Proceedings Paper:
Xue, Bowen, Guarnera, Claudio, Zhao, Shuang et al. (1 more author) (2024) ReflectanceFusion:Diffusion-based text to SVBRDF Generation. In: Eurographics Symposium on Rendering. Rendering 2024 - Symposium Track, 03-05 Jul 2024, Imperial College London. Eurographics Association , GBR

https://doi.org/10.2312/sr.20241152

Eurographics Symposium on Rendering (2024)
E Garces and E. Haines (Editors)

ReflectanceFusion: Diffusion-based text to SVBRDF Generation

Bowen Xue[1], Giuseppe Claudio Guarnera[2], Shuang Zhao[3], Zahra Montazeri[1]

[1]University of Manchester, UK , [2]University of York, UK, [3]University of California, Irvine, USA

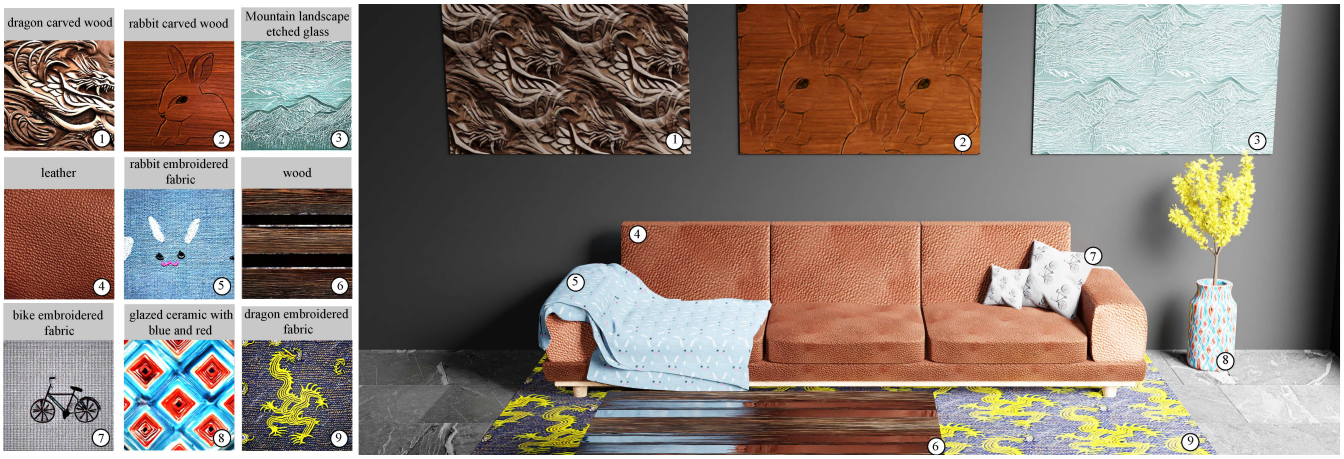


Figure 1: *Given a textual description of the desired material appearance, our model generates plausible Spatially-Varying Bidirectional Reflectance Distribution Function (SVBRDF) parameter maps. These maps can accurately represent a wide range of materials. Unlike static images, SVBRDFs enable relighting, and editing and can be applied on any object surface. Here we show rendered images under both side and environment lighting to demonstrate the adaptability of our model to diverse lighting configurations.*

Abstract

We introduce ReflectanceFusion (Reflectance Diffusion), a new neural text-to-texture model capable of generating high-fidelity SVBRDF maps from textual descriptions. Our method leverages a tandem neural approach, consisting of two modules, to accurately model the distribution of spatially varying reflectance as described by text prompts. Initially, we employ a pre-trained stable diffusion 2 model to generate a latent representation that informs the overall shape of the material and serves as our backbone model. Then, our ReflectanceUNet enables fine-tuning control over the material's physical appearance and generates SVBRDF maps. ReflectanceUNet module is trained on an extensive dataset comprising approximately 200,000 synthetic spatially varying materials. Our generative SVBRDF diffusion model allows for the synthesis of multiple SVBRDF estimates from a single textual input, offering users the possibility to choose the output that best aligns with their requirements. We illustrate our method's versatility by generating SVBRDF maps from a range of textual descriptions, both specific and broad. Our ReflectanceUNet model can integrate optional physical parameters, such as roughness and specularity, enhancing customization. When the backbone module is fixed, the ReflectanceUNet module refines the material, allowing direct edits to its physical attributes. Comparative evaluations demonstrate that ReflectanceFusion achieves better accuracy than existing text-to-material models, such as Text2Mat, while also providing the benefits of editable and relightable SVBRDF maps.

CCS Concepts

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

1. Introduction

Reproducing the visual appearance of material reflectance is challenging—especially for materials with detailed spatially varying variations. Conventionally, authoring spatially varying reflectance requires the use of professional software (e.g., Substance) and can be highly time-consuming. The most promising recent so-

lutions leverage machine learning to generate images corresponding to target text prompts. Existing generative models primarily produce static images under a specific type of lighting, lacking the editability of materials such as SVBRDF maps. Consequently, the rendered images have to be used as-is, with no control over the material.

To tackle this limitation, we formulate SVBRDF estimation as a diffusion task conditioned on an input text description. Our objective is to develop a model that facilitates text-to-SVBRDF conversion, substantially simplifying the complexity of SVBRDF design. Existing diffusion-based models rely on pre-trained, large-scale image diffusion models to sample the distribution of natural images. However, the distribution of SVBRDFs differs significantly from natural images. Therefore, we introduce a novel generative diffusion model tailored for spatially varying materials. We introduce an initial diffusion model to produce an overall appearance and synthesize the input text into a latent space. The output is then passed to a second phase, named ReflectanceUNet, that estimates SVBRDF parameter maps consisting of 10 channels (i.e. normals, diffuse and specular albedo, and specular roughness). Our tandem approach allows us to first generate a rough estimate of the appearance and then refine it in the second stage, governed by physical parameters.

Furthermore, training diffusion models typically require a significantly larger training set than conventional neural networks. We use the first phase as pre-trained Standard Diffusion 2, serving as a backbone model. To train our second phase, the SVBRDF diffusion model, we supplement the INRIA synthetic SVBRDF dataset [DAD*18] with the UBO dataset [MK20], amounting to about 200,000 unique training exemplars. We employ Euler as our scheduler and v-prediction for our prediction type, meaning our diffusion model outputs predicted velocity. The loss function used is the root mean square error function. After obtaining the latent representation from the first phase, it goes through Variational Autoencoder (VAE) and Visual Geometry Group (VGG), added together as a connecting module, and fed into the second phase which is a diffusion module, then outputs SVBRDF maps.

Concretely, our contributions include:

- Text-to-Texture Pipeline: We have developed a new pipeline that translates textual descriptions into detailed SVBRDF maps, enabling precise and customizable representations of material appearances.
- ReflectanceUNet: Our new diffusion-based network generates complex reflectance properties.
- Dual-Phase Architectural Strategy: Our method is a two-phase approach, initiating with a backbone network for foundational texture generation, subsequently refined by our ReflectanceUNet for superior detail and accuracy.

2. Related Works

2.1. Material Properties Acquisition and Generation

The SVBRDF [Nic65] is a surface reflectance model that characterizes how light reflects off non-homogeneous, opaque surfaces, capturing spatially varying properties such as shininess and texture at different points on the material. Traditionally, acquiring SVBRDF involves documenting the variations in a material's appearance under various lighting and viewing conditions through extensive photographic processes [GGG*16]. Such image-based approach necessitates high-quality imaging equipment and controlled lighting setups, capturing images from multiple angles and under varied lighting conditions to fully represent the material's reflective behavior [MK20; MXZ*23]. In traditional methods, acquiring material properties necessitates expensive and complex professional hardware, such as camera domes and computer-controlled robots, often making it unaffordable for individual artists and forcing them to rely on existing material libraries.

In the context of material property acquisition, the Bidirectional Texture Function (BTF) [DvGNK99] merits discussion. Like SVBRDF, BTF captures the texture's appearance under varying lighting and viewing conditions but is tailored for different types of surfaces, as it accounts for effects arising from small-scale geometry, such as inter-reflection and self-shadowing. BTF datasets, such as the UBO dataset [MK14], require specialized equipment and extended collection time, demanding significant storage and retrieval resources. While some studies [XZJM23] [KMX*21] have successfully compressed BTF using neural networks, achieving good results and speed, they lack generalizability as each material requires training a separate model. Moreover, many methods employ analytical and procedural appearance models [MGZJ20; ZMA*23], suitable for complex materials like fabric. Yet, these models are costly and necessitate detailed implementations specific to material types.

With the advent of neural approaches, predicting SVBRDF from images has been significantly simplified, reducing the number of required images to sometimes just a few, a pair, or even a single one, streamlining material acquisition. Recent works[GLT*21; ZK21; GSH*20; VPS21; KCW*18; VMR*23] utilized Generative Adversarial Networks (GANs) for effective image-to-material property transformation. Despite these advancements, accurately replicating materials with complex textures remains a challenge. Hu et al. [HGH*23] uses a CLIP-based encoder to generate high-quality, procedural material node graphs from text or image prompts. Matfuse and Controlmat [VSPS23; VMR*23] explores the application of diffusion processes for controlled texture generation in materials. Sartor et al. [SP23] demonstrate new possibilities by generating SVBRDF from images using diffusion models. Although methods based on procedural material synthesis tools, such as [HDR19; SLH*20], can produce high-quality material maps, they lack scalability. Text2Mat [GZT*23] facilitates the creation of materials from the text but is limited to simpler materials and cannot generate materials with complex meanings.

2.2. Text to Image Generation

Variational AutoEncoders (VAE) [KW13] and GAN [GPM*20] have been used for image generation tasks, but Diffusion models [DN21; HJA20] stand out by producing high-fidelity and diverse images through a multi-step denoising process. In contrast, VAEs and GANs typically generate images in a single step. VAEs often yield blurry images due to overlapping latent code distributions and the averaging effect in pixel reconstruction, while GANs face challenges like mode collapse and difficulties in training convergence.

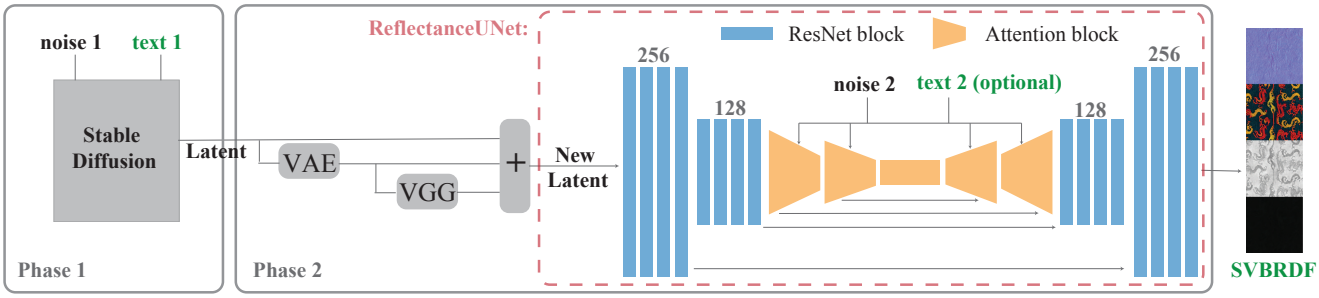


Figure 2: *Our dual-phase architecture. The system consists of two parts: Stable Diffusion 2.0 and a custom UNET. The former processes Gaussian noise, text, and time (t), while the latter, a diffusion network, outputs ten-channel SVBRDF. Inputs to the second part include the first's output, VAE, and the initial layers of VGG network contributions, material properties (specularity, roughness), and time step (t).*

Diffusion models represent a new generation of generative models that add and then remove noise from images in stages. Unlike traditional networks like VAE and GAN, they operate in lower-dimensional latent spaces to reduce computational load. Latent Diffusion [RBL*22] employs cross-attention [HZL*17] for multimodal data processing, enhancing model flexibility. Stable Diffusion, built on Latent Diffusion, is optimized for text-to-image tasks and improves generalization through extensive training. It comprises a VAE encoder compression module that encodes images into latent space and a conditional generation module that denoises and diffuses in this space to produce images matching text conditions.

Stable Diffusion XL (SDXL) [PEL*23] is a deep learning model based on the latent diffusion architecture, consisting of two main parts: the Base model and the Refiner model. The Base model handles the overall composition of the image. In contrast, the Refiner model is dedicated to adding finer visual details to further enhance image quality, representing a two-stage diffusion process where each part has a distinct function.

The recent models, Stable Diffusion 3 [EKB*24] and Sora (Video generation models as world simulators) [BPH*24], utilize the same underlying architecture, MMDiT (Multimodal Diffusion Transformer), which employs separate sets of weights for image and language representations. The MMDiT architecture uses two independent weight sets for text and image modalities, merging the sequences of both modalities in the attention mechanism. This allows each representation to operate in its own space while considering the other. Stable Diffusion achieved improved results over its predecessors, while Sora demonstrated remarkable text-to-video generation capabilities.

3. Our Method

3.1. Architectural Design

Through experimentation, we found that employing a single diffusion model does not simultaneously maintain the original diffusion model's generalization capabilities and meet the high-quality SVBRDF generation requirements. Given the relatively limited SVBRDF dataset (~200,000 samples), to ensure the model's generality, we aimed to maximally preserve the versatility of Stable Diffusion [RBL*22]. The pipeline design was inspired by Stable Diffusion XL [PEL*23] and other works that use diffusion for generating images and textures. We designed a serial network architecture composed of two diffusion modules, as shown in Fig 2, with the first part being Stable Diffusion 2.0 and the second part our custom-designed U-NET followed by a connecting module. Due to the lack of a sufficiently large dataset to train the SVBRDF VAE structure, we opted for Diffusion instead of latent Diffusion, a network that directly outputs results without an encoder and decoder. This is because latent diffusion operates on denoising in the latent space, requiring a corresponding VAE network to extract the SVBRDF's latent features. However, the limited size of SVBRDF datasets does not provide enough data to effectively train a VAE architecture. In contrast, direct diffusion models denoise in the noise space rather than the latent space, alleviating the issue of having an insufficient amount of data for training. This design choice allows us to maximally retain the original Stable Diffusion's generalization capability. Adopting a serial pipeline with two Diffusion modules instead of one retains the diffusion model's ability to represent appearances. Our dual-phase sequential structure maintains the appearance generation capacity of Stable Diffusion while integrating a second Diffusion module focused on physical parameters, ensuring excellent generalizability and the production of realistic, controllable SVBRDFs.

The first part of the network employs Stable Diffusion 2.0 [RBL*22], primarily responsible for denoising the latent appearance aspect. It takes Gaussian noise, textual description, and time step (t) as inputs. The output remains in its latent form and is not transformed back into pixel space through a decoding process. The second phase is a diffusion network designed followed by a connecting module to output ten channels, focusing on physical parameters and SVBRDF maps. The output of the first diffusion process, augmented with values from the VAE and the lower initial of VGG network [SZ14], along with material properties such as specularity and roughness, and the time step (t), is fed into the second diffusion model. We use an unmodified VAE decoder from Stable Diffusion 2.0 here. The material properties such as roughness and specular could be used as input of the model to control the final SVBRDF. This phase generates a ten-channel output, comprising normals, diffuse albedo, and specular reflection (each with three channels), and one channel for roughness.

3.2. Our ReflectanceUNet

Our network design draws on the structures proposed in previous works [DN21; RBL*22; SP23], as depicted in Fig 2. Our U-Net follows an encoder-decoder architecture, with both the input and output resolutions set to 256. Our network is trained on a resolution of 256x256, primarily constrained by the resolution of the dataset, which is 288x288. While it is feasible for our model to support higher resolutions, doing so would require a dataset with greater resolution and lead to long training time. The output features ten channels, representing normal map, diffuse albedo, roughness, and specular components. To reduce computational demands, the network blocks sized 256 and 128 only contain ResNet blocks; from the 64-sized block onwards, all blocks incorporate attention mechanisms. Each trapezoidal-shaped attention block consists of two ResNet layers and two spatial transformers, ending with a convolutional layer to adjust the output size. The sequence is ResNet layer-spatial transformer-ResNet layer-spatial transformer. Therefore, the input for the leftmost trapezoidal attention block is sized 64, and the output is sized 32, with the trapezoidal shape indicating that the output layer size is halved. The central square-shaped attention layers consist of multiple sequences of ResNet layers and spatial transformers in an alternating pattern. Each attention block has Skip Connections present between the U-Net's encoder and decoder.

We have opted for Euler as our scheduler and v-prediction [SH22] for our prediction type. V-prediction predicts velocity instead of noise. Let v be the velocity vector, which can be estimated from the model-learned parameters x and noise ϵ to estimate: $V = \alpha_t \epsilon - \sigma_t x$. V-prediction helps decrease the variance in generated samples, mitigating issues with limited training data. Additionally, it simplifies the training of diffusion models, resulting in a more stable and interpretable training process. Parameters for roughness and specular are flexible, allowing for inclusion or exclusion as needed. The optimization objective function is defined as follows:

$$\mathbb{E}_{t,x\sim p_{\text{data}}(x),\epsilon\sim\mathcal{N}(0,I)} \left\| \tilde{v}_\theta(t, \mathbf{z}_t, c) - \mathbf{v} \right\|_2^2 \tag{1}$$

where $\tilde{v}_\theta$ is the output of the diffusion model.

4. Implementation

4.1. Dataset and Training

To train our network, we utilized the SVBRDF dataset introduced by INRIA synthetic SVBRDF dataset [DAD*18] and UBO dataset [MK20], which comprises approximately 200,000 SVBRDFs alongside their corresponding rendered images. These datasets encapsulate a wide array of material types such as leather, fabric, and stone, with the specular maps employing Schlick's approximation for reflection modeling. The grey blocks shown in Fig.2 are used as pre-trained while the rest are fully trained using our dataset. During the training of this network, it initially extracts rendered images from the dataset. These images are then processed separately through a VAE network and the initial layers of a VGG network to extract features. It is worth noting that we did not utilize the entire VGG network; instead, we employed only the lower layers of VGG for feature extraction. Subsequently, the obtained latent representations are reshaped and used as inputs for the network. We stack them with spaces in between and reshape some of the latent while filling the gaps with zeros. Throughout the training, noise is incrementally introduced into the SVBRDF maps until it morphs into Gaussian noise. Moreover, the training dynamically assesses the necessity of inputting roughness and specular parameters into the network; if required, these parameters are calculated based on the specular and roughness maps and then fed as inputs. We employed a standardized input format of "flat texture of + text". We omitted the repeated phrase "flat texture of" in all results for brevity. Our training process was conducted on four A100 GPUs, utilizing FP16 precision, with each training cycle lasting approximately ten days.

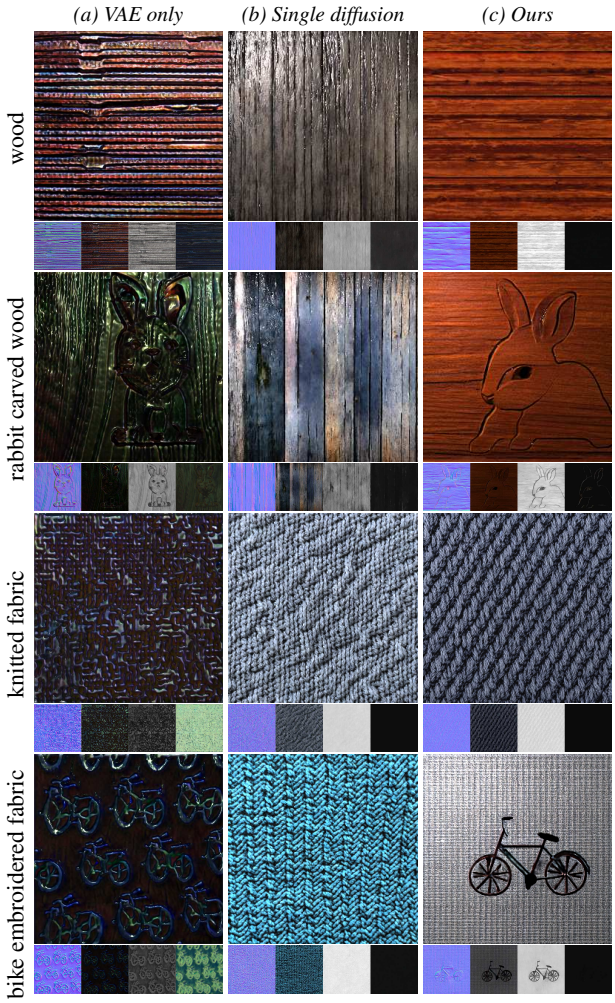


Figure 3: ***Ablation Study.*** *Comparison of the current method with two prior approaches described in the Experiments section: (a) Training only the VAE decoder to output SVBRDF maps results in invalid maps despite understanding text beyond the training set. (b) A retrained single diffusion model yields slightly better-quality maps but fails to comprehend text outside the training set.*

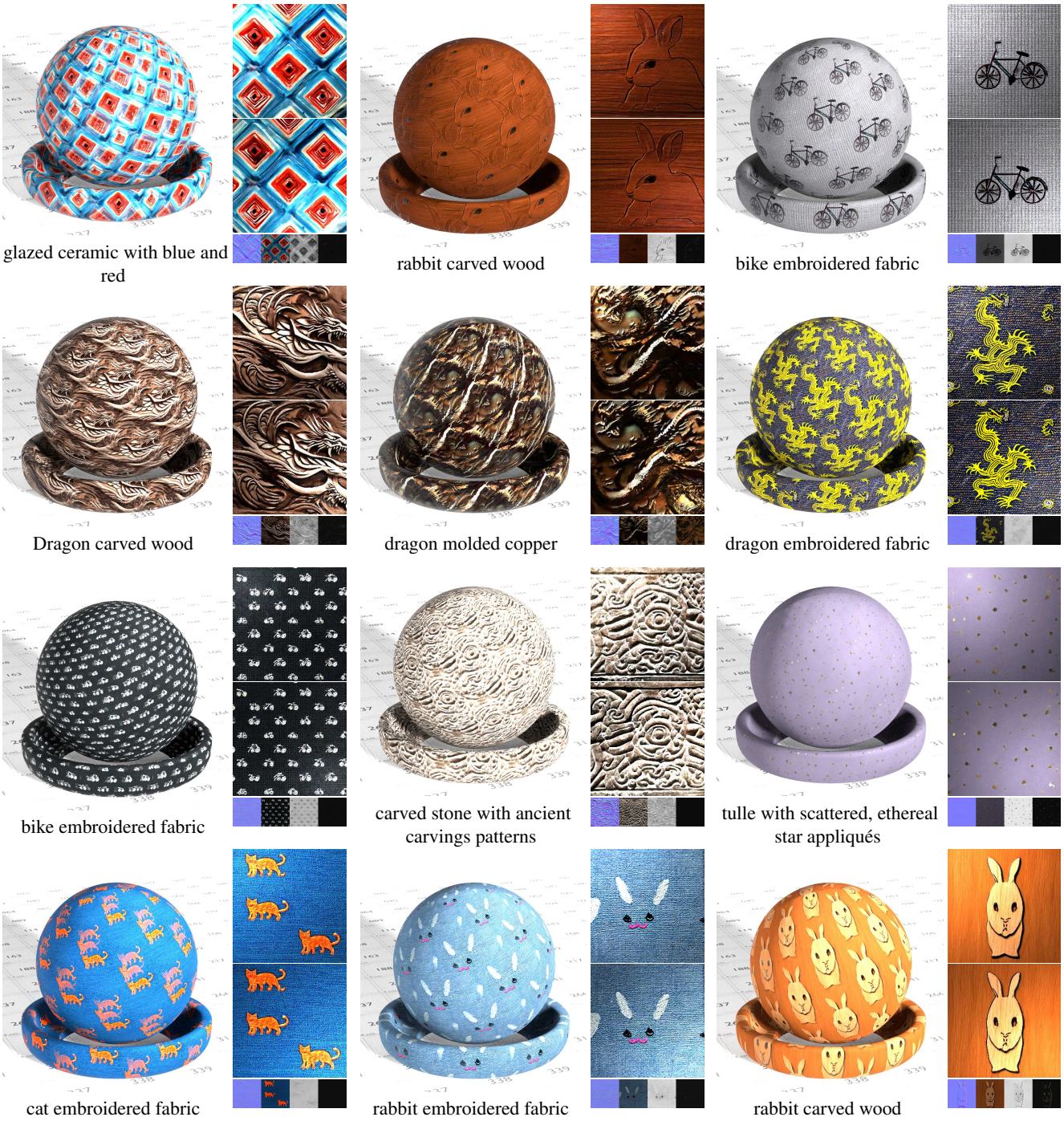


Figure 4: ***Non-stationary results.*** *We demonstrate our method for generating SVBRDF maps, with two flat renderings illuminated from the top left and bottom right corners shown on the right. The same texture, tiled and applied to a 3D shape, is displayed on the left.*

4.2. Performance

In conducting 100 network queries on an A100 GPU and averaging the time expenditure, Stable Diffusion 2 [RBL*22] takes 3.21 seconds to produce a static rendered image, whereas our architecture requires 5.16 seconds to be queried and generate the four SVBRDF maps. Although our approach is marginally slower than models dedicated to direct image generation, the disparity is minimal. The advantage gained, however, is the provision of fully editable and freely utilizable SVBRDF maps instead of single images.

Figure 5: ***Stationary results.*** *We demonstrate our method for generating SVBRDF maps, with two flat renderings illuminated from the top left and bottom right corners shown on the right. The same texture, tiled and applied to a 3D shape, is displayed on the left.*

4.3. Experiments

Initially, we experimented with three approaches, the results of which are presented in an ablation study illustrated in Fig. 3. The first strategy was to transfer learning from the Stable Diffusion 2 [RBL*22] model, starting by training its VAE. Given our goal to ensure model versatility, our first idea was to freeze the encoder part of the VAE's model, train the VAE's decoder, adjust it to output ten channels, increase the network layers, and train with the SVBRDF dataset. After the VAE was trained and its performance deemed acceptable, we proceeded to train the entire network end-to-end with the SVBRDF dataset. However, after sufficient training convergence, the network's performance was poor, producing unrealistic SVBRDF. Two examples are shown in Fig. 3.a demonstrate that the model was unable to distinguish wood material from metal due to color specular maps. The second method we attempted was to directly train a single diffusion model for text-to-SVBRDF conversion was hindered by lacking the training dataset, resulting in a model that could only represent a limited range of specific textures and demonstrated poor text description comprehension. For instance, as depicted in Fig. 3.(b), it failed to generate materials not present in the dataset, such as a carved bunny, since the model was only trained on plain wood. The final approach, which we currently employ, involves using a two-phase diffusion model pipeline shown in 3.(c).

5. Results

5.1. Main Results

Non-stationary: Stationary materials have uniform properties, easily represented statistically and suitable for seamless tiling, whereas non-stationary materials display structured, varying patterns that are less predictable. In Fig. 4, we demonstrate our method for generating SVBRDF maps followed by two flat renderings illuminated

Figure 6: ***Metal and glass results.*** *The upper half displays three flat renderings of the material, illuminated from the top left, center, and bottom right. For a continuous light interaction, please refer to the supplementary video. The lower half presents the corresponding SVBRDF maps.*

from the top left and bottom right corner. On the left, we show the same texture applied to a 3D shape. It is observable that our model is capable of producing meaningful textures following the text, such as bike-embroidered fabric and rabbit-carved wood; which presents certain challenges for artists. Furthermore, we can guide the texture generation with more complex language instructions, like glazed ceramic with blue and white, allowing us to control both the material and color. In order to seamlessly tile the textures, we synthesize a gradient mask to blend the original and quarter-shifted versions of the image. This tiling algorithm facilitates the creation of a continuous pattern that is suitable for repetitive tiling applications. The gradient mask is applied when we map the materials to the 3D ball which makes the SVBRDF tileable. We do not use this mask in the flat result.

Stationary: In Fig. 5, we have generated stationary results including leather, brick, wood, rubber, fabric, and stone. The rubber outcome has a diffuse map that is almost entirely black and is primarily dominated by the specular map, showcasing our network's capability to generate challenging materials accurately without baking the highlights in the maps.

Metal and glass: In Fig. 6, we use text to control the generation of metal and glass materials. The metal materials have colored specular maps unlike other results, and the rendered results exhibit

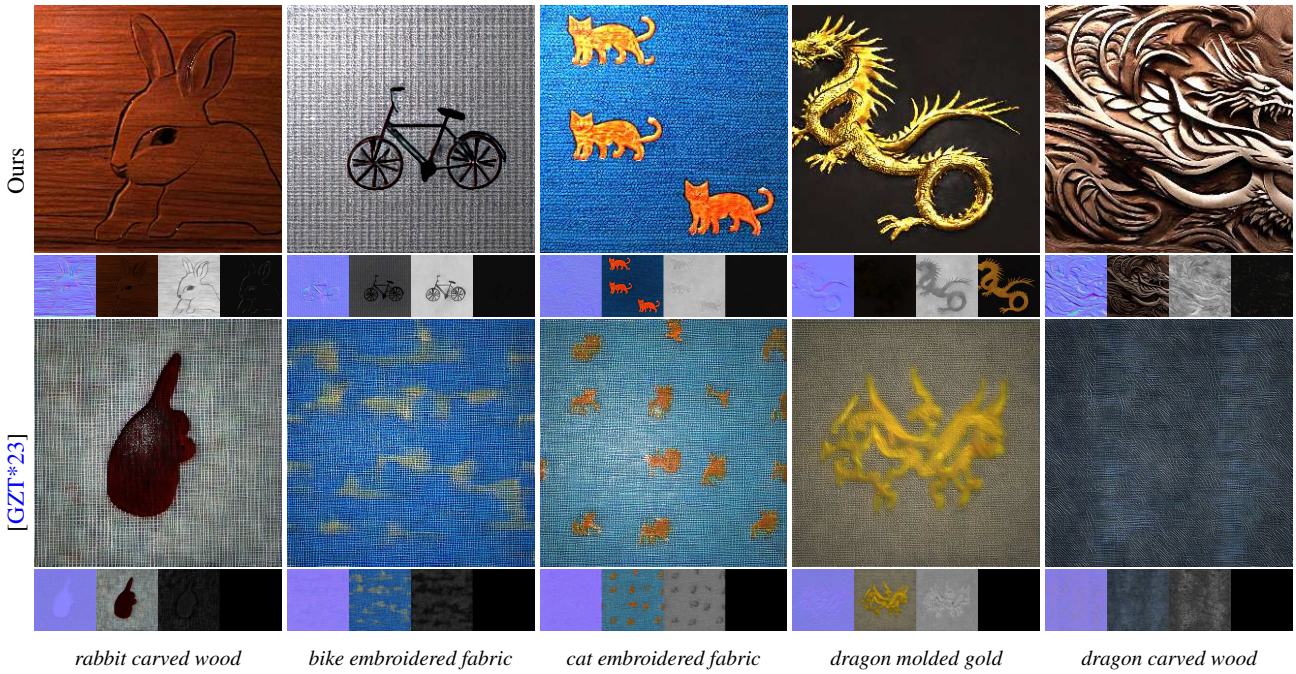


Figure 7: ***Comparison to previous model.*** *The upper half displays our flat render results, with the corresponding SVBRDF maps shown beneath, compared to results from Text2Mat [GZT*23]. Our results exhibit a significant improvement in texture generation quality over Text2Mat, which struggles to accurately convey the meaning of the text.*

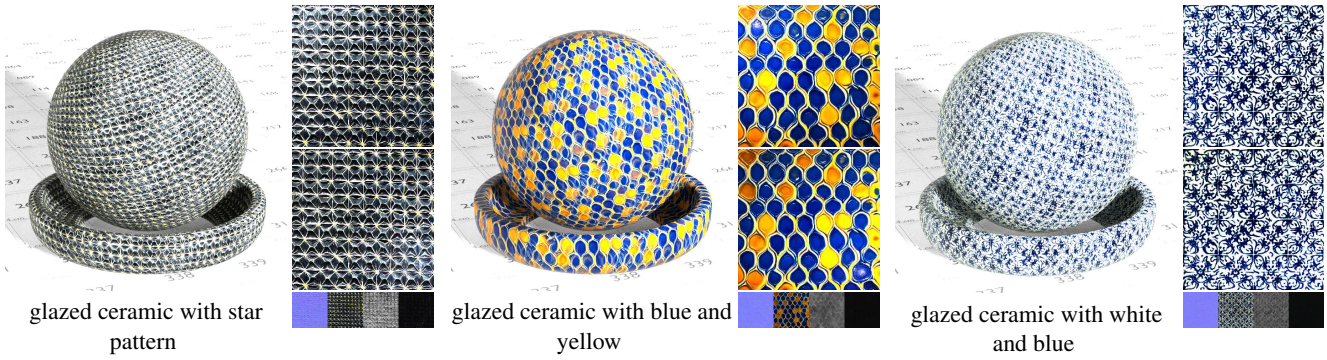


Figure 8: ***Editing shape and color.*** *We demonstrate how text can be used to control the appearance of materials, ceramic shown here, by using detailed descriptions to dictate color and pattern variations.*

metallic characteristics. In this figure, the three rendered images show the light source moving from the top left to the bottom right. Additionally, we have also generated glass materials that present a dielectric feeling. For simplicity, we chose to only show the flat renderings of metal and glass examples with the absence of light transmission.

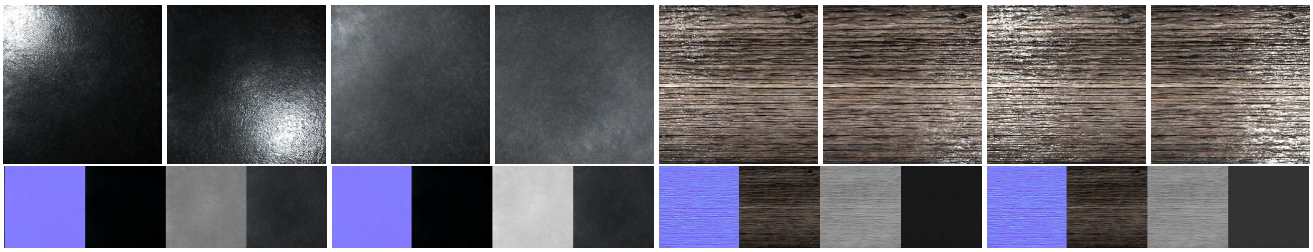
For additional results and detailed analyses, please refer to the supplementary materials.

5.2. Comparison to Previous Work

To our knowledge, the only existing model for converting text to materials is "Text2Mat: Generating Materials from Text." [GZT*23]. We compared our results with theirs, as shown in Fig. 7, where the text input into both models is provided at the bottom. It can be seen that Text2Mat has difficulty creating meaningful textures; it can only generate basic stationary materials and cannot fulfill complex requirements. Since Text2Mat is not open source, these results were obtained by sharing our text directly with the authors and requesting generation, which we would like to appreciate.

5.3. Editing Results

Text-controlled appearance features: Our model can manipulate the appearance features of the generated SVBRDF map, such as color and pattern, using text prompts. For single text input, our two-



(a) *metal with low roughness* **(b)** *metal with high roughness* **(c)** *wood with low specular* **(d)** *wood with high specular*

Figure 9: ***Controlling physical parameters with text.*** *We show two flat renderings illuminated from the top left and top right corners, with corresponding SVBRDF maps displayed below. The metal example demonstrates roughness adjustment for the same sample by modifying the text prompt, and the wood example illustrates control over specularity.*

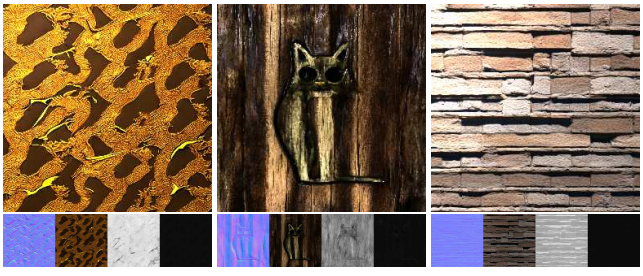
stage diffusion-based methods can produce multiple different textures for user to select. As illustrated in Fig. 8, we can control the texture's color or local shape through text by keeping the material constant while varying the detailed descriptions.

Text-Controlled physical parameters: Our network can also control physical parameters through text, specifically roughness, and specular levels. In Fig. 9, we demonstrate our method for controlling roughness and specular followed by two flat renderings illuminated from the top left and bottom right corner. By using a text controller, we can adjust these parameters simply by inputting text commands like "high/low roughness/specular" to manipulate the output's roughness and specular intensity, marked as Text 2 in Fig. 2. This can be also done by post-processing the generated SVBRDF maps that allow further editing of the physical parameters.

6. Discussion and Conclusion

6.1. Limitation and future work

Our model occasionally bakes highlights into the diffuse map. Furthermore, because diffusion occurs directly in image space, the resolution of our model is limited and currently low. This issue might be addressed by exploring latent-based diffusion. Similarly to other diffusion models, our model sometimes generates irrelevant or nonsensical results relative to the input text. Examples of these limitations, including failed cases, are illustrated in Fig. 10. However, we believe these issues can be overcome with further research.



dragon molded gold *metal cat on rough wood* *brick*

Figure 10: ***Failure cases*** *The first two cases did not successfully output textures as per the given text, while the last demonstrate the baked highlights and shadows*

6.2. Conclusion

Our model achieves generation from text to material, capable of producing specified textures or colors, thus filling a gap in related research. Furthermore, we can also create materials with special semantics, such as a Rabbit carved wood. To the best of our knowledge, we are the only ones able to perform this type of generation. Lastly, we would like to thank the organizations that provided the UBO and INRIA synthetic SVBRDF datasets and express our gratitude to Text2Mat for providing comparison images.

References

[BPH*24] BROOKS, TIM, PEEBLES, BILL, HOLMES, CONNOR, et al. "Video generation models as world simulators". (2024). URL: https://openai.com/research/video-generation-models-as-world-simulators 3.

[DAD*18] DESCHAINTRE, VALENTIN, AITTALA, MIIKA, DURAND, FREDO, et al. "Single-image svbrdf capture with a rendering-aware deep network". *ACM Transactions on Graphics (ToG)* 37.4 (2018), 1–15 2, 4.

[DN21] DHARIWAL, PRAFULLA and NICHOL, ALEXANDER. "Diffusion models beat gans on image synthesis". *Advances in neural information processing systems* 34 (2021), 8780–8794 2, 4.

[DvGNK99] DANA, KRISTIN J., van GINNEKEN, BRAM, NAYAR, SHREE K., and KOENDERINK, JAN J. "Reflectance and Texture of Real-World Surfaces". *ACM Trans. Graph.* 18.1 (Jan. 1999), 1–34. ISSN: 0730-0301. DOI: 10.1145/300776.300778. URL: https://doi.org/10.1145/300776.300778 2.

[EKB*24] ESSER, PATRICK, KULAL, SUMITH, BLATTMANN, ANDREAS, et al. "Scaling rectified flow transformers for high-resolution image synthesis". *arXiv preprint arXiv:2403.03206* (2024) 3.

[GGG*16] GUARNERA, D., GUARNERA, G.C., GHOSH, A., et al. "BRDF Representation and Acquisition". *Computer Graphics Forum* 35.2 (2016), 625–650. DOI: https://doi.org/10.1111/cgf.12867 2.

[GLT*21] GUO, JIE, LAI, SHUICHANG, TAO, CHENGZHI, et al. "Highlight-aware two-stream network for single-image SVBRDF acquisition". (2021) 2.

[GPM*20] GOODFELLOW, IAN, POUGET-ABADIE, JEAN, MIRZA, MEHDI, et al. "Generative adversarial networks". *Communications of the ACM* 63.11 (2020), 139–144 2.

[GSH*20] GUO, YU, SMITH, CAMERON, HAŠAN, MILOŠ, et al. "MaterialGAN: Reflectance capture using a generative SVBRDF model". *arXiv preprint arXiv:2010.00114* (2020) 2.

[GZT*23] GUO, ZHEN HE1 JIE, ZHANG, YAN, TU, QINGHAO, et al. "Text2Mat: Generating Materials from Text". (2023) 2, 8.

[HDR19] HU, YIWEI, DORSEY, JULIE, and RUSHMEIER, HOLLY. "A novel framework for inverse procedural texture modeling". *ACM Transactions on Graphics (ToG)* 38.6 (2019), 1–14 2.

[HGH*23] HU, YIWEI, GUERRERO, PAUL, HASAN, MILOS, et al. "Generating Procedural Materials from Text or Image Prompts". *ACM SIGGRAPH 2023 Conference Proceedings*. 2023, 1–11 2.

[HJA20] HO, JONATHAN, JAIN, AJAY, and ABBEEL, PIETER. "Denoising diffusion probabilistic models". *Advances in neural information processing systems* 33 (2020), 6840–6851 2.

[HZL*17] HAO, YANCHAO, ZHANG, YUANZHE, LIU, KANG, et al. "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge". *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, 221–231 3.

[KCW*18] KANG, KAIZHANG, CHEN, ZIMIN, WANG, JIAPING, et al. "Efficient reflectance capture using an autoencoder". *ACM Trans. Graph.* 37.4 (July 2018). ISSN: 0730-0301. DOI: 10.1145/3197517.3201279. URL: https://doi.org/10.1145/3197517.3201279 2.

[KMX*21] KUZNETSOV, ALEXANDR, MULLIA, KRISHNA, XU, ZEXIANG, et al. "NeuMIP: Multi-Resolution Neural Materials". *Transactions on Graphics (Proceedings of SIGGRAPH)* 40.4 (July 2021) 2.

[KW13] KINGMA, DIEDERIK P and WELLING, MAX. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114* (2013) 2.

[MGZJ20] MONTAZERI, ZAHRA, GAMMELMARK, SØREN B, ZHAO, SHUANG, and JENSEN, HENRIK WANN. "A practical ply-based appearance model of woven fabrics". *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–13 2.

[MK14] MICHAEL WEINMANN, JUERGEN GALL and KLEIN., REINHARD. "Material Classification based on Training Data Synthesized Using a BTF Database". *ECCV* (2014) 2.

[MK20] MERZBACH, SEBASTIAN and KLEIN, REINHARD. "Bonn Appearance Benchmark." *MAM@ EGSR*. 2020, 21–24 2, 4.

[MXZ*23] MA, XIAOHE, XU, XIANMIN, ZHANG, LEYAO, et al. "OpenSVBRDF: A Database of Measured Spatially-Varying Reflectance". *ACM Trans. Graph.* 42.6 (Dec. 2023). ISSN: 0730-0301. DOI: 10.1145/3618358. URL: https://doi.org/10.1145/3618358 2.

[Nic65] NICODEMUS, FRED E. "Directional reflectance and emissivity of an opaque surface". *Applied optics* 4.7 (1965), 767–775 2.

[PEL*23] PODELL, DUSTIN, ENGLISH, ZION, LACEY, KYLE, et al. "Sdxl: Improving latent diffusion models for high-resolution image synthesis". *arXiv preprint arXiv:2307.01952* (2023) 3.

[RBL*22] ROMBACH, ROBIN, BLATTMANN, ANDREAS, LORENZ, DOMINIK, et al. "High-resolution image synthesis with latent diffusion models". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 10684–10695 3–6.

[SH22] SALIMANS, TIM and HO, JONATHAN. "Progressive distillation for fast sampling of diffusion models". *arXiv preprint arXiv:2202.00512* (2022) 4.

[SLH*20] SHI, LIANG, LI, BEICHEN, HAŠAN, MILOŠ, et al. "Match: Differentiable material graphs for procedural material capture". *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–15 2.

[SP23] SARTOR, SAM and PEERS, PIETER. "Matfusion: a generative diffusion model for svbrdf capture". *SIGGRAPH Asia 2023 Conference Papers*. 2023, 1–10 2, 4.

[SZ14] SIMONYAN, KAREN and ZISSERMAN, ANDREW. "Very deep convolutional networks for large-scale image recognition". *arXiv preprint arXiv:1409.1556* (2014) 3.

[VMR*23] VECCHIO, GIUSEPPE, MARTIN, ROSALIE, ROULLIER, ARTHUR, et al. "ControlMat: A Controlled Generative Approach to Material Capture". *arXiv preprint arXiv:2309.01700* (2023) 2.

[VPS21] VECCHIO, GIUSEPPE, PALAZZO, SIMONE, and SPAMPINATO, CONCETTO. "Surfacenet: Adversarial svbrdf estimation from a single image". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 12840–12848 2.

[VSPS23] VECCHIO, GIUSEPPE, SORTINO, RENATO, PALAZZO, SIMONE, and SPAMPINATO, CONCETTO. "Matfuse: Controllable material generation with diffusion models". *arXiv preprint arXiv:2308.11408* (2023) 2.

[XZJM23] XUE, BOWEN, ZHAO, SHUANG, JENSEN, HENRIK WANN, and MONTAZERI, ZAHRA. *A Hierarchical Architecture for Neural Materials*. 2023. arXiv: 2307.10135 [cs.GR] 2.

[ZK21] ZHOU, XILONG and KALANTARI, NIMA KHADEMI. "Adversarial Single-Image SVBRDF Estimation with Hybrid Training". *Computer Graphics Forum*. Vol. 40. 2. Wiley Online Library. 2021, 315–325 2.

[ZMA*23] ZHU, JUNQIU, MONTAZERI, ZAHRA, AUBRY, JEAN-MARIE, et al. "A Practical and Hierarchical Yarn-based Shading Model for Cloth". *Computer Graphics Forum* (2023). ISSN: 1467-8659. DOI: 10.1111/cgf.14894 2.