

4.3. Training and implementation

We implemented our framework in TensorFlow. The generator and discriminator were trained using Adam optimizers with a fixed learning rate of 2e-5. We set the hyper-parameter λ (in Eq.(1)), λ_1, λ_2 to 0.1, 0.1, 0.2 respectively. At the first stage, we train our network on an arbitrary input image for 10,000 iterations from scratch to obtain a pretrained model. Then with a new image as input, we fine-tune the pretrained network for another 3,000 iterations to get a plausible result. It takes about 2 hours to get the pretrained model, which we then use for any new input images. It then costs about 30 minutes to train the model on a new image, using a RTX 2080Ti GPU. Note that training from scratch for a new input image, rather than using the pretrained model, requires 20,000 iterations with 4 hours, using the same GPU. Hence, the pretrained model achieves an approximate 8 times speedup.

5. Results and discussion

We first compare the results of our network with Gao et al. [GLD*19], Guo et al. [GSH*20], Guo et al. [GLT*21] and Zhao et al. [ZWX*20] on both synthetic images and captured photos (Sec. 5.1). Then we show the influence of different loss terms by a loss ablation experiment (Sec. 5.2) and the effects of our two-stage training strategy (Sec. 5.3).

5.1. Comparison with previous works

We ran our experiments on images with strong highlights to show the effectiveness of our approach. The input real photos and reference maps are from the two-shot dataset [AWL*15] and free material websites†. For the two-shot dataset, we cropped the 3264×2448 SVBRDF maps to 1600×1600 and resized them to 1024×1024 resolution. We render the input synthetic images using the Cook-Torrance reflectance model [CT82] with a point light

† https://texturehaven.com
† https://ambientcg.com

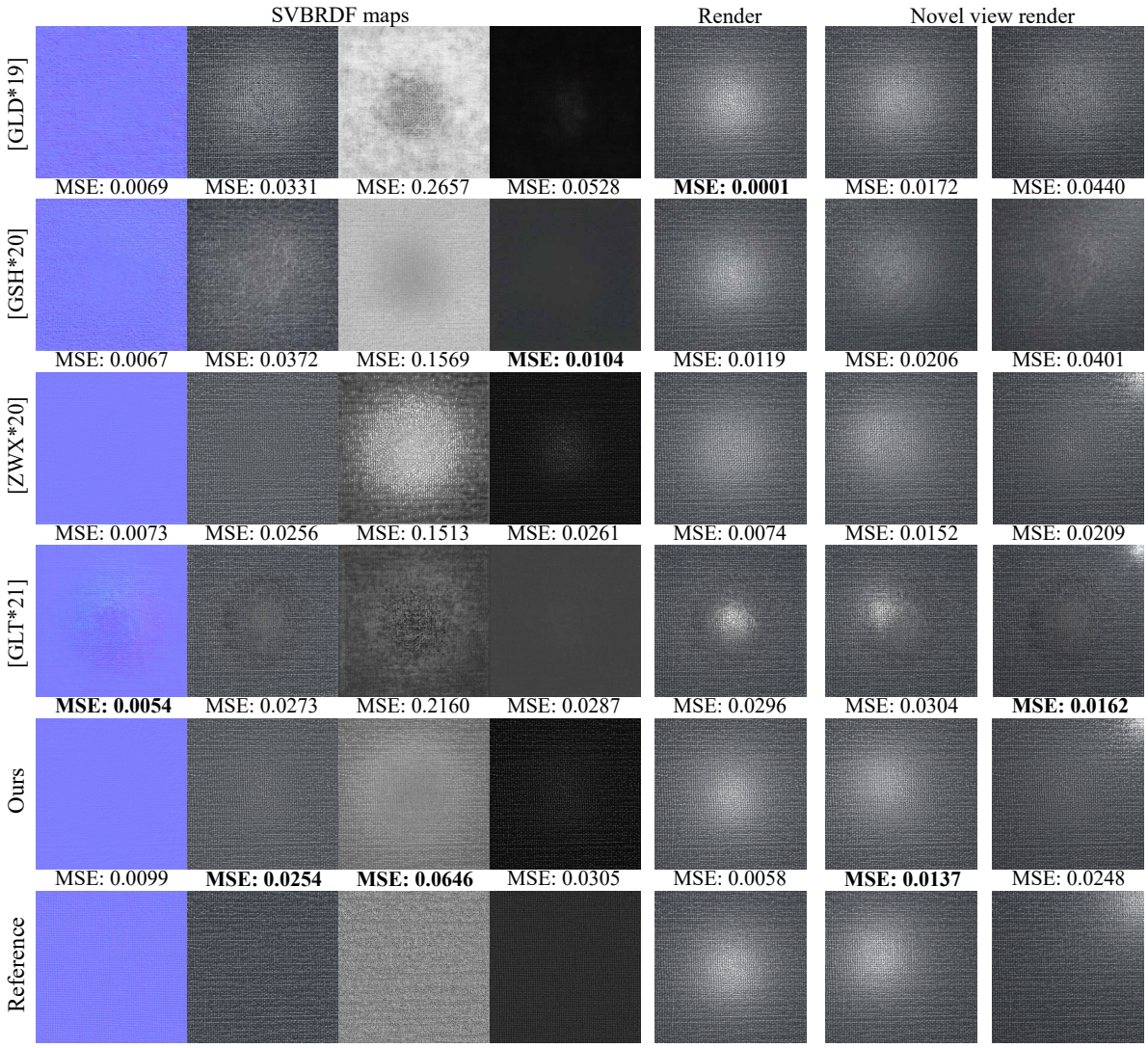

Figure 4: *SVBRDF maps recovered from synthetic images of* 1024×1024*, compared with Gao et al. [GLD*19], Guo et al. [GSH*20], Guo et al. [GLT*21] and Zhao et al. [ZWX*20]. Note how the specular highlight at the center of the image is challenging for all acquisition methods, resulting in dark or bright areas in the specular albedo map, and flattened areas in the normal map. We use the network of [DAD*18] as initialization for [GLD*19] and set the number of input images to one. Guo et al. [GSH*20] can only produce* 256×256 *maps, so we scale them to proper size for comparison. The lowest error (MSE) is marked in bold.*

and camera right above the center of the plane, consistent with our re-rendering process in the network.

Comparison on synthetic images. In Figure 4 and Figure 5, we compare our results on synthetic images with Gao et al. [GLD*19], Guo et al. [GSH*20], Guo et al. [GLT*21] and Zhao et al. [ZWX*20]. For Gao et al. [GLD*19], we use the network of Deschaintre et al. [DAD*18] as an initialization and set the number of inputs to one for fair comparison. The errors (mean square error, MSE) between the maps / rendered images and the references are shown below the images. Although Gao et al. [GLD*19] achieve plausible rendering results through optimization steps, strong artifacts still exist in the recovered SVBRDF maps, leading to poor performance in novel view rendering. Guo et al. [GSH*20] also produce unpleasant SVBRDF maps and novel view rendering. Zhao et al. [ZWX*20] preserve some details in SVBRDF maps but still suffer from highlight regions, especially in roughness map and specular map. Our method recovers the stationarity in SVBRDF maps, which are comparable to the reference maps, thus lead to more plausible appearance in novel view rendering.

We provide more results on synthetic images in Figure 6. Comparing to Zhao et al. [ZWX*20], our method performs better on both recovered SVBRDF maps and rendering results visually and quantitatively. As shown in the first row of the six examples, with only one image as input, Zhao et al. [ZWX*20] tend to produce

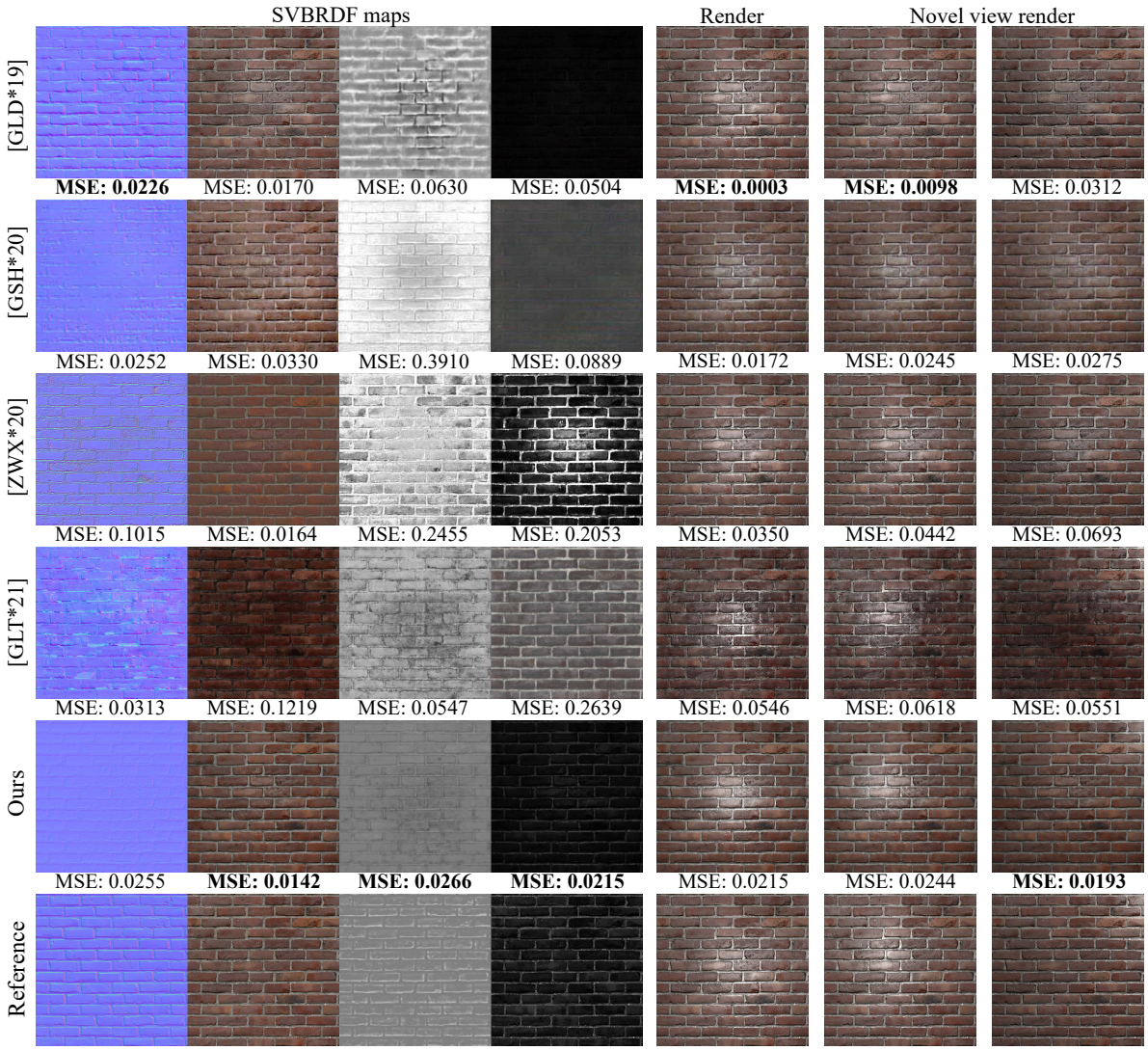

Figure 5: *SVBRDF maps recovered from synthetic images of* 1024×1024*, compared with Gao et al. [GLD*19], Guo et al. [GSH*20], Guo et al. [GLT*21] and Zhao et al. [ZWX*20]. Note how the specular highlight at the center of the image is challenging for all acquisition methods, resulting in dark or bright areas in the specular albedo map, and flattened areas in the normal map. We use the network of [DAD*18] as initialization for [GLD*19] and set the number of input images to one. Guo et al. [GSH*20] can only produce* 256×256 *maps, so we scale them to proper size for comparison. The lowest error (MSE) is marked in bold.*

"polluted" SVBRDF maps and unpleasant novel view rendering results. By comparison, our method produces clearer roughness maps and specular maps which are much less affected by highlight regions, leading to more vivid rendering results.

Comparison on captured photos. In Figures 7 and 8, we validate our method on three captured photos, comparing to Gao et al. [GLD*19], Guo et al. [GSH*20], Guo et al. [GLT*21] and Zhao et al. [ZWX*20]. The photos are from the two-shot dataset [AWL*15], cropped so that the brightest part of the image is at the center. There are no reference SVBRDFs for the captured photos, we only provide the MSE between input photos and rendering results. As shown in Figure 7, Gao et al. [GLD*19] and Guo et al. [GLT*21] produce "polluted" SVBRDF maps that are highly affected by highlight regions, while Guo et al. [GSH*20] produce blurred results that lack detailed structure in SVBRDF maps. Zhao et al. [ZWX*20] produce plausible diffuse maps, but still suffer from highlight regions in the other maps. All the previous works fail to handle the ambiguity in reflectance and illumination, yielding unpleasant novel view rendering results. Our method produces more stationary SVBRDF maps and more plausible rendering results under novel views. As shown in Figure 8, our method recovers detailed variations in normal map and suppresses the bright spots in other three maps. Thus, our method better handles the ambiguity in

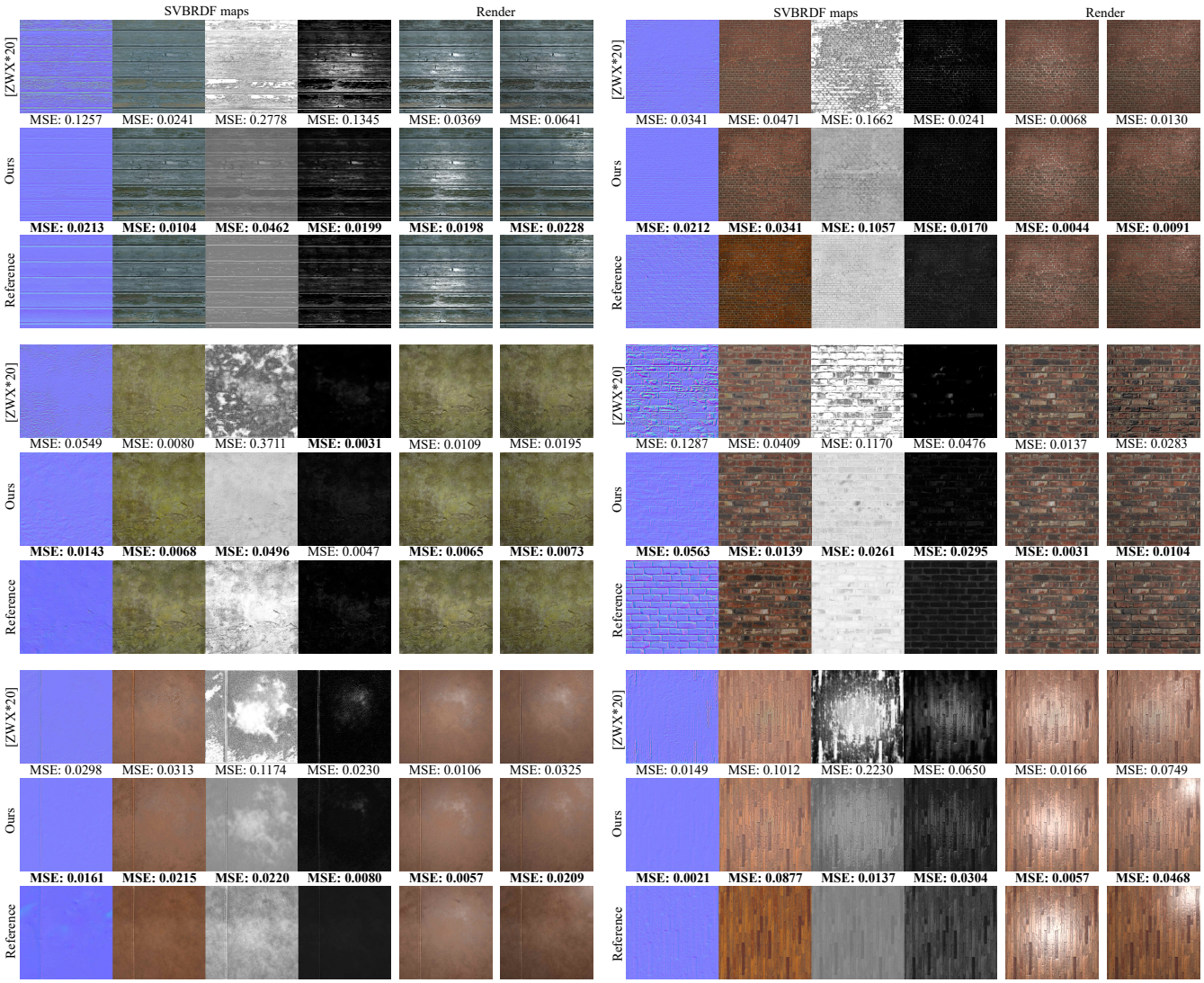

Figure 6: *More results on synthetic images, comparing with Zhao et al. [ZWX*20]. Our method could reduce the impact of highlights on SVBRDFs recovery. With only one image as input, it could produce SVBRDF maps closer to the references, leading to more vivid rendering results.*

reflectance and illumination, producing more reasonable rendering results under novel views.

5.2. Ablation Study

There are two important components in our SVBRDF recovery network: Fourier loss and perceptual loss. We ran an ablation study to validate the impacts of these components in Figure 9. We compare Zhao et al. [ZWX*20], our model (with Fourier loss $\mathcal{L}_F$ only), our model (with both Fourier loss $\mathcal{L}_F$ and perceptual loss $\mathcal{L}_P$), our model (without "guessed" diffuse loss $\mathcal{L}_d$) and the reference.

Zhao et al. [ZWX*20] (first row) suffer from bright spots in the roughness and specular maps, and over-flat normal map, leading to obvious difference from the reference SVBRDF maps, where the material properties are stationary. By introducing the Fourier loss, the predicted maps (second row) become more uniform and have less artifacts at the center. The novel view rendering results have more pronounced variations in illumination due to the more uniform SVBRDF maps, thus our method produces more plausible appearance compared to Zhao et al. [ZWX*20]. However, we found that the texture variation is blurred in the highlight region. Further introducing perceptual loss (third row) solves this issue, via measuring the semantic similarity between the input image and re-rendering result. The joint loss function with both Fourier loss and perceptual loss produces the best results. The bright spots in the SVBRDF maps have been removed, making them decoupled from

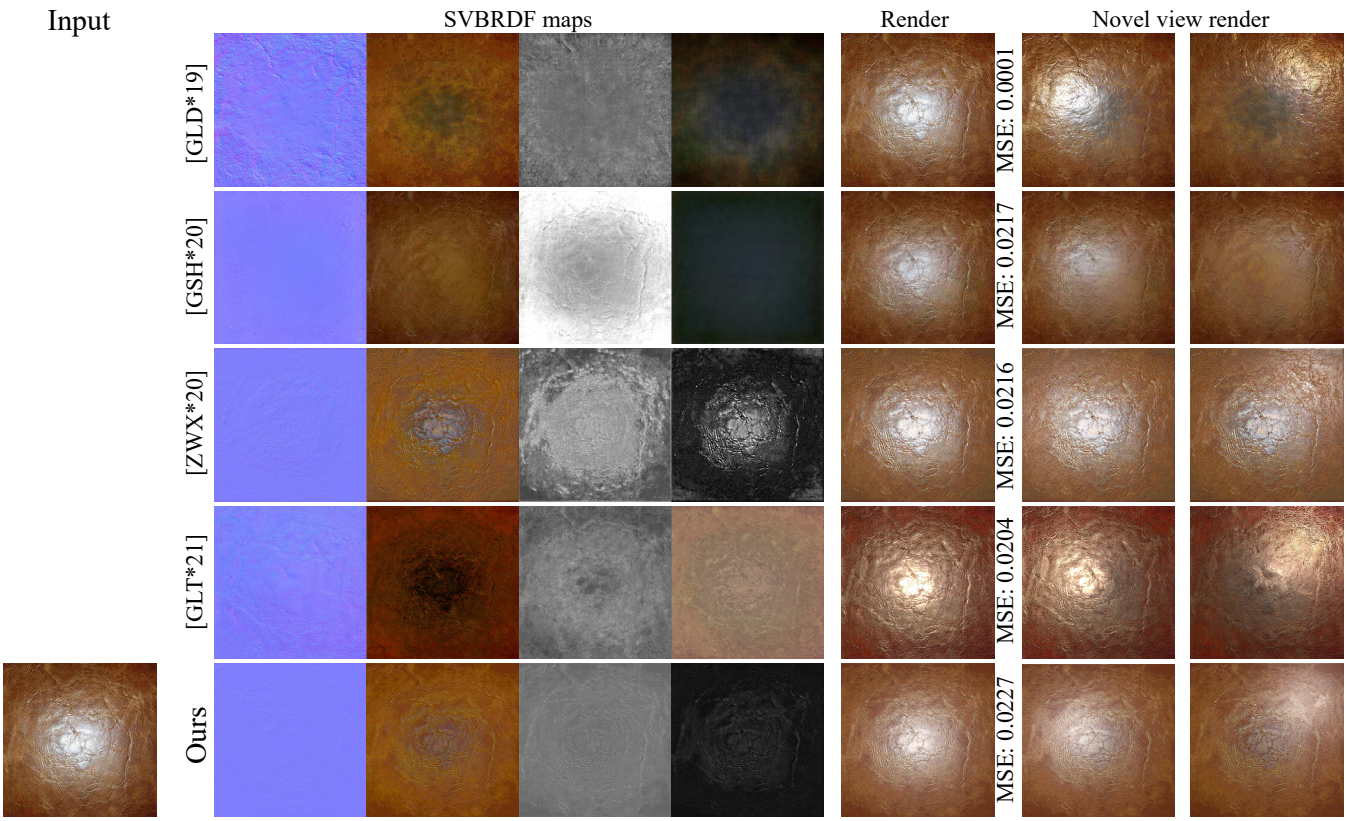

Figure 7: *SVBRDF maps recovered from real photos of* 1024×1024*, compared with Gao et al. [GLD*19], Guo et al. [GSH*20], Guo et al. [GLT*21] and Zhao et al. [ZWX*20]. Note how the highlight at the center of the image is challenging for all acquisition methods. Our method produces more stationary SVBRDF maps and more plausible rendering results, compared to previous works.*

the illumination in the input images, thus leading to more plausible re-rendering results under different view and light directions. We also validate the impact of the "guessed" diffuse loss, as shown in the fourth row of Figure 9, without the guessed diffuse map as a guidance, the predicted maps would be terribly wrong, leading to unpleasant novel view rendering results. More results are shown in the supplemental materials.

We also tried using the Fourier loss on only some of the SVBRDF maps, such as the roughness or specular albedos. The results of this study are shown in Figure 10. The maps that were computed without the Fourier loss are highly affected by the bright spot, while the other maps are not. We used the Fourier loss on all four maps to ensure the stationarity in all SVBRDF maps.

5.3. Validation of the two-stage training strategy

In Figure 11, we compare the SVBRDF maps recovered with the two-stage training strategy and one-stage training strategy (without pretraining). For the two-stage training strategy, we show the results recovered from two pretrained models which are trained with different images (book-red and brick). For the one-stage training strategy (without pretraining), the training is performed on the input image from scratch, with 20,000 iterations. By comparison, we find that the difference between the results of one-stage and two-stage training strategies is subtle, and the difference with different pretrained models is also not obvious. At the bottom row of Figure 11, we also compare the MSE of both SVBRDF maps and rendering results w.r.t. two different strategies. The network without pretraining takes about 10,000 iterations to converge, while the two-stage training strategy takes only about 1,200 iterations. Also, their converged losses are quite similar. Thus, we believe that our two-stage training strategy greatly reduces the training time, without any quality degradation and the pretrained model can be trained on arbitrary single image under our assumption.

5.4. Limitations

We have identified some limitations for our method.

First, we assume that all maps have the same frequency in their patterns, such that we can use the guessed diffuse map as a guide for learning in the other maps. If the SVBRDF maps have different

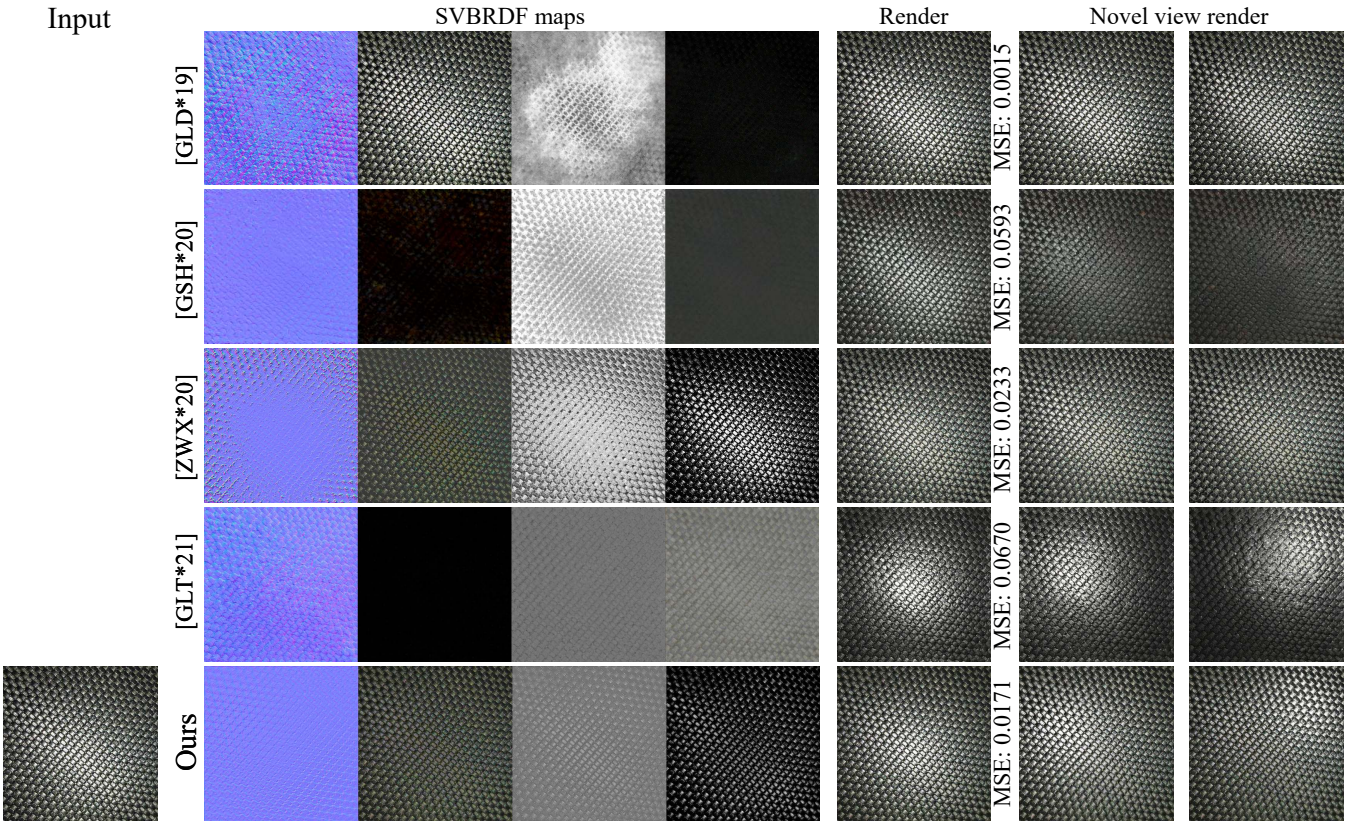

Figure 8: *SVBRDF maps recovered from real photos of* 1024×1024*, compared with Gao et al. [GLD*19], Guo et al. [GSH*20], Guo et al. [GLT*21] and Zhao et al. [ZWX*20]. Our method produces more stationary SVBRDF maps and more plausible rendering results, comparing to previous works.*

frequency, our Fourier loss could provide the wrong guidance, as shown in Figure 12.

Second, our method does not work well for input images with sharp contrast, as shown in Figure 13. The highlights are so strong that the texture information in this region is almost obscured, which prevent us from getting a plausible guessed diffuse map. This suggests that our method is not completely unaffected by the highlights, but in most cases it is able to produce more plausible SVBRDF maps and rendering results, comparing to previous works.

Last, our method tends to produce blurred normal map when the material has strong normal variations. This is due to the inherent ambiguity between normal variation and albedo — there is only one shotted image as input and no datasets for training. Without any other information, the network tends to incorrectly bake the shading effects onto the diffuse map.

6. Conclusion

In this paper, we improve the unsupervised SVBRDF recovery generative adversarial network, by introducing two new loss functions: a Fourier loss function and a perceptual loss function to enforce the stationarity of SVBRDFs, yielding better quality for the SVBRDF maps, especially for input images with intense highlights. Then, we propose a two-stage training strategy to reduce the training time with $8\times$ speedup. In the end, our method is able to generate high-quality SVBRDFs and produce more plausible rendering results, compared with the acknowledged state-of-the-art methods.

Acknowledgments

We thank the editors and reviewers for the valuable comments and suggestions. This work has been partially supported by the National Natural Science Foundation of China under grant No. 62172220 and the Fundamental Research Funds for the Central Universities under grant No. 30920021133.

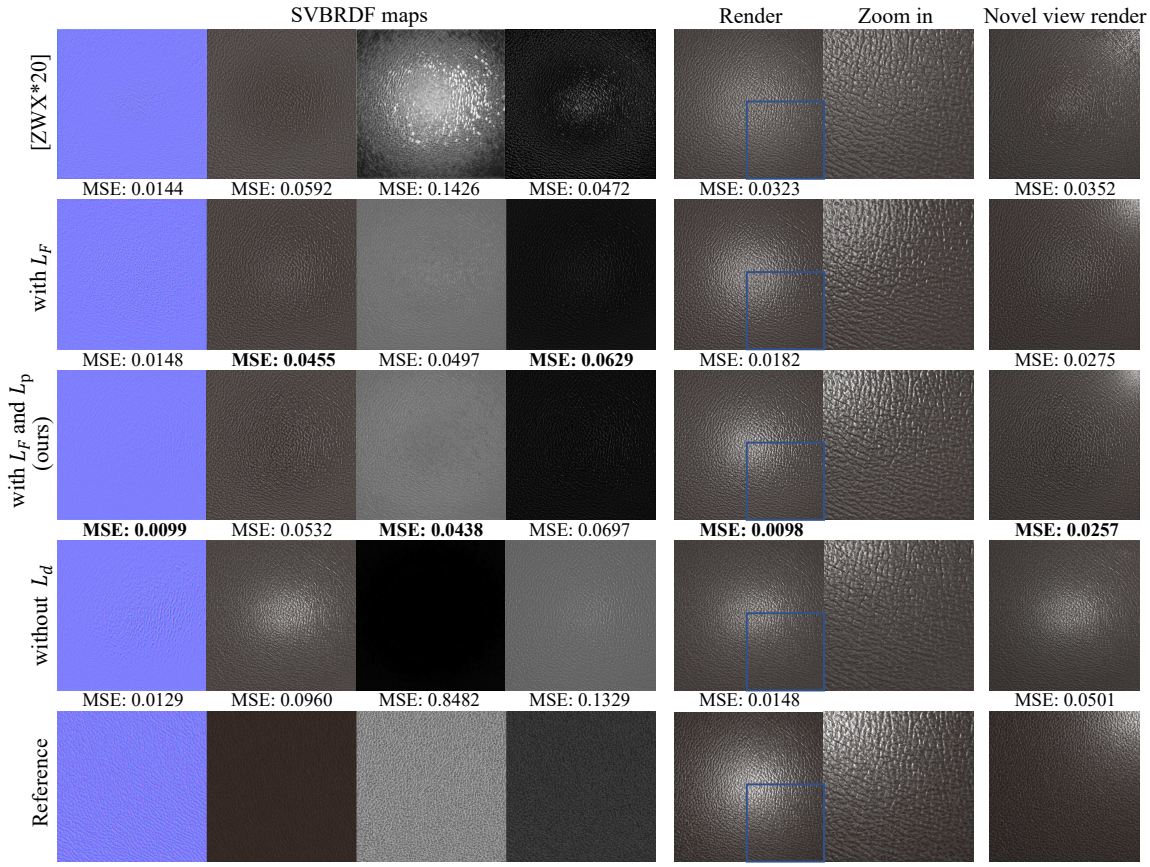

Figure 9: *Ablation study on several maps to validate the impacts of Fourier loss, perceptual loss and "guessed" diffuse loss in our SVBRDF recovery network.*

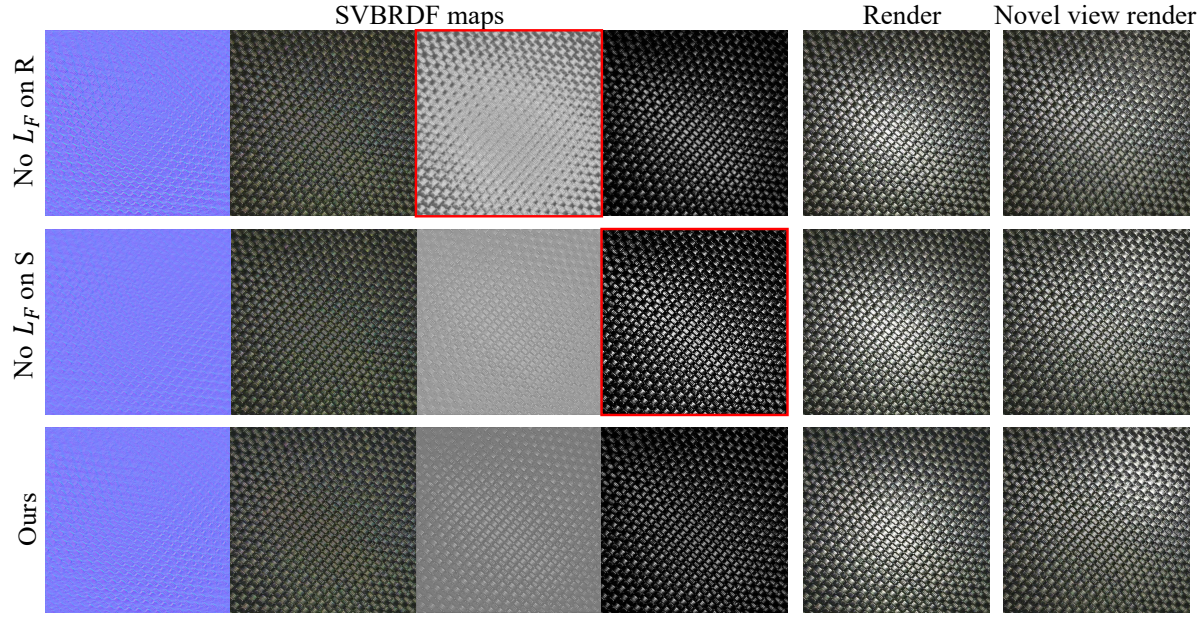

Figure 10: *Influence of the Fourier loss on different maps. Without Fourier loss on roughness map (R) or specular map (S), the bright spot still exist.*

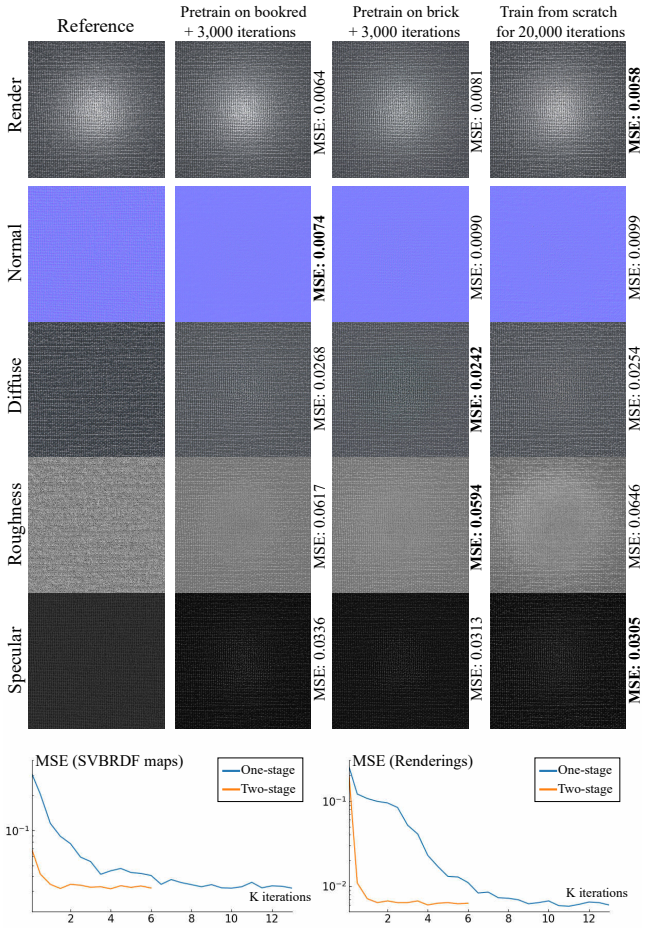

Figure 11: *Comparison on different pretrained models and different training strategies. There is little difference in the three columns of results. The bottom row shows the mean square error as a function of iterations on two different training strategies.*

References

[AAL16] AITTALA M., AILA T., LEHTINEN J.: Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (TOG) 35*, 4 (2016), 65. 3

[AWL*15] AITTALA M., WEYRICH T., LEHTINEN J., ET AL.: Two-shot svbrdf capture for stationary materials. *ACM Trans. Graph. 34*, 4 (2015), 110–1. 2, 5, 7

[BJK*20] BOSS M., JAMPANI V., KIM K., LENSCH H. P., KAUTZ J.: Two-shot spatially-varying brdf and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020). 2

[Cha14] CHANDRAKER M.: On shape and material recovery from motion. In *European Conference on Computer Vision* (2014), Springer, pp. 202–217. 2

[CT82] COOK R. L., TORRANCE K. E.: A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG) 1*, 1 (1982), 7–24. 3, 5

[DAD*18] DESCHAINTRE V., AITTALA M., DURAND F., DRETTAKIS G., BOUSSEAU A.: Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 128. 2, 3, 6, 7

[DAD*19] DESCHAINTRE V., AITTALA M., DURAND F., DRETTAKIS G., BOUSSEAU A.: Flexible svbrdf capture with a multi-image deep network. *Computer Graphics Forum 38*, 4 (2019). 2

[DCP*14] DONG Y., CHEN G., PEERS P., ZHANG J., TONG X.: Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting. *ACM Transactions on Graphics (TOG) 33*, 6 (2014), 193. 2

[GGG*16] GUARNERA D., GUARNERA G., GHOSH A., DENK C., GLENCROSS M.: Brdf representation and acquisition. *Computer Graphics Forum 35*, 2 (2016), 625–650. 2

[GLD*19] GAO D., LI X., DONG Y., PEERS P., XU K., TONG X.: Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 134. 2, 5, 6, 7, 9, 10

[GLT*21] GUO J., LAI S., TAO C., CAI Y., WANG L., GUO Y., YAN L.-Q.: Highlight-aware two-stream network for single-image svbrdf acquisition. *ACM Transactions on Graphics (TOG) 40* (2021), 1 – 14. 2, 3, 5, 6, 7, 9, 10

[GSH*20] GUO Y., SMITH C., HAŠAN M., SUNKAVALLI K., ZHAO S.: Materialgan: reflectance capture using a generative svbrdf model. *ACM Transactions on Graphics (TOG) 39*, 6 (2020), 1–13. 2, 5, 6, 7, 9, 10

[HDR19] HU Y., DORSEY J., RUSHMEIER H.: A novel framework for inverse procedural texture modeling. *ACM Transactions on Graphics (TOG) 38*, 6 (2019), 1–14. 3

[HS15] HUI Z., SANKARANARAYANAN A. C.: A dictionary-based approach for estimating shape and spatially-varying reflectance. In *2015 IEEE International Conference on Computational Photography (ICCP)* (2015), IEEE, pp. 1–9. 2

[HSL*17] HUI Z., SUNKAVALLI K., LEE J.-Y., HADAP S., WANG J., SANKARANARAYANAN A. C.: Reflectance capture using univariate sampling of brdfs. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 5362–5370. 2

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (2016), Springer, pp. 694–711. 2

[LDPT17] LI X., DONG Y., PEERS P., TONG X.: Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 45. 2, 3

[LSC18] LI Z., SUNKAVALLI K., CHANDRAKER M.: Materials for masses: Svbrdf acquisition with a single mobile phone image. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 72–87. 2, 3

[LSR*20] LI Z., SHAFIEI M., RAMAMOORTHI R., SUNKAVALLI K., CHANDRAKER M.: 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* (2020), pp. 2475–2484. 3

[LXR*18] LI Z., XU Z., RAMAMOORTHI R., SUNKAVALLI K., CHANDRAKER M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG) 37*, 6 (2018), 1–11. 3

[RPG16] RIVIERE J., PEERS P., GHOSH A.: Mobile surface reflectometry. *Computer Graphics Forum 35*, 1 (2016), 191–202. 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 4

[XDPT16] XIA R., DONG Y., PEERS P., TONG X.: Recovering shape and spatially-varying surface reflectance under unknown illumination. *ACM Trans. Graph. 35*, 6 (Nov. 2016). 2

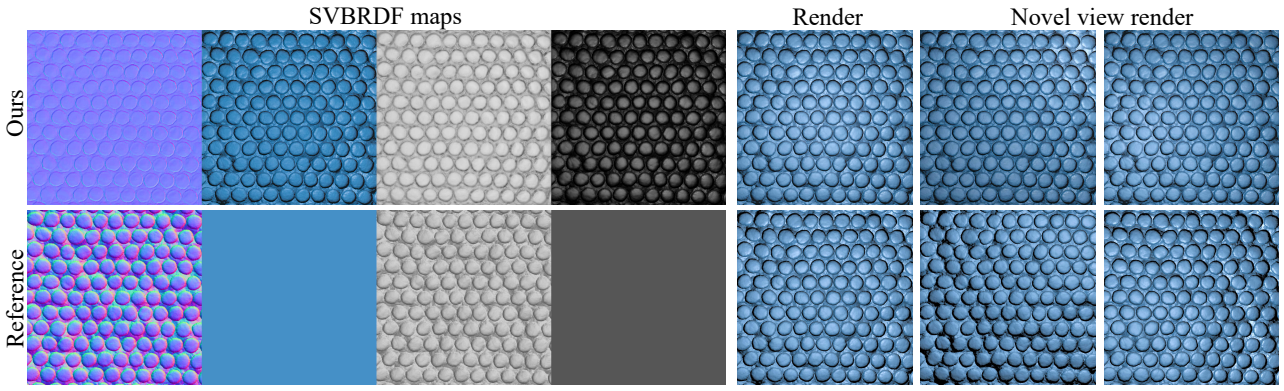

Figure 12: *Failure case. Our method generates SVBRDF maps with similar texture variation, while the reference diffuse map and specular map are both constant.*

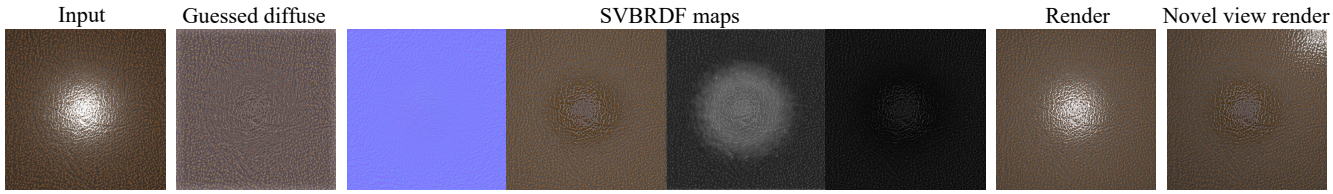

Figure 13: *Failure case. Images with sharp contrast may not be well handled since our method fails to get a stationary "guessed" diffuse map.*

[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). 2

[YLD*18] YE W., LI X., DONG Y., PEERS P., TONG X.: Single image surface appearance modeling with self-augmented cnns and inexact supervision. In *Computer Graphics Forum* (2018), vol. 37, Wiley Online Library, pp. 201–211. 3

[ZWX*20] ZHAO Y., WANG B., XU Y., ZENG Z., WANG L., HOLZSCHUCH N.: Joint svbrdf recovery and synthesis from a single image using an unsupervised generative adversarial network. In *EGSR* (2020). 2, 3, 4, 5, 6, 7, 8, 9, 10