Computational Visual Media
https://doi.org/10.1007/s41095-022-0289-1 Vol. 9, No. x, month 2022, xx–xx
Research Article

An Attention-embedded GAN for SVBRDF Recovery from a Single Image

Zeqi Shi[1], **Xiangyu Lin**[1], **and Ying Song**[1](✉)

Abstract Learning based approaches have made substantial progress in capturing spatially-varying bidirectional reflectance distribution functions (SVBRDFs) from a single image with unknown lighting and geometry. However, most existing networks only consider per-pixel losses which limit their capability to recover local features such as smooth glossy regions. A few generative adversarial networks use multiple discriminators for different parameter maps, increasing network complexity. We present a novel end-to-end GAN to recover appearance from a single picture of a nearly-flat surface lit by flash. We use a single unified adversarial framework for each parameter map. An attention module guides the network to focus on details of the maps. Furthermore, the SVBRDF map loss is combined to prevent paying excess attention to specular highlights. We demonstrate and evaluate our method on both public datasets and real data. Quantitative analysis and visual comparisons indicate that our method achieves better results than the state-of-the-art in most cases.

Keywords SVBRDF; appearance capture; generative adversarial network; attention mechanism

1 Introduction

The complex interaction between light and the surfaces of objects with various appearances explain the variations in photographs captured in the real world. Acquiring the surface reflection parameters of objects is one of the major tasks in computer vision and realistic graphics [1–5], with various applications including appearance transfer, restoration, and augmented reality. Breaking appearance down into reflectance and lighting leads to powerful image editing applications, such as material transfer [6] and illumination editing [7].

There is a trend towards data-driven approaches when capturing surface appearance, as they are more expressive than analytic models. Recently, lightweight capture processes using consumer devices and uncalibrated lighting have drawn more attention. It would benefit content creation for virtual worlds if a user could create the appearance from a few or even a single image. Since different lighting and material properties may produce the same visual effects, recovering material parameters based on a single image is an ill-posed problem.

Researchers have proposed many learning-based approaches to address the task; some focus on estimating spatially-varying surface material parameters from a single image [8–11]. The numerical pixel error in these works is low, but the visual effect is not satisfactory in certain cases. Others use multiple image inputs to capture materials [12, 13], but require tedious optimization to support an arbitrary number of input images. In this paper, inspired by the significant progress in using GANs [14] in image processing, we propose an end-to-end learning framework to reconstruct SVBRDFs from a single image.

Traditional convolutional neural networks (CNNs) for SVBRDF recovery cannot effectively express smooth highlight details in the reconstructed appearance. Another challenge is generalizability, especially when trained on synthetic datasets. Some methods [15, 16] add multiple adversarial losses to optimize individual material parameters. Considering the inherent instability in GAN training, increasing the adversarial loss inevitably increases redundancy and the difficulty of determining network parameters. We aim to obtain high-quality material parameters to efficiently reconstruct the appearance of the input, so in our framework we use a unified discriminator for all maps. Specifically, we apply a binary classification discriminant network to judge the authenticity of the generated results, and use it to optimize the latent features in the generative network, so that the generator can produce images of high perceptual quality.

Parameter estimation not only needs to consider overall

1 Zhejiang Sci-Tech University, 2011 Collaborative Innovation Center for Garment Personal Customization of Zhejiang Province, Hangzhou, 310018, China. E-mail: Z. Shi, shizeqi6@gmail.com; X. Lin, linxiangyu@zstu.edu.cn; Y. Song, ysong@zstu.edu.cn.

Manuscript received: 2022-01-20; accepted: 2022- -

quality, but also must be optimized for certain local details, which are often the main visual clues for human determination of material properties: when a person observes an image, they first scan it globally, and then pay attention to details. This focusing of attention tends to allocate more resources to the target of interest, while suppressing the less useful features. Therefore, we design an attention mechanism to improve the quality of SVBRDF maps, especially for local high-frequency details.

However, the attention mechanism may also increase error, especially in specular regions, because it focuses on highlights and cannot decompose the result of the joint contributions of multiple maps, resulting in false bright regions in the specular map. We thus add an SVBRDF-map loss to guide weight learning after adopting the attention mechanism. Our new joint loss consists of adversarial loss, rendering loss and SVBRDF map loss. We have validated that our GAN-based reconstruction framework can produce convincing results.

To summarize, our contributions include:

- a unified GAN framework for supervised high-quality SVBRDF map recovery,
- an attention mechanism specifically designed to improve visual quality resulting from the reconstructed SVBRDFs, and
- a new joint loss consisting of a weighted sum of the adversarial loss, rendering loss and SVBRDF map loss.

2 Related work

This overview mainly discusses works that allow non-expert users to employ image-driven appearance modeling tools with commodity devices and lightweight capturing processes. Some use a few or multiple images as input and fit the SVBRDF parameters using them, without any prior knowledge. Others are deep learning-based, and use large-scale datasets to train network parameters.

2.1 Multi-Image Appearance Modeling

Several works use multiple images as input to capture SVBRDFs. Chandraker et al. [17] use motion cues to jointly optimize the shape and reflectance of objects, under known lighting conditions. Hui et al. [18] capture several images of an object from a fixed viewing angle under varied lighting to estimate shape and reflectance. Riviere et al. [19] use a handheld camera or mobile phone to collect spatially varying material samples and utilize hand-crafted heuristics to estimate specular and diffuse reflectance. Xia et al. [20] estimate SVBRDFs and detailed geometric shapes from videos of rotating objects under unknown natural lighting. Although these methods can obtain accurate reflectance, they need heuristic regularization or assumptions due to the paucity of input samples. In contrast, our method can perform well even for a single input image.

2.2 Single or Few Image Appearance Modeling

Other works aim to estimate material properties by inputting one or a minimal number of images. Boivin et al. [21] use a single image and a 3D geometric model to recover surface reflectance. Their algorithm first classifies the materials in the image, and finds the optimal values of material parameters by continuous layering and iteration. Aittala et al. [22] recover reflectance from only two photos assuming that each local area is statistically similar. Xu et al. [23] obtain the BRDF of a homogeneous plane sample from two photos; the approach can be can also be simply extended to acquisition of SVBRDFs by clustering the materials. These methods usually require strict constraints and complex fitting or optimization to achieve their goals.

2.3 Learning-based Appearance Modeling

At present, most successful works are deep learning based, as this allows use of prior knowledge to help solve this ill-posed problem. Li et al. [8] obtain diffuse albedo and normal maps from nearly-planar samples under global illumination, using a self-augmentation strategy to train the model using a small training set. Deschaintre et al. [9] proposed a parallel network structure that combines a fully connected layer with a traditional U-Net [24] to extract global features, in order to reduce artifacts. They extended this work to flexibly allow a varying number of images by using an order-independent fusion layer [25]. Li et al. [26] designed a complicated cascade network that can recover shape and SVBRDFs simultaneously. They added a rendering layer to the network to estimate global illumination effects; this is essential for real-world scenes. The method proposed by Gao et al. [12] can estimate SVBRDF maps for a flat sample from any number of input photos. They train an autoencoder to build the latent space of the SVBRDF maps, and then optimize the material maps within it. Using more input images, the SVBRDF maps become more accurate, but takes more time.

Zhao et al. [27] proposed an unsupervised generative adversarial neural network that can generate high-quality SVBRDFs from a single photograph with a repetitive structure. Guo et al. [13] proposed MaterialGAN to solve the problem of SVBRDF reconstruction from multiple input images. Generally, the multi-view method requires relatively correct viewing angles and light directions to

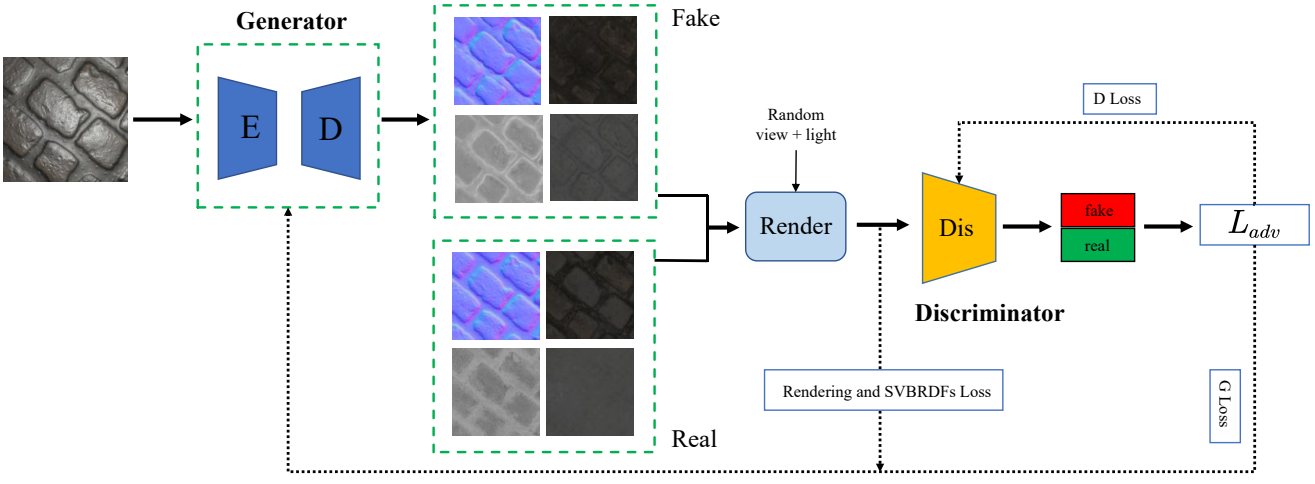


Fig. 1 Overview of our GAN architecture. The input is a single image. Solid arrows indicate the direction of data flows, while dashed arrows indicate the direction of gradient propagation.

provide high-quality output, but non-expert users cannot accurately determine these parameters, increasing difficulty and decreasing robustness of the method. Asselin et al. [28] use a new portable capture device to obtain real datasets, and estimate material maps based on the deep learning architecture of StyleGANv2 [29]. Zhou et al. [15] adopt multiple adversarial losses and add some real-world images during training to improve the quality of the reconstructed parameter maps, but the generated maps have artifacts in saturated highlight regions. Guo et al. [16] designed a two-stream neural network to obtain SVBRDF maps, which contains two independent feature extraction modules and four feature fusion modules to reduce artifacts caused by input highlights, but there are still large errors in some cases. Our solution is an end-to-end GAN architecture; an attention mechanism is embedded in the generator to keep details. A comprehensive comparison of results demonstrates the superiority of this method.

3 Method

Inspired by the progress of GANs in image processing tasks, we propose a new GAN architecture, which can generate reliable SVBRDF maps from a single image. The main structure of our GAN architecture is shown in Fig. 1. We input a picture taken by a mobile phone or camera into the generation network to give an initial result. We then input the predicted SVBRDF maps and ground truth maps into the rendering layer to randomly render multiple images and concatenate them. The discriminant network is used to distinguish between true and false, and the final difference is combined as the loss function.

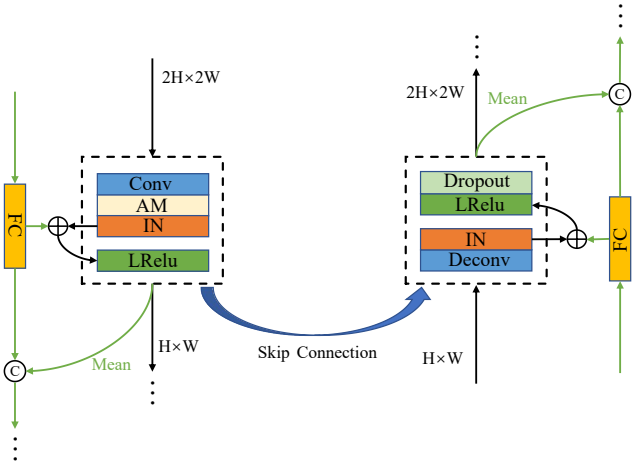


Fig. 2 Generative network architecture for one level, including down-sampling components, up-sampling components and a fully connected layer. The attention mechanism module is embedded within the down-sampling component.

3.1 Network Structure

Our generator consists of an encoder and a decoder, and finally generates a normal map, a diffuse albedo map, a roughness map, and a specular albedo map. For convenience, we denote the layers producing outputs of the same resolution as a level. Fig. 2 shows our generator architecture for one such level, which includes down-sampling, up-sampling, and a parallel fully connected layer to fuse global information. A typical down-sampling block contains 4 layers: a convolution layer, an attention layer, an InstanceNorm layer and a Leaky ReLU [30] activation layer. The attention mechanism module will be explained in Section 3.2. A typical up-sampling block contains 4 layers: a deconvolution layer, an InstanceNorm layer, a Leaky ReLU activation layer and a dropout layer.

For image generation tasks like ours, the generated result mainly depends on one image instance, so we use instance normalization instead of normalizing the entire batch. In order to increase the nonlinear relationship between the layers of the network, the choice of activation function is crucial. We use the Leaky Relu function with a weight of 0.2, as, compared to Relu, doing so can speed up convergence and effectively avoid gradient vanishing and dead neurons during training. We introduce skip connections to sample blocks of the same size to reintroduce missing high-frequency details. Deschaintre et al. [9] showed that the current task cannot be readily be solved using only a network structure similar to U-Net, as the convolution operation is usually used to extract spatially local features. The practical receptive field of the CNN is actually much smaller than the theoretical value especially at high levels, as shown by Zhou et al. [31]. Therefore, a global feature extraction module is needed to fuse far-away information. Specifically, we add a network composed of fully connected (FC) layers parallel to the U-Net as the global feature extractor, following [9]. The output of the InstanceNorm layer in each sampling block is added to the output of the FC layer in the current level and then passed to the activation layer. The output of the FC layer is further concatenated with the mean vector of the activation layer output, as the input to the FC layer in the next level. The discriminator follows Isola et al. [32].

3.2 Attention Module

Attention mechanisms [33, 34] are widely used in natural language processing, speech and image processing, etc. [35–37]. We designed an attention module considering multi-scale features (see Fig. 3) to improve both global and local quality of the reconstructed results. Intuitively speaking, 'high-level' attention tends to concentrate on highlight saturated area, while 'low-level' attention focuses on details like local high-frequency variations in SVBRDF maps. Experimental results of ablation studies of the attention module are given in Section 4.3.

X' in the attention module is given by

$$X'_{h,w,c} = \sigma(\sum_{i=1}^{N} f_i(W_c \otimes \alpha_i(X^l_{h,w,c}))) \odot X^l_{h,w,c}, \quad (1)$$

where α_i represents average pooling, W_c denotes a 1×1 convolution kernel, f_i is bilinear interpolation for up-sampling, σ is the activation function, $\otimes$ represents convolution, and $\odot$ represents element-wise product.

After adding this mechanism, detailed features are enhanced, as can be seen later from the results in Fig. 6.

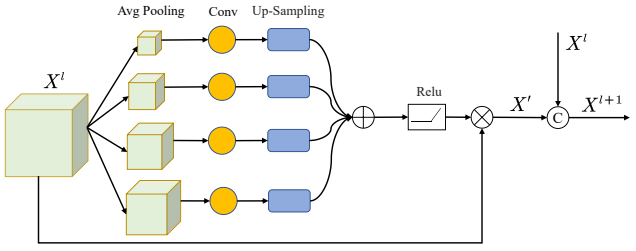


Fig. 3 Attention module. Arrows indicate directions of data flows. X^l is the input to level l. X' is the result of multiplying the output of the activation function and the input feature map. X^{l+1} is the output of the entire attention module to the next level, concatenating X^l and X'.

3.3 Loss Function

Choice of loss function is critical for generators. Deschaintre et al [9] showed that L1 loss using SVBRDF maps alone cannot recover appearances relatively consistent with the ground truth, so they used the rendering loss instead of L1 loss. Although re-rendering of the restored SVBRDF maps can produce an appearance relatively consistent with the input, there are still large errors for some reflection parameters, especially the roughness map and the specular albedo map. In order to account for both per-pixel error and consistency, we apply a joint loss function which is a weighted sum of adversarial loss, rendering loss and SVBRDF map loss:

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{\text{adv}} + \lambda_2 \mathcal{L}_{\text{render}} + \lambda_3 \mathcal{L}_{\text{svbrdf}} \quad (2)$$

The optimized objective function for the generative adversarial network is

$$\min_G \max_D V(D,G) = E_{x\sim p_{\text{data}}(x)}[\log(D(x))] + E_{z\sim p_z(z)}[\log(1-D(G(z)))], \quad (3)$$

where G is the generative network, which represents the mapping of training samples to generated data. D is the discriminant network, which discriminates the input samples and maximizes the distance between the real data and the generated data. $x \sim p_{\text{data}}(x)$ is the real data, and $z \sim p_z(z)$ is the input data. The two networks optimize the objective function through alternated iterative training.

The rendering loss

$$\mathcal{L}_{\text{render}} = \frac{1}{N}\sum_{i=1}^{N} \|\log(R_i(\mathbf{x})) - \log(R_i(\mathbf{y}))\|_1 \quad (4)$$

is an L1 loss between the rendering result of the predicted SVBRDF maps and the ground truth SVBRDF maps under several lighting and viewing directions. The logarithmic transformation aims to enhance details especially in dark regions, following [38]. N represents the number of generated images rendered in random directions, and R_i is the rendering layer in the network. The rendering layer acts as a pixel shader

that evaluates the rendering equation at each pixel of the SVBRDFs, given a pair of viewing and lighting directions. The process is performed in SVBRDF coordinate space. We use the Cook-Torrance [39] BRDF model to render the image following Aittala et al. [40]. The SVBRDF map loss is defined as:

$$\mathcal{L}_{\text{svbrdf}} = \lambda_n \mathcal{L}_n + \lambda_d \mathcal{L}_d + \lambda_r \mathcal{L}_r + \lambda_s \mathcal{L}_s, \quad (5)$$

where $\mathcal{L}_n$, $\mathcal{L}_d$, $\mathcal{L}_r$, $\mathcal{L}_s$ are the L1 loss of the normal map, diffuse albedo map, roughness map, and specular map. In our experiments, $\lambda_n = \lambda_d = 1$, $\lambda_r = \lambda_s = 0.5$.

4 Experiments

We now introduce the dataset and experimental parameters used in our experiments, and give a quantitative and qualitative evaluation of our proposed methods.

4.1 Datasets and Implementation

For evaluation we use a synthetic dataset provided by Deschaintre et al. [9]. It contains approximately 200,000 synthetic samples, including training and test samples. Each sample contains the original image, normal map, diffuse albedo map, roughness map, specular albedo map; the size of the image is 256.

We implemented our model using the Tensorflow [41] deep learning framework. Training was performed on a Tesla V100 GPU. The generator and the discriminator were trained alternately, and the discriminator was updated once after the generator was trained 5 times, on average. Training used the Adam optimizer [42], with initial learning rate 0.00002, reduced by half every two epochs. All other hyperparameters were set to the default values for Tensorflow. We set $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$ for $\mathcal{L}_{\text{adv}}$. The batch size was 8. We trained our GAN architecture for 20 epochs, which took about 7 days.

4.2 Comparison

We conducted quantitative and qualitative comparisons on the synthetic dataset to verify the validity. For further experimental results, refer to the electronic supplementary material. The test dataset is a selection from the synthetic dataset to provide ground truth, and was not used in training.

4.2.1 Synthetic Data

We chose several state-of-the-art methods [9, 12, 15, 16] for comparison. Gao et al. [12] can use an arbitrary number of input images; to be fair, we set $N = 1$. The root mean square error (RMSE) on each reflectance map was calculated. Merely calculating the numerical error of the material map does not suffice to show the superiority of our method, so the RMSE between the re-rendered image and the original image under 6 random lighting and viewing directions was also calculated. Average results for all test sets are given in Tab. 1. To further demonstrate the superiority of our method, we also considered two more advanced evaluation metrics, SSIM and LPIPS, with results shown in Tabs. 2 and 3. Our method achieves better results on several error metrics.

Table 1 Comparison to various baseline methods. RMSE for each map is given; the maps were rendered using 6 random lighting and viewing directions. The best scores are highlighted in bold.

Method	Normal	Diffuse	Roughness	Specular	Render
RADN	**0.061**	0.018	0.118	0.036	0.052
DIR	0.064	0.021	0.106	0.031	0.057
ASSE	0.066	0.045	0.111	0.040	0.077
HATN	0.066	0.021	0.109	0.030	0.056
Ours	0.062	**0.017**	**0.103**	**0.028**	**0.051**

Table 2 Comparison using SSIM.

Method	Normal	Diffuse	Roughness	Specular	Render
RADN	**0.791**	0.916	0.701	0.864	0.862
DIR	0.763	0.872	0.699	0.897	0.848
ASSE	0.758	0.804	0.689	0.894	0.760
HATN	0.774	0.885	0.704	0.932	0.842
Ours	0.784	**0.921**	**0.709**	**0.944**	**0.872**

To make a qualitative comparison, we randomly selected some synthetic data, and provide visual results for each algorithm in Fig. 4. When the method of Deschaintre et al. [9] processes input images with strong highlights, the output material maps have noticeable artifacts. Gao et al. [12] need a reconstruction method to give the initial input material maps; we used the material maps provided by Deschaintre et al. [9]. If these initial maps are not close to the ground truth, the optimization result is likely to fall into a local minimum. Furthermore, their method is not designed for a single input only. Zhou et al. [15] suffer from artifacts or color distortion when dealing with areas with significant highlight saturation, as shown in the upper row of Fig. 4. Guo et al. [16] can effectively suppress artifacts, but there are still noticeable visual errors in some cases, as shown in the left column of Fig. 4. Our results are better than previous methods in terms of overall visual appearance and some local details, due to our joint loss and attention mechanism, which help to restore global and local consistency of feature details.

4.2.2 Real Data

We compared our results to those of other methods [9, 12, 15, 16] using the collected real samples [9, 43]. Pictures taken by cameras or mobile phones were used as inputs, and the output SVBRDF maps were re-rendered under the same lighting conditions; final results are displayed in Fig. 5. As can be

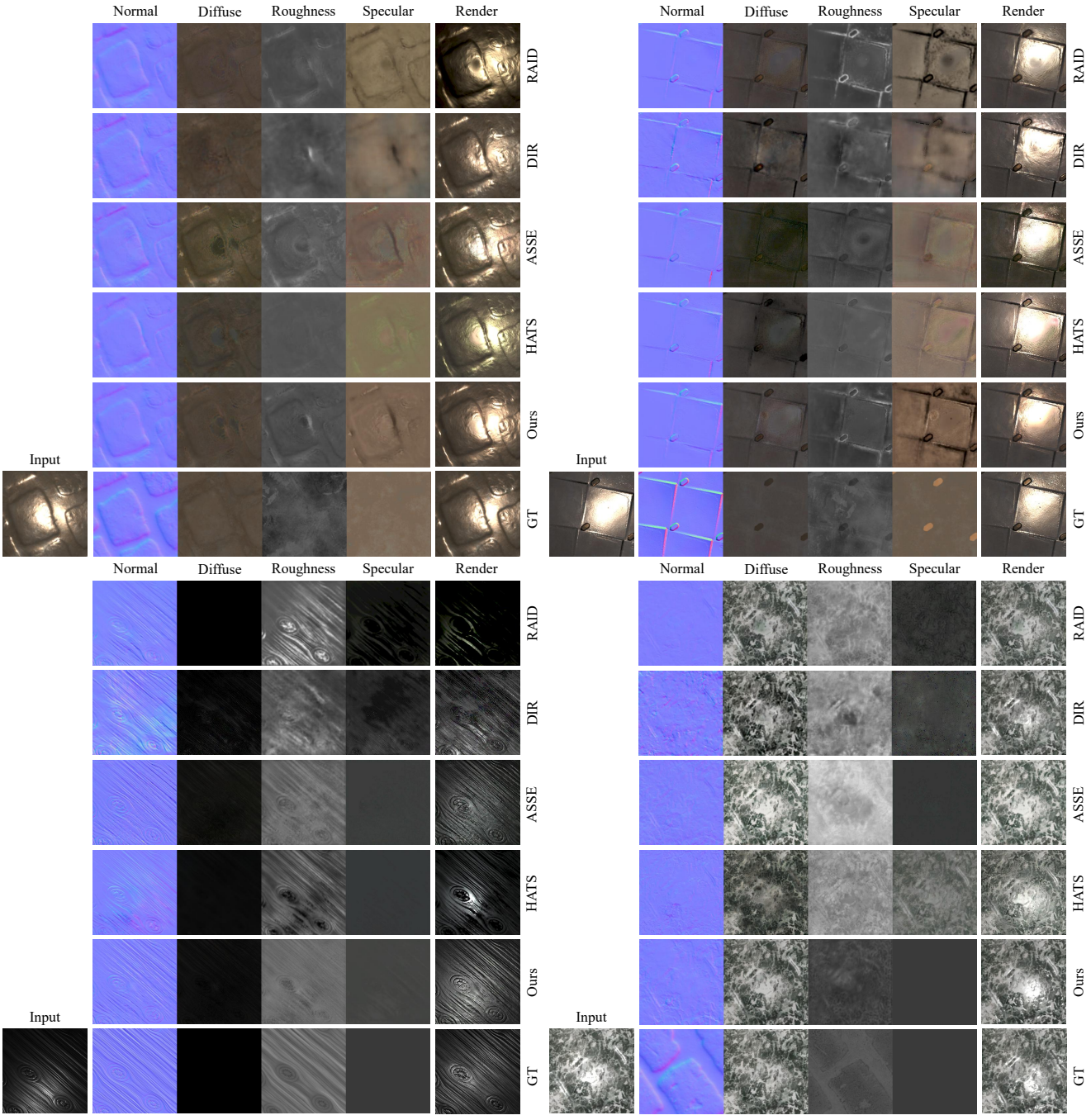


Fig. 4 Comparison to RADN of Deschaintre et al. [9], DIR of Gao et al. [12], ASSE of Zhou et al. [15] and HATS of Guo et al. [16], using synthetic data. The parameters from our method and its results after re-rendering are closer to the ground truth.

Table 3 Quantitative comparison using LPIPS.

Method	Normal	Diffuse	Roughness	Specular	Render
RADN	0.284	0.127	0.442	0.349	0.290
DIR	0.286	0.149	0.476	0.379	0.298
ASSE	**0.267**	0.254	0.414	0.271	0.274
HATN	0.274	0.128	0.387	**0.257**	0.271
Ours	0.278	**0.124**	**0.384**	0.341	**0.268**

seen, results of our method after re-rendering are closer to the real input.

4.3 Ablation Studies

We performed a set of ablation experiments to verify the contribution of each component of our method, and compared our GAN architecture to ablated versions. The variants considered were:

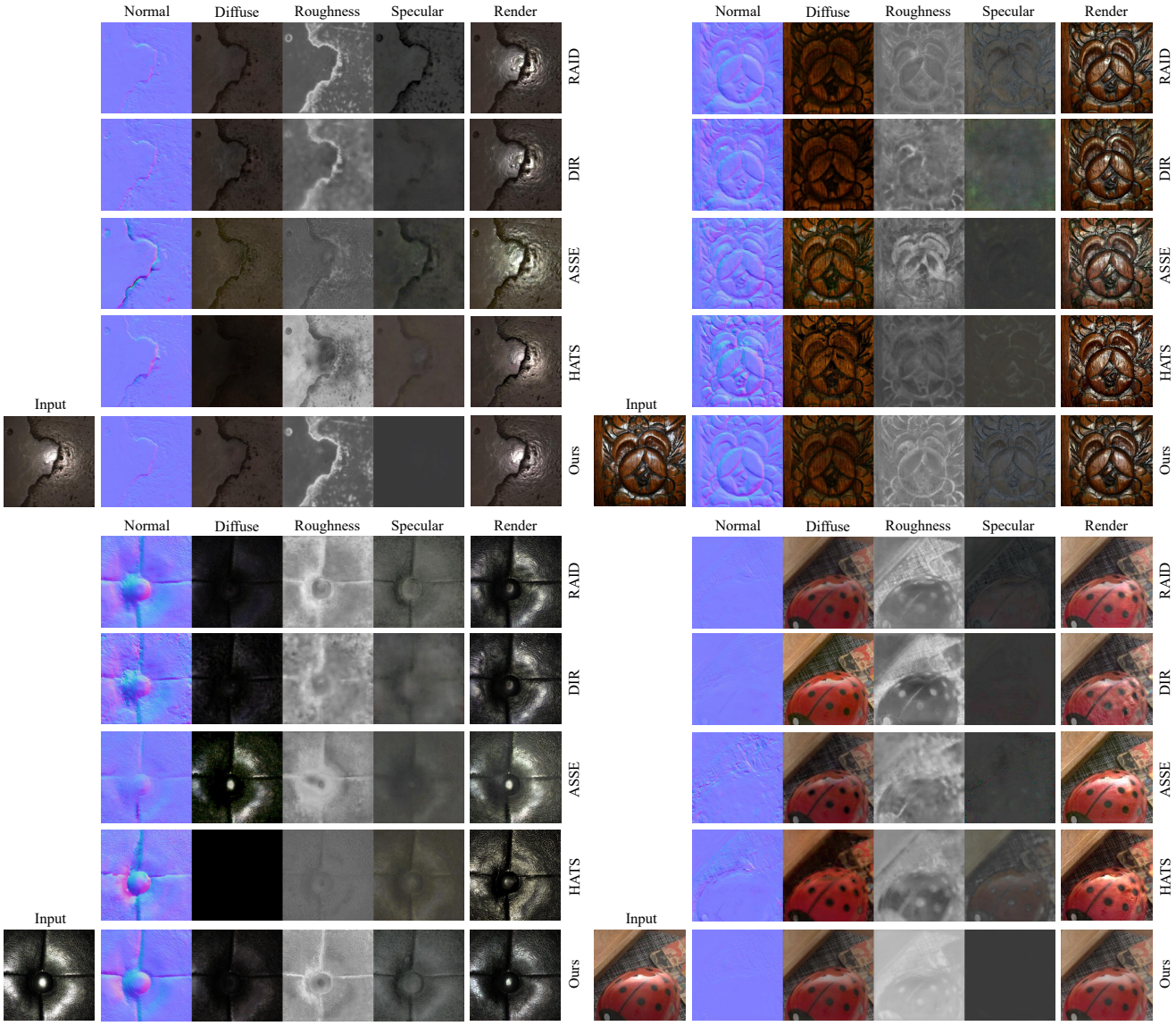


Fig. 5 Comparison to RADN of Deschaintre et al. [9], DIR of Gao et al. [12], ASSE of Zhou et al. [15] and HATS of Guo et al. [16] using a single real image as input. All re-rendered results were generated with the same lighting conditions and viewing direction. It can be seen that our results are closest to the input pictures, effectively reconstructing the SVBRDF maps.

- without $\mathcal{L}_{\text{adv}}$. To analyze the importance of the adversarial loss in restoring material maps, we removed the discriminator network and set $\lambda_1 = 0$.
- without AM. To verify that the attention mechanism module can effectively enhance details, we removed the module.
- without $\mathcal{L}_{\text{svbrdf}}$. To verify that, if SVBRDF map loss is not used ($\lambda_3 = 0$), there will be incorrect bright regions in the specular map.
- MultiGAN. To demonstrate that our unified GAN framework can produce better results, multiple adversarial loss experiments were conducted.

Deschaintre et al. [9] have already demonstrated the importance of rendering loss.

Table 4 RMSE evaluation of ablation studies; w/o means without.

Method	Normal	Diffuse	Roughness	Specular	Render
w/o $\mathcal{L}_{\text{adv}}$	0.067	0.018	0.109	0.029	0.055
w/o AM	0.063	0.018	0.112	0.038	0.052
w/o $\mathcal{L}_{\text{svbrdf}}$	0.063	0.017	0.110	0.050	0.054
MultiGAN	0.065	0.018	0.149	0.036	0.059
Ours	**0.062**	**0.017**	**0.103**	**0.028**	**0.051**

We trained the above models on the same dataset under the same training conditions as before. Quantitative results of the ablation studies are shown in Tab. 4.

The visual comparison in Fig. 6 further shows that our approach provides the best results. Fig. 6(a) shows that when

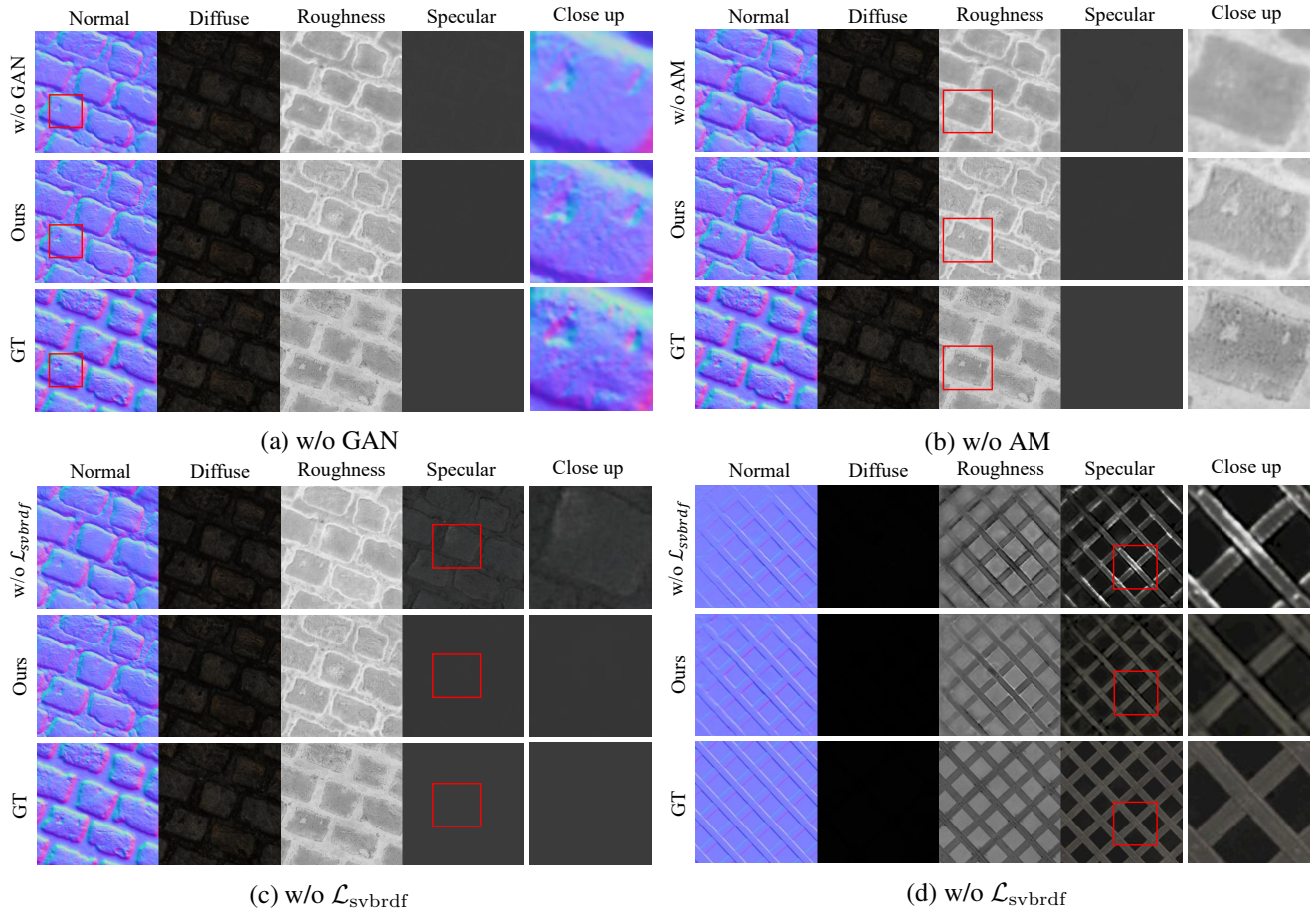


Fig. 6 Examples of qualitative displays of ablation experiments. The difference has been shown in red and zoomed in.

the adversarial loss is removed, the overall error increases, especially in the normal map: the adversarial loss has a significant impact on the quality of the generator. In Fig. 6(b), there is higher error in the roughness map: it lacks details after removing the attention module. Simply using the rendering loss to update the network parameters, without $\mathcal{L}_{\text{svbrdf}}$, leads to weight reduction for the specular albedo map, resulting in higher errors, as shown in Fig. 6(c). If SVBRDF loss is omitted, if a picture with specular highlights is input, the network will pay more attention to it due to the existence of the attention mechanism, but cannot decompose the maps, resulting in the false highlight in the specular albedo map, as shown in Fig. 6d. When adding multiple adversarial losses for training as in MultiGAN, the recovery of SVBRDF maps is not significantly improved. This is due to the inherent difficulty in achieving the Nash equilibrium between the generator and multiple discriminators, so it is difficult and time-consuming to obtain thane optimal solution.

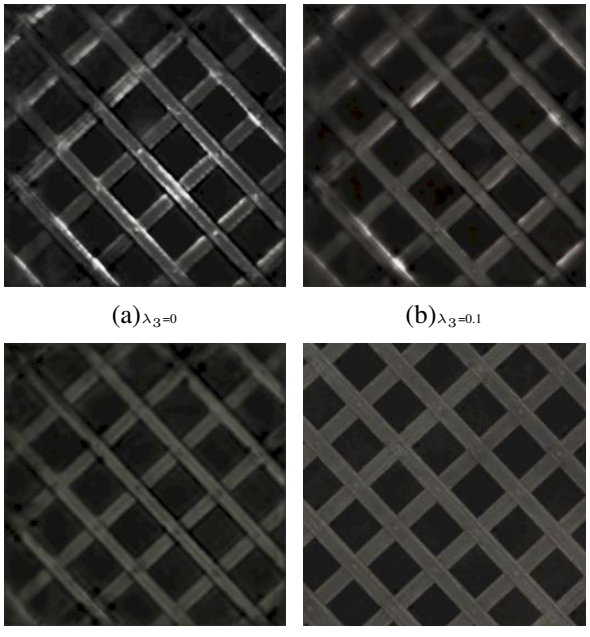
In order to further study the role of SVBRDF map loss in the training process when the input image has specular highlights, we analyzed its effect by changing the value of λ_3, as shown in Fig. 7. As λ_3 increases, the false highlights in the specular albedo map are eliminated, and the learned results become closer to the ground truth.

5 Conclusions

This paper presents a novel solution based on a GAN to recover SVBRDF maps from a single image. It can generate more accurate material maps and re-rendered appearance is more realistic. However, our method also has limitations.

Although we achieve reasonable results for input images containing specular highlights, when the highlights are too large, low-saturation pixels will dominate the entire image. In this case, the network cannot learn enough features to generate plausible SVBRDF maps, resulting in color distortion around highlights in the re-rendered image. One possible way to improve this is to use multiple inputs from different lighting and viewing directions. A more flexible network structure should also be designed to support multiple inputs.

Currently, a synthetic dataset is used for training, then the model is applied to real data to generate material maps, because labelled real datasets are scarce and difficult to



(a)$_{\lambda_3=0}$ (b)$_{\lambda_3=0.1}$ (c)$_{\lambda_3=0.5}$ (d)$_{\text{Ground Truth}}$

Fig. 7 Effect of λ_3 on results.

acquire. Therefore, another potential direction for future work is to find a way to utilize real datasets in training.

Acknowledgements

The authors would like to thank Jie Guo from Nanjing University for his kind help with the comparison. Ying Song was partially supported by the National Natural Science Foundation of China (No. 61602416), and Shaoxing Science and Technology Plan Project (No. 2020B41006).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Electronic Supplementary Material

Supplementary material including more experimental results and a video of re-rendering results is available in the online version of this article.

References

[1] Weyrich T, Lawrence J, Lensch HP, Rusinkiewicz S, Zickler T. Principles of appearance acquisition and representation. *Foundations and Trends in Computer Graphics and Vision*, 2009, 4(2): 75–191.

[2] Dorsey J, Rushmeier H, Sillion F. *Digital modeling of material appearance*. 2010.

[3] Weinmann M, Klein R. Advances in geometry and reflectance acquisition (course notes). In *SIGGRAPH Asia 2015 Courses*, 2015, 1–71.

[4] Guarnera D, Guarnera GC, Ghosh A, Denk C, Glencross M. BRDF representation and acquisition. In *Computer Graphics Forum*, 2016, 625–650.

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

[6] Deschaintre V, Drettakis G, Bousseau A. Guided Fine-Tuning for Large-Scale Material Transfer. In *Computer Graphics Forum*, 2020, 91–105.

[7] Yang S, Yanli L. An algorithm generating human face cartoon portrait including light editing based on photo. *Journal of Graphics*, 2015, 36(1): 83–89.

[8] 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)*, 2017, 36(4): 1–11.

[9] 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)*, 2018, 37(4): 1–15.

[10] 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, 72–87.

[11] 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, 201–211.

[12] 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)*, 2019, 38(4): 1–15.

[13] Guo Y, Smith C, Hašan M, Sunkavalli K, Zhao S. MaterialGAN: reflectance capture using a generative SVBRDF model. *arXiv preprint arXiv:2010.00114*, 2020.

[14] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. *Advances in neural information processing systems*, 2014, 27.

[15] Zhou X, Kalantari NK. Adversarial Single-Image SVBRDF Estimation with Hybrid Training. In *Computer Graphics Forum*, 2021, 315–325.

[16] Guo J, Lai S, Tao C, Cai Y, Wang L, Guo Y, Yan LQ. Highlight-aware two-stream network for single-image SVBRDF acquisition. *ACM Transactions on Graphics (TOG)*, 2021, 40(4): 1–14.

[17] Chandraker M. On shape and material recovery from motion. In *European Conference on Computer Vision*, 2014, 202–217.

[18] Hui Z, Sankaranarayanan AC. A dictionary-based approach for estimating shape and spatially-varying reflectance. In *2015 IEEE International Conference on Computational Photography (ICCP)*, 2015, 1–9.

[19] Riviere J, Peers P, Ghosh A. Mobile surface reflectometry. In *Computer Graphics Forum*, 2016, 191–202.

[20] Xia R, Dong Y, Peers P, Tong X. Recovering shape and spatially-varying surface reflectance under unknown illumi-

nation. *ACM Transactions on Graphics (TOG)*, 2016, 35(6): 1–12.

[21] Boivin S, Gagalowicz A. Inverse rendering from a single image. In *Conference on Colour in Graphics, Imaging, and Vision*, 2002, 268–277.

[22] Aittala M, Weyrich T, Lehtinen J, et al.. Two-shot SVBRDF capture for stationary materials. *ACM Transactions on Graphics (TOG)*, 2015, 34(4): 110–1.

[23] Xu Z, Nielsen JB, Yu J, Jensen HW, Ramamoorthi R. Minimal brdf sampling for two-shot near-field reflectance acquisition. *ACM Transactions on Graphics (TOG)*, 2016, 35(6): 1–12.

[24] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015, 234–241.

[25] Deschaintre V, Aittala M, Durand F, Drettakis G, Bousseau A. Flexible svbrdf capture with a multi-image deep network. In *Computer Graphics Forum*, 2019, 1–13.

[26] 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)*, 2018, 37(6): 1–11.

[27] 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 (DL)*, 2020, 53–66.

[28] Asselin LP, Laurendeau D, Lalonde JF. Deep SVBRDF estimation on real materials. In *2020 International Conference on 3D Vision (3DV)*, 2020, 1157–1166.

[29] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 8110–8119.

[30] Maas AL, Hannun AY, Ng AY, et al.. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the International Conference on Machine Learning*, 2013, 3–8.

[31] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[32] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 1125–1134.

[33] Nguyen TV, Zhao Q, Yan S. Attentive systems: A survey. *International Journal of Computer Vision*, 2018, 126(1): 86–110.

[34] Chaudhari S, Mithal V, Polatkan G, Ramanath R. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2021, 12(5): 1–32.

[35] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[36] Chorowski J, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*, 2015.

[37] Mnih V, Heess N, Graves A, et al.. Recurrent models of visual attention. In *Advances in neural information processing systems*, 2014, 2204–2212.

[38] Maini R, Aggarwal H. A comprehensive review of image enhancement techniques. *arXiv preprint arXiv:1003.4053*, 2010.

[39] Cook RL, Torrance KE. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1982, 1(1): 7–24.

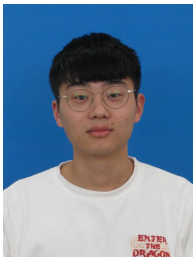
[40] Aittala M, Aila T, Lehtinen J. Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (ToG)*, 2016, 35(4): 1–13.

[41] Team T. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. *URL http://tensorflow. org/. Software available from tensorflow. org*, 2015.

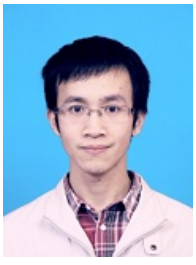
[42] Kingma DP, Ba J. Adam: A method for stochastic optimization. ICLR 2015. *arXiv preprint arXiv:1412.6980*, 2015, 9.

[43] Fu G, Zhang Q, Zhu L, Li P, Xiao C. A multi-task network for joint specular highlight detection and removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 7752–7761.

Author biographies



Zeqi Shi is a master's student at the School of Information Science and Technology of Zhejiang Sci-Tech University. He received his B.S. from Zhejiang Sci-Tech University in 2019. His research interests include deep learning and computer graphics.



Xiangyu Lin is a lecturer in the School of Information Science and Technology of Zhejiang Sci-Tech University. He obtained his B.S. and Ph.D. degrees in electronic information technology and instruments from Zhejiang University. His main research interests are image processing and machine learning.

Ying Song is an associate professor in the School of Information Science and Technology of Zhejiang Sci-Tech University. She obtained her B.S. and Ph.D. degrees in computer science and technology from Zhejiang University. Her main research interests are appearance modeling and realistic rendering.