UMat: Uncertainty-Aware Single Image High Resolution Material Capture

Carlos Rodriguez-Pardo[1,2]   Henar Dominguez-Elvira[1,2]   David Pascual-Hernandez[1]   Elena Garces[1,2]

[1]SEDDI, Spain   [2]Universidad Rey Juan Carlos, Spain

Abstract

We propose a learning-based method to recover normals, specularity, and roughness from a single diffuse image of a material, using microgeometry appearance as our primary cue. Previous methods that work on single images tend to produce over-smooth outputs with artifacts, operate at limited resolution, or train one model per class with little room for generalization. In contrast, in this work, we propose a novel capture approach that leverages a generative network with attention and a U-Net discriminator, which shows outstanding performance integrating global information at reduced computational complexity. We showcase the performance of our method with a real dataset of digitized textile materials and show that a commodity flatbed scanner can produce the type of diffuse illumination required as input to our method. Additionally, because the problem might be ill-posed –more than a single diffuse image might be needed to disambiguate the specular reflection– or because the training dataset is not representative enough of the real distribution, we propose a novel framework to quantify the model's confidence about its prediction at test time. Our method is the first one to deal with the problem of modeling uncertainty in material digitization, increasing the trustworthiness of the process and enabling more intelligent strategies for dataset creation, as we demonstrate with an active learning experiment.

1. Introduction

Virtual design, online marketplaces, product lifecycle workflows, AR/VR, videogames, . . . , all require lifelike digital representations of real-world materials (*i.e.*, digital twins). Acquiring these digital copies is typically a cumbersome and slow process that requires expensive machines and several manual steps, creating roadblocks for scalability, repeatability, and consistency. Among the many industries requiring digital twins of materials, the fashion industry is in a critical position; facing the demand to digitize hundreds of samples of textiles in short periods, which cannot be achieved with the current technology.
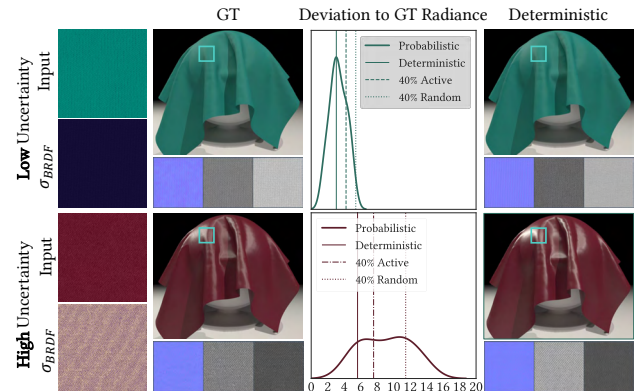

Figure 1. Our method digitizes a material taking as input a single scanned image. Further, it returns a pixel-wise metric of uncertainty σ_{BRDF}, computed at test time through probabilistic sampling, proven useful for active learning. In the plot we compare the average deviations of the radiance of different renders in the blue crop w.r.t the ground truth (GT) of: 1) the distribution of the probabilistic samples of a model trained with 100% of the data; 2) the deterministic output of that model; 3) the output of a model trained using 40% of the training dataset, sampled by active learning guided by σ_{BRDF} and; 4) a model trained using 40% of the training dataset, randomly sampled. The material at the bottom, for which the model shows a higher uncertainty, generates more varied renders and differs most from the ground truth.

In this context, casual capture systems for optical digitization provide a promising path for scalability. These systems leverage handheld devices (such as smartphones), one or more different illuminations, and learning-based priors to estimate the material's diffuse and specular reflection lobes. However, existing approaches present several drawbacks that make them unsuitable for practical digitization workflows. Generative solutions [18, 63] typically produce unrealistic artifacts. Despite recent attempts to improve tileability and controllability [73], these solutions are slow to train and to evaluate (requiring online optimization iterations), are limited in resolution, and present challenges for generalization (requiring one model per material class). Further, the fact that these methods build on perceptual losses –not pixel losses– to compare the input photo with the generated material entails extra difficulties when it comes to guaranteeing the repeatability and consistency required for build-

Figure 2. Scanner images vs fitted albedos.

ing a digital inventory (i.e., color swatches, prints, or other variations). On the other hand, methods that build on differentiable node graphs [21] overcome the tileability and resolution limitations, yet, they share the problems derived from using perceptual losses and category-specific training.

In this work, we present *UMat*, a practical, scalable, and reliable approach to digitizing the optical appearance of textile material samples using SVBRDFs. Commonly used SVBRDFs typically contain two reflection terms: a diffuse term, parameterized by an albedo image, and a specular one, parameterized by normals, specularity, and roughness. Prior work typically estimates both components, which becomes a very challenging problem, obtaining over-smooth outputs, and being prone to artifacts [7, 16, 74]. Instead, in this paper, we demonstrate that it is possible to provide accurate digitizations of materials leveraging as input a single diffuse image that acts as albedo and estimating the specular components using a neural network. Our key observation is to realize that most of the appearance variability of textile materials is due to its microgeometry and that a commodity flatbed scanner can approximate the type of diffuse illumination that we require for the majority of textile materials (see Figure 2).

Nevertheless, single-image material estimation is still an ill-posed problem in our setting, as reflectance properties may not be directly observable from a single diffuse image. To account for these non-directly observable properties, we propose a novel way to measure the model's confidence about its prediction at test time. Leveraging Monte Carlo (MC) Dropout [10], we propose an uncertainty metric computed as the variance of sampling and evaluating multiple estimations for a single input in a render space. We show that this confidence directly correlates with the accuracy of the digitization, which helps identify ambiguous inputs, out-of-distribution samples, or under-represented classes. Besides increasing the trustworthiness of the capture process, our confidence quantification enables smarter strategies for dataset creation, as we demonstrate with an active learning experiment.

We pose the estimation as an Image-to-Image Translation problem (I2IT) that directly regresses roughness, specular, and normals, from a single input image. Under the hood, our novel residual architecture has a single encoder enhanced with lightweight attention modules [45, 66] for improving global consistency and reducing artifacts, specialized decoders for each target reflectance map, and a U-Net discriminator [57], which enhances generalization.

In summary, we present the following contributions:

- A novel material capture system which leverages the diffuse illumination provided by flatbed scanners for high-resolution, scalable and reliable digitizations.
- An attention-enhanced GAN model and training procedure designed for maximizing accuracy and sharpness, and removing undesired artifacts.
- A generic uncertainty quantification framework for material capture algorithms which correlates with prediction error on a render space.

2. Related Work

Single Image SVBRDF Capture Capturing accurate SVBRDFs from a single image is a challenging problem that requires predicting the photometric response of a material given only a sample of it. The most common approximation is to use a flash-lit front planar image captured with a smartphone. Extending neural style transfer [14] to material capture, Aittala *et al.* [1] leverage pre-trained CNNs and texture priors for smartphone material acquisition. Relatedly, Henzler *et al.* [21] use style losses for training generative models for BRDF synthesis. These approaches are optimized for synthesis, which allow for seamlessly tileable outputs, and do not require supervised training. However, they work best for stochastic materials, limiting their scope.

Leveraging datasets of labeled materials, different methods have trained autoencoders for SVBRDF capture. Li *et al.* [38] reconstruct spatially-varying albedo and normals using a U-Net [53] and homogeneous specular albedo and roughness using a CNN regressor. This work was extended through Self-Augmented CNNs in [70]. By leveraging CRFs, a material classifier and one decoder per map, Li *et al.* [39] reconstruct spatially-varying albedo, roughness and normals. Deschaintre *et al.* [7] propose a modified U-Net, synthetic datasets and a render loss. Cascaded models [40, 56]; and deep latent spaces optimized using inverse rendering [11] have also shown success for this problem. Recently, Generative Adversarial Networks (GANs) have shown improved capabilities compared to more naive losses. These require a discriminator, which can be trained on renders [67], SVBRDF maps [16, 63], or both [74].

Our approach differs from previous work in several factors. Importantly, we use flatbed scanners instead of smartphones for material capture. While they limit the materials which can be captured, they provide adequate illumination for easier digitizations, and a higher level of resolution and detail. We hypothesize that material specularity can be estimated accurately by leveraging its microgeometry. From this assumption, we build a GAN which, in contrast with previous work, leverages state-of-the-art attention mechanisms and discriminator design for obtaining a more holistic understanding of its inputs. Further, we train exclusively on real data and propose a more comprehensive evaluation.
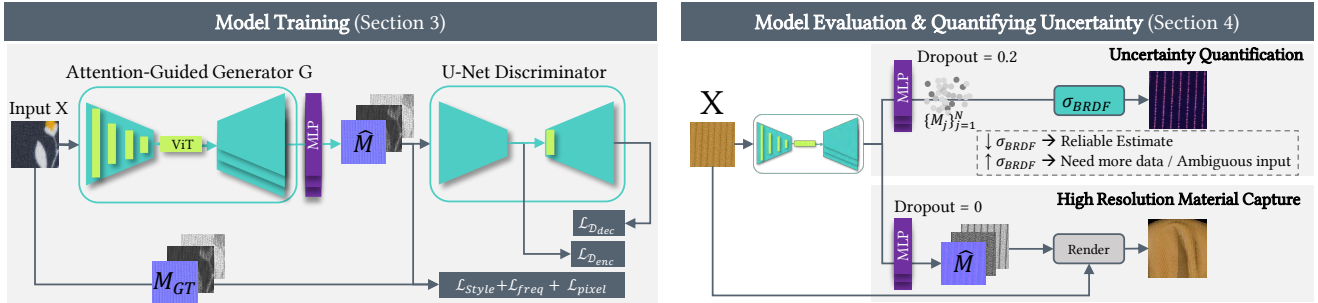
Figure 3. Overview of *UMat*. We propose an attention-guided generator with a U-Net discriminator trained with style, frequency, pixel-wise, and adversarial losses. In green, we show the components that include any form of attention mechanism. The supplementary material contains the detailed architectures. On the right, we show two applications of our method: First, thanks to our test-time uncertainty evaluation, we can provide a measure of reliability of the estimation. Second, the maps that we produce can be used by any render engine.

Procedural graphs have also been used for material generation. Instead of relying on material priors, these approaches work by optimizing a material graph through a differentiable pipeline. These provide interesting capabilities, such as easy edition or tiling, but are limited by the expressiveness of the procedural model. These have been explored for general SVBRDF estimation [17, 24, 59], or for high-quality woven fabric digitizations [29].

Uncertainty Quantification in Deep Learning Measuring the confidence of deep learning models is an active research area [32] with multiple applications, including safety-critical problems, like self-driving [25] or medicine [34]; and active dataset creation [37, 60]. In computer vision, uncertainty quantification has focused on image classification [37, 58], segmentation [33], and depth regression [3, 23]. To overcome the computational intractability of Bayesian Neural Networks, different approximations have been proposed, including MC Dropout [10], Deep Ensembles [35] or Variational Inference [46]. Orthogonal alternatives exist, including Evidential Deep Learning [3, 58] or frequentist approaches like Constrained Ordinal Regression [23]. We refer the reader to recent surveys [15, 30] for more comprehensive reviews.

Quantifying uncertainties allows to communicate the end users that the model predictions may be inaccurate, suggesting alternative pathways; as well as cheaper dataset creation through *active learning*. Bayesian material parameter estimation has been proposed for procedural frameworks [17], but, to the best of our knowledge, it has not been explored for SVBRDF estimation. We aim to propose an efficient uncertainty quantification framework for deep SVBRDF capture methods which accounts for material perception, and is agnostic to the material model.

3. Method

Our method starts from an input image X of a material taken under diffuse lighting and outputs the parameters of the specular lobe of the SVBRDF, *i.e.*, $\mathbf{M} = \{\mathrm{M}_i\}_{i=1}^{3}$ corresponding to the material roughness, specularity, and nor-

mals. We illustrate this process in Figure 3. Following previous work [31, 48], we use the physically-based material model from Disney [5], which aggregates a diffuse term with an isotropic, microfacet specular GGX lobe $s(\mathbf{M})$ [64], such that, $f_{l,v}(\mathbf{M}, \mathbf{X}) = \frac{\mathbf{X}}{\pi} + s_{l,v}(\mathbf{M}) \in \mathbb{R}^{x \times y}$ is the shading model for a light l and view position v. We formulate this estimation as an Image-to-Image Translation problem (I2IT). We train a U-Net [53] generator $G(\mathbf{X}) = \hat{\mathbf{M}}$ within a GAN framework, extending the adversarial loss with pixel-wise, style, and frequency losses. We design the generator to maximize accuracy, sharpness, and robustness. To do so, we train a single encoder, enhanced with self-attention layers and a transformer, and use one decoder per map. Sections 3.1, 3.2, and 3.3 present the network design, the loss and data augmentation choices, respectively. Implementation details are provided on the supplementary material.

Using as input a single image taken under diffuse lighting presents extra challenges when estimating the SVBRDF; we lack the extra cues provided by more complex illumination patterns (e.g. flash lighting). Therefore, in Section 4, we explain how to compensate this potential ambiguity by introducing an uncertainty metric that can be computed at test time. Section 5 presents the evaluation, which includes the description of our dataset and metrics (Sections 5.1 and 5.2), an ablation study that validates the design (5.3), qualitative and quantitative results (5.4 and 5.5), an application of our uncertainty metric in an active learning setting (5.6), and comparisons with previous work (5.7).

3.1. Network Design

We use a U-Net [53] with residual connections [8, 9, 20] in all of our convolutional blocks, an individual decoder per map [8, 12, 51, 74], and group normalization [69]. To each decoder, we add a pixel-wise dropout-regularized MLP [61], aimed at increasing the accuracy of the predictions and to allow us to measure uncertainty at test-time. While this multi-decoder residual U-Net is relatively accurate, it is limited by its receptive field, as is common on fully-convolutional architectures. Previous work [7] proposed the use of a *global track* for fusing spatially distant

information. We instead draw inspiration from recent advances in attention and diffusion models [22, 52, 55], and add a self-attention module with linear complexity [66] to the output of every convolutional block in the encoder. Finally, we add a lightweight MobileViT [45] to the bottleneck to provide the model with a global understanding of its input. By performing the most complex computations at the encoder, we provide the specialized decoders with dense inputs which are computed only once.

3.2. Loss Function

Our loss function is comprised of four terms: pixel-wise losses, an adversarial loss, a style loss, and a frequency loss:

$$\mathcal{L}_G = \sum_i \lambda_i \mathcal{L}_{pixel_i} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{style}\mathcal{L}_{style} + \lambda_{freq}\mathcal{L}_{freq} \tag{1}$$

$\mathcal{L}_{pixel}$ is the $\mathcal{L}_1$ norm weighted per map, λ_i. $\mathcal{L}_1$ produces sharper results than higher-order alternatives, such as $\mathcal{L}_2$. We introduce an adversarial loss to handle the intrinsic ambiguity of ill-posed problems [12, 27, 62, 63]. In our case, the choice of the discriminator is a critical design decision. Recent work [57] proposed U-Net architectures for discriminators, which allows to better learn both low and high-level features, and to introduce further regularization. These result in more conservative albeit less diverse generations [19]. We use a U-Net discriminator [57] with attention [68], which outputs two estimations: a scalar output $\mathcal{D}_{enc}$, provided by its encoder, and a 2D estimation $\mathcal{D}_{dec}$, provided by its decoder. $\mathcal{D}_{enc}$ provides a global estimate of the quality of the stack $\mathbf{M}$, while $\mathcal{D}_{dec}$ gives pixel-wise estimations. As in [57], we add a *regularization* term $\mathcal{L}_{\mathcal{D}_{dec}}^{cons}$, and leverage *cut-mix* as for discriminator data-augmentation. Our discriminator and adversarial losses are:

$$\mathcal{L}_{\mathcal{D}} = \mathcal{L}_{\mathcal{D}_{enc}} + \mathcal{L}_{\mathcal{D}_{dec}} + \lambda_{cons}\mathcal{L}_{\mathcal{D}_{dec}}^{cons} \tag{2}$$

$$\mathcal{L}_{adv} = \log(\mathcal{D}_{enc}(G(X))) + \log(\mathcal{D}_{dec}(G(X))) \tag{3}$$

where the implementation of $\mathcal{L}_{D_{enc}}$ and $\mathcal{L}_{D_{dec}}$ follows [57].

We further add two losses to improve the accuracy and sharpness of the results: a frequency loss $\mathcal{L}_{freq}$ and a style loss $\mathcal{L}_{style}$. $\mathcal{L}_{freq}$ is estimated by averaging the *focal frequency loss* [28] computed over each individual channel of $\mathbf{M}$. This is designed to help GANs preserve high-frequency details. Further, as shown in prior work, working in the frequency domain is beneficial when handling textures [2, 43]. Our style loss $\mathcal{L}_{style}$ is inspired by the success of neural losses when dealing with textures [13, 14]. However, off-the-shelf metrics which are designed for 3-channel images, are not immediately usable in SVBRDF. While it is possible to compute them for each map separetely [51], this does not necessarily preserve inter-map consistency. We follow
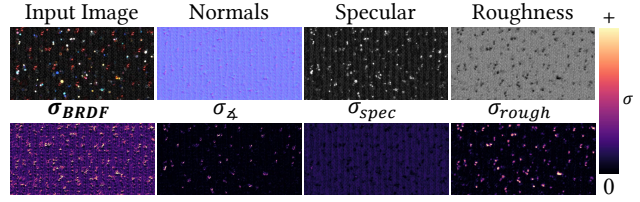

Figure 4. Top: input image of a *rib* material with metallic sequins. Bottom: σ_{BRDF} and per-map uncertainties.

recent work [6] and use LPIPS [71] taking as input a 3-channel image created by randomly sampling three channels from the set of five available channels of $\mathbf{M}$.

3.3. Data Augmentation

We follow two strategies for data augmentation. First, we perform patch-based training and affine transforms with randomly cropped patches [50, 62, 63]. We also apply random rescales for generalization at lower resolutions, and rotations to account for possible misalignments that may arise when capturing the samples. Second, we apply several image transformations to increase model robustness: random intensity changes in HSV space to the inputs, gaussian noise and blurs, and random erasing [72] for regularization.

4. Uncertainty Estimation

Material acquisition from a single diffuse image is potentially an ill-posed problem, since it assumes that the microgeometry is a sufficient cue to predict the material appearance. While pure learned priors have been proven to work well for inverse problems of single images object shape estimation [26, 40–42, 47], they are typically combined with render losses that guarantee consistency in the reconstruction. We lack the necessary input to include this kind of supervision, therefore, we propose an alternative approach aimed at quantifying the confidence of the prediction. This is valuable for several purposes. It provides a way of communicating possible inaccuracies to users of these systems; and more importantly, it enables for efficient dataset creation through active learning, as we show in Section 5.6.

We propose an uncertainty quantification mechanism which is applied to invidual per-map estimations, and also globally in a render space. It is possible to measure uncertainty, among other methods, through deep ensembles [35], evidential learning [4, 58, 65], or ordinal regression [23]. However, they are costly to train and evaluate, and would imply major changes in our method. Instead, we follow a probabilistic approach called MC Dropout [10], with which, for a particular input, we sample a set of predictions by adding randomness to the forward pass of our model. This process has no impact in the regular deterministic evaluation and implies no changes in our model architecture. Specifically, while measuring uncertainty, we randomly deactivate 20% of the neurons of the MLP of our decoders,

obtaining a set of outputs $U = \{\hat{M}_j \sim G(X)\}_{j=1}^N$ that we use to compute several metrics.

First, we compute the pixel-wise standard deviation for each invididual map separately, obtaining $\sigma_\angle$, σ_{spec}, and σ_{rough}, for the normals, specular, and roughness maps, respectively. Then, inspired by perceptual metrics for computing BRDF differences [36], we define our novel perceptually-aware uncertainty metric σ_{BRDF} as follows:

$$\sigma_{\mathrm{BRDF}} = \frac{1}{|xy|} \sum_{xy} \log \left(\frac{1}{|S|} \sqrt{ \sum_{(l,v) \in S} \sqrt[3]{\sigma_{l,v}(\{f_{l,v}(\mathrm{U}_j, K)\cos(\theta_l)\}_{j=1}^N)} } \right) \tag{4}$$

where K is a 2D image of a constant neutral grey value, $\sigma_{l,v} \in \mathbb{R}^{x \times y}$ is the pixel-wise standard deviation of the renders $f_{l,v}$ obtained for the set of sampled maps U at light l and view position v, and S is the set of 50 views optimized in [49] for efficient BRDF capture. The log spatially-varying result is averaged across the spatial dimensions xy to obtain a single value as output. This equation introduces perceptual components to our uncertainty metric in two forms: first, by applying cosine weighting to the light position to compensate for light attenuation at grazing angles, and second, by taking the cubic root of these differences to attenuate peak reflectances. Figure 4 showcases an example of the per-map and BRDF uncertainties for a rib material with sequins. While the uncertainty is low at the yarns because it is a common material in our dataset, it appears high in the sequins since that effect had not been observed during training.

5. Experimental Results

5.1. Dataset

We gathered a novel dataset for training and testing our method. It comprises 2000 textile materials of a variety of families and microstructures that we divided into 14 families: crepe, jacquard, pile, plain, satin, and twill for *wovens*; fleece, french terry, interlock, jersey, milano, pique, and rib for *knits*; and, finally, *leathers*. Each family differs in its construction pattern (*i.e.,* its microstructure), which directly impacts its optical appearance. In the supplementary material we show a more detailed analysis of the dataset. For each material, we have an image scanned with the flatbed scanner EPSON V850 whose lighting configuration is close to diffuse (as we show in Figure 2), and its corresponding ground truth specular maps of the SVBRDF (normals, specular, and roughness). To obtain these maps, first, we digitized the material with an optical gonioreflectometer and then, propagated the maps to the scan using map propagation techniques [50]. All our images and maps have a resolution of 1000 PPIs, allowing us to leverage the full semantics of the microstructure for inference. We split our dataset in 90-10 for train and test, making sure that every family is equally represented in both splits.

5.2. Metrics

We quantify individual *per-map accuracy*, *rendered perceptual accuracy*, and *artifacts*. **Per-map accuracy** is computed differently depending on the semantics of the map: the roughness and specular maps are evaluated using Mean Absolute Error ($\mathcal{L}_1$), and the normals are evaluated using the angular distance in vector space ($\mathcal{L}_\angle$). Further, to account for the possibility that the model always returns an accurate average value, resulting in relatively low $\mathcal{L}_1$, we also measure Pearson correlation ρ.

We evaluate **rendered perceptual accuracy** $\mathcal{L}_{\mathrm{BRDF}}$ between the ground truth stack $\mathbf{M}_{GT}$ and the estimation $\hat{\mathbf{M}}$ following existing metrics [36],

$$\mathcal{L}_{\mathrm{BRDF}} = \frac{1}{|xy|} \sum_{xy} \sqrt{ \frac{1}{|S|} \sum_{(l,v) \in S} \sqrt[3]{\cos^2(\theta_l)\left(f_{l,v}(\mathbf{M}_{GT}, K) - f_{l,v}(\hat{\mathbf{M}}, K) \right)^2} } \tag{5}$$

where the terms are the same as in Equation 4.

Finally, we found that some architectural improvements in the neural network introduce artifacts in the specular and roughness maps that were not present in the input image, as illustrated in Figure 5. Thus, we provide an **artifacts detection** metric to quantify them. We start by defining a metric of homogeneity for an input image I,

$$\mathcal{H}(\mathrm{I}) = \frac{1}{|xy|} \sum_{xy} \frac{1}{|d|} \sum_{d=\{\uparrow,\downarrow,\leftarrow,\rightarrow\}} \|F_{\mathrm{Box}}(\mathrm{I}) - F_{\mathrm{Box}}(\mathrm{I}^d)\|_1 \tag{6}$$

where I^d is the image shifted up, down, left, and right by a number of pixels equal to the kernel size of the box filter. Then, we define three metrics that we compute per map: $e_1(\mathrm{M}) = \mathcal{H}(\mathrm{M})$, $e_2(\mathrm{M}) = \frac{\mathcal{H}(\mathrm{M})}{\mathcal{H}(\mathrm{X})}$, and $e_3(\mathrm{M}) = MI(\mathrm{X},\mathrm{M})^{-1}$. X is the original input image, and *MI* is the Mutual Information [54], which helps discern artifacts that appear in a single map from semantic patterns (*e.g.,* plaids, prints). Each map is labeled as having artifacts if the majority of metrics exceed their corresponding thresholds, $t_m(\mathrm{M})$. If any of the maps contain artifacts, the entire stack $\mathbf{M}$ is classified as having artifacts. The values for the thresholds and filter are included in the supplementary material.

5.3. Ablation Study

In Figure 5 and Table 1, we present an ablation study to validate our model design. From a baseline U-Net [53] trained with a pixel-wise loss and no data augmentation other than rescales, we add different components to the model to improve its generalization. First, we observe that using a PatchGAN [27,63] discriminator provides a significant increase in accuracy. However, using a similarly-sized U-Net discriminator [57], we achieve better results, particularly when using discriminator regularization. Further,
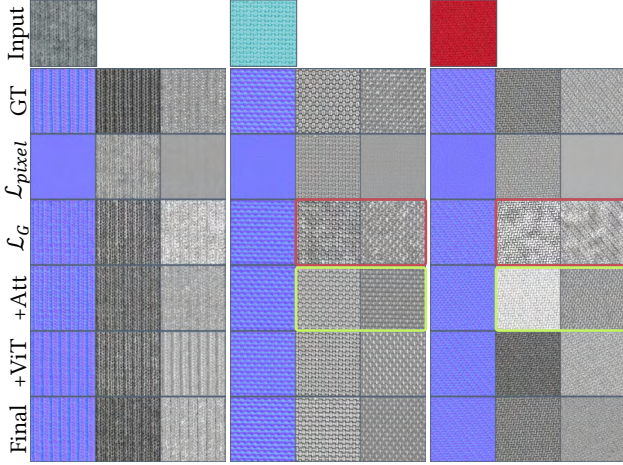
Figure 5. Qualitative results of some configurations of our ablation study. In red, we show that the baseline generator architecture trained on the full loss introduces artifacts, which are removed using attention on the encoder. Further results are included in the supplementary material.

Figure 6. Embeddings of the transformer of our generator for data in the training set, reduced using UMAP [44].

	Configuration	$\rho^S\uparrow$	$\rho^R\uparrow$	$\mathcal{L}_\angle\downarrow$	$\mathcal{L}_1^S\downarrow$	$\mathcal{L}_1^R\downarrow$	$\mathcal{L}_{BRDF}\downarrow$	Art.$\downarrow$
	Baseline	0.615	0.329	7.570	0.130	0.070	0.325	0.0
Loss	Baseline + PatchGAN	0.810	0.510	3.729	0.089	0.069	0.307	12.0
	Baseline + U-Net D.	0.854	0.658	2.950	0.085	0.068	0.299	28.0
	U-Net D. + $\mathcal{L}^{cons}_{\mathcal{D}_{dec}}$	0.858	0.653	2.860	0.088	0.068	0.296	19.0
	$\mathcal{L}^{cons}_{\mathcal{D}_{dec}}$ + $\mathcal{L}_{style}$	0.831	0.655	2.790	0.091	0.060	0.289	23.0
	$\mathcal{L}_{style}$ + $\mathcal{L}_{freq}$	0.856	0.677	2.410	0.086	0.059	0.288	20.1
Model	$\mathcal{L}_{freq}$ + Residual	0.855	0.665	2.310	0.089	0.060	0.285	25.5
	Residual + Decoders	0.863	0.699	2.120	0.079	0.054	0.276	18.5
	Decoders + Attention	0.860	0.665	2.080	0.080	0.059	0.275	0.5
	Attention + ViT	0.863	0.665	2.040	0.079	0.057	0.271	0.5
Augment.	ViT + Color	0.899	0.692	1.969	0.068	0.054	0.269	0.0
	Color + Rotations	0.870	0.682	2.050	0.078	0.055	0.271	0.2
	Rotations + Distortion	0.876	0.699	2.001	0.074	0.053	0.268	0.0
	Distortion + Erasing	0.893	0.727	1.941	0.067	0.052	0.265	0.0

Table 1. Results of our ablation study, across a variety of metrics. *Art.* refers to our artifact detection metric. We use a color code to highlight best and worst cases.

$\mathcal{L}_{style}$ and $\mathcal{L}_{freq}$ yield significant improvements, most notably in the normal map. To the baseline U-Net trained with the full loss, adding residual connections and one decoder per map increases accuracy. However, this setup tends to produce artifacts as it struggles to integrate global information. By adding self-attention to the encoder, we remove the artifacts; and with the MobileViT [45], we achieve higher quality results. Our final model contains full data augmentation, which provides small gains on generalization.

5.4. Qualitative Analysis

We aim to understand which features are exploited by our model for making its predictions. In Figure 6, we show the embeddings of our generator using UMAP. Despite some overlap, it seems that our model learns to separate between material *families* (*e.g. leathers* from *wovens*). Interestingly, the *thickness* of the material is also a relevant
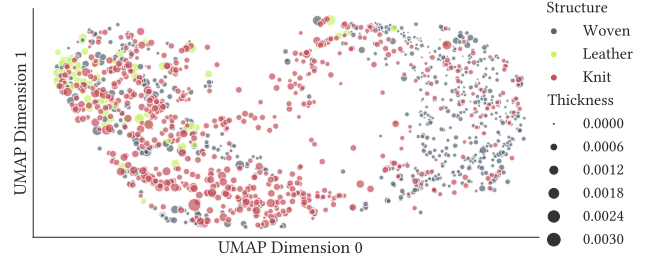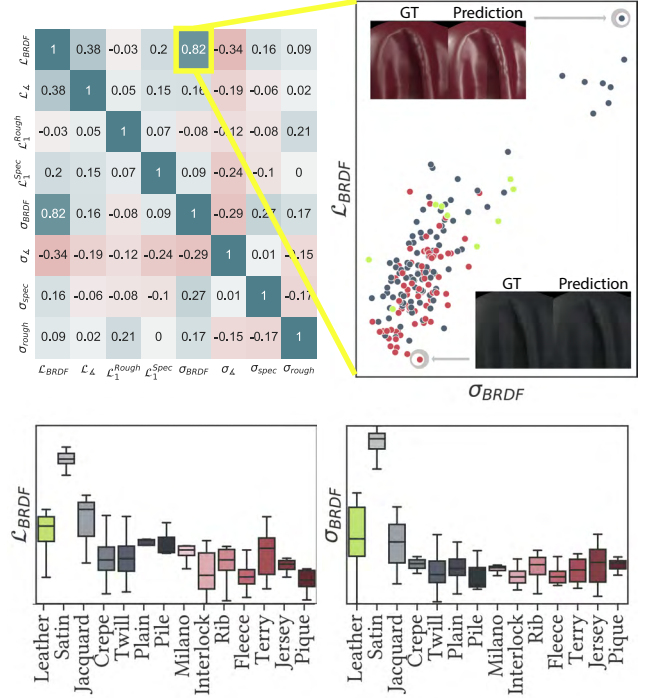

Figure 7. Top-left, correlation between render and pixel-wise losses, and uncertainties. Top-right, plot showing the correlation between our uncertainty metric and the error in render space; the renders illustrate the worst and best cases. Bottom, uncertainty and errors for different material families of the test set.

parameter. The model is exploiting these semantic patterns without explicit supervision, providing evidence that material microgeometry plays important an important role for its optical appearance.

5.5. Uncertainty Evaluation

In Figure 7, we show the Pearson correlation matrix between the uncertainty and error for each map, our uncertainty metric (Equation 4), and the render perceptual metric (Equation 5), on our test set. As shown, neither the error nor uncertainties per map explain errors on the render space. Our proposed σ_{BRDF} achieves a remarkable correlation of 82% with the render error, validating that this metric is useful to predict errors at test time with reasonably high precision. In the same plot at the bottom, we distill the un-
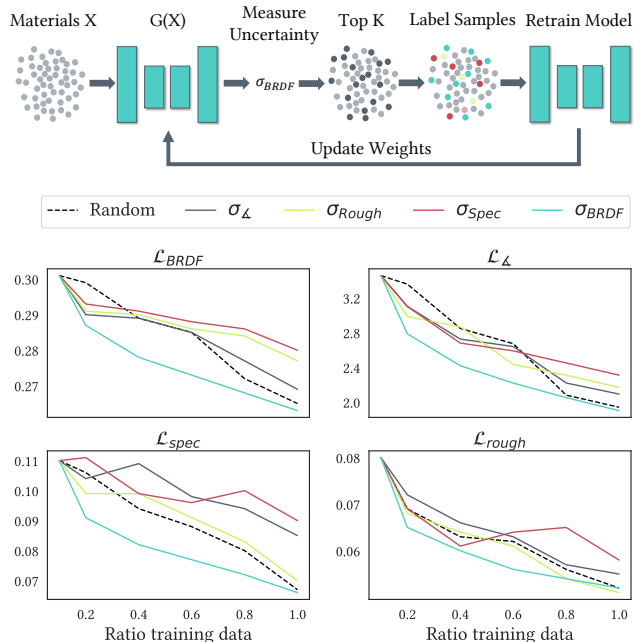
Figure 8. On top, illustration of our active learning algorithm. On the bottom, results of our active learning experiments. Leveraging σ_{BRDF} for actively selecting the top-k samples with the highest uncertainty, we achieve better results than a random sampling strategy for every metric we measure.

certainty and render error per family, in which we observe that our model struggles to accurately and confidently predict the reflectance of *satins*, *jacquards*, and *leathers* more than it does for any other structure in our dataset. These structures have complex optical behaviour that make their digitization more challenging (for example, satins exhibit anisotropy, which we do not support in our material model), and are relatively uncommon in our training dataset. Figures 1 and 4 show further examples of our uncertainty estimation for a diverse set of materials.

5.6. Active Learning

We leverage σ_{BRDF} for active learning [60] to identify the samples that contribute most to reduce the error of the model. Figure 8 (top) illustrates the process. We start by training a model with 10% of the available training data and measure the uncertainties in the remainder of the dataset. We then select the samples with the highest uncertainties and re-train the model with 20%. We repeat the process with 40, 60, 80 and 100% of the training dataset. For each subset, we compare the performance of this model with four baselines: random-sampling, sampling by the highest uncertainty in normals, roughness, and specular. The results are shown in Figure 8 (bottom). Active sampling based on σ_{BRDF} provides significant gains on sample efficiency, obtaining better accuracy for every map. For instance, an actively trained model with our metric which uses 20% of the training data obtains comparable results to a model trained

Method	Size (MB)	Output Dims	Eval Time (s)
Deep Inverse Rendering* [11]	167	256x256	603.5
Generative Modeling* [21]	1095.3	512x384	218.8
Diff. Material Graphs* [59]	-	512x512	1209.8
Adversarial Estimation [74]	11552.2	256x256	0.078
UMat (Ours)	**22.6**	256x256 512x512	**0.036** **0.131**

Table 2. Model sizes for different methods, and their evaluation time in seconds (average of 100 evaluations on an RTX 2080 GPU), for different output sizes. The methods with * perform test-time optimization. DiffMat [59] does not use a pre-trained model.

on three times more (but randomly sampled) data. While $\sigma_{\angle}$ is typically more informative than σ_{rough} and σ_{spec}, using per-map uncertainties does not provide better results than a random strategy. Finally, Figure 1 shows the variation of the probabilistic samples with respect to the ground truth radiance for two materials with high and low uncertainty.

5.7. Comparisons with Previous Work

In Table 3, we compare our method with previous work on single image material capture. First, we made sure that the training data for these methods included textile materials similar to the ones we choose for testing. Emulating their capture conditions, we took images with a smartphone, with flash and ambient lighting. Note that this capture conditions are not ideal for our method, affecting the final renders if the albedo has shading gradients. However, our goal in this experiment is to evaluate the overall preservation of the material structure in the inferred maps, particularly visible in the normals. For Shi *et al.* [59], we initialize the graph using a fabric material, provided in their open-source implementation and include the metallic map. Our model provides sharper and more accurate estimations without requiring optimization during test. Methods trained on style losses [21,59] degrade the semantic structure, while Zhou *et al.* [74] generate similar albedos to ours (note that ours are captured), but provide over-smooth estimations. We provide a comparison of timings and model sizes in Table 2. With our efficient model design, we can provide real time estimations without needing any optimization, which also enables our sampling-based uncertainty metric. Besides, it can also handle larger resolutions than previous work. Further results and re-rendered images are included in the supplementary material.

5.8. Limitations

We show some limitations in Figure 9. The illumination in the scanner hides the wrinkles in the *seersucker* fabric, and our model predicts a flat surface. The *organza* fabric at the bottom is very transparent with visible holes between the yarns. Since the background is white, the model has mistakenly treated the light regions as yarn centers. For the

Table 3. Comparisons of our method with previous work on images captured under different lighting conditions. Top: a smartphone flash-lit image. Middle: a smartphone image with ambient light. Bottom: a flatbed scanner capture. Our method produces the best results preserving the microstructure even when capture conditions degrade due to the sensor resolution. Note that we do not estimate albedos and that absolute intensities for specular and roughness maps are not directly comparable due to differences in the material model.
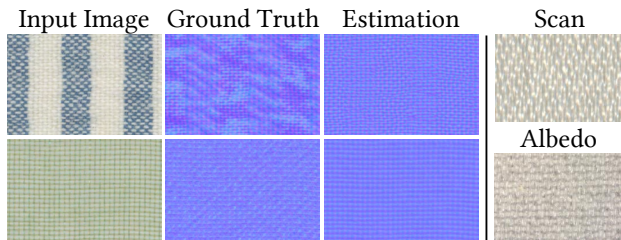

Figure 9. Limitations cases of our method. On the left, we show a *seersucker*, with wrinkles that are hidden by the diffuse illumination of the device, and a translucent *organza* with holes between the yarns that appear very bright due to the white background of the scanner, and are therefore mistakenly treated as yarn centers. On the right, we show that for highly directional materials, such as *satins*, the diffuse-like illumination in our capture device sometimes introduces specular highlights.

satin at the right, the scan image exhibits specular highlights due to the directionality of the yarns. While this image may be problematic to use as an albedo, it does not affect our metrics as we use constant albedos to compute them.

6. Conclusions

We have presented a GAN-based method to digitize materials which leverages microgeometry appearance and a flatbed scanner as capture device. Our method has shown better performance than state-of-the-art solutions that require a single image as input, when it comes to textile materials. To account for potential ambiguities derived from the capture setting, we have presented a method to model the uncertainty in the estimation at test time.

Managing uncertainty in machine learning projects is important to guarantee robust and functional solutions. However, this typically comes at the cost of complex or slow models. In this work, we have presented the first method to quantify uncertainty in single image material digitization, while introducing minimal impact in the training and evaluation processes. While it is currently not possible to discern the source of the uncertainty, whether it is epistemic (uncertainty which can be reduced by increasing the dataset size) or aleatoric (which is derived by a noisy data generation process), our metric has proven useful to identify ambiguous inputs, underrepresented classes, or out-of-distribution data.

We could extend our work in several ways. The most obvious extension is to estimate real albedos, so that we can deal with other type of scanning devices. Further, expanding our material model to give support for more reflectance properties, such as transmittance or anisotropy, could be useful to improve the realism in render of textiles. Finally, we will keep growing our dataset according to our active sampling process to add more families.

References

[1] Miika Aittala, Timo Aila, and Jaakko Lehtinen. Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 2

[2] Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. Practical svbrdf capture in the frequency domain. *ACM Transactions on Graphics (ToG)*, 32(4):110–1, 2013. 4

[3] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. 3

[4] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13349–13358, 2021. 4

[5] Brent Burley. Physically-based shading at disney. In *In SIGGRAPH Courses: Practical Physically Based Shading in Film and Game Production*, 2012. 3

[6] Thomas Chambon, Eric Heitz, and Laurent Belcour. Passing multi-channel material textures to a 3-channel loss. In *ACM SIGGRAPH 2021 Talks*, pages 1–2. 2021. 4

[7] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)*, 37(4):1–15, 2018. 2, 3

[8] Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3d shape and svbrdf acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15567–15576, 2021. 3

[9] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 3

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016. 2, 3, 4

[11] Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images. *ACM Transactions on Graphics (ToG)*, 38(4):134:1–134:15, July 2019. 2, 7, 8

[12] Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno. A Survey on Intrinsic Images: Delving Deep into Lambert and Beyond. *International Journal of Computer Vision*, 2022. 3, 4

[13] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in Neural Information Processing Systems*, 28, 2015. 4

[14] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 2, 4

[15] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 3

[16] Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. Highlight-aware two-stream network for single-image svbrdf acquisition. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2

[17] Yu Guo, Miloš Hašan, Lingqi Yan, and Shuang Zhao. A bayesian inference framework for procedural material parameter estimation. In *Computer Graphics Forum*, volume 39, pages 255–266. Wiley Online Library, 2020. 3

[18] Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. Materialgan: reflectance capture using a generative svbrdf model. *ACM Transactions on Graphics (TOG)*, 39(6):1–13, 2020. 1

[19] Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*, 2022. 4

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[21] Philipp Henzler, Valentin Deschaintre, Niloy J Mitra, and Tobias Ritschel. Generative modelling of brdf textures from flash images. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6), 2021. 2, 7, 8

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[23] Dongting Hu, Liuhua Peng, Tingjin Chu, Xiaoxing Zhang, Yinian Mao, Howard Bondell, and Mingming Gong. Uncertainty quantification in depth estimation via constrained ordinal regression. 2022. 3, 4

[24] Yiwei Hu, Chengan He, Valentin Deschaintre, Julie Dorsey, and Holly Rushmeier. An inverse procedural modeling pipeline for svbrdf maps. *ACM Transactions on Graphics (TOG)*, 41(2):1–17, 2022. 3

[25] Zhiyuan Huang, Mansur Arief, Henry Lam, and Ding Zhao. Evaluation uncertainty in data-driven self-driving testing. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1902–1907. IEEE, 2019. 3

[26] Inseung Hwang, Daniel S Jeon, Adolfo Muñoz, Diego Gutierrez, Xin Tong, and Min H Kim. Sparse ellipsometry: portable acquisition of polarimetric svbrdf and shape with unstructured flash photography. *ACM Transactions on Graphics (TOG)*, 41(4):1–14, 2022. 4

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 4, 5

[28] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13919–13929, 2021. 4

[29] Wenhua Jin, Beibei Wang, Milos Hasan, Yu Guo, Steve Marschner, and Ling-Qi Yan. Woven fabric capture from a

single photo. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 3

[30] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022. 3

[31] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013. 3

[32] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017. 3

[33] Michael C Krygier, Tyler LaBonte, Carianne Martinez, Chance Norris, Krish Sharma, Lincoln N Collins, Partha P Mukherjee, and Scott A Roberts. Quantifying the unknown impact of segmentation uncertainty on image-based simulations. *Nature Communications*, 12(1):1–11, 2021. 3

[34] Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrtens, Eva Krieghoff-Henning, Achim Hekler, Jakob Nikolas Kather, Stefan Fröhling, Christof von Kalle, Titus Josef Brinker, et al. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*, 10(8):e36427, 2022. 3

[35] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017. 3, 4

[36] Guillaume Lavoué, Nicolas Bonneel, Jean-Philippe Farrugia, and Cyril Soler. Perceptual quality of brdf approximations: dataset and metrics. In *Computer Graphics Forum*, volume 40, pages 327–338. Wiley Online Library, 2021. 5

[37] Zhao Lei, Yi Zeng, Peng Liu, and Xiaohui Su. Active deep learning for hyperspectral image classification with uncertainty learning. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 3

[38] Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 36(4):1–11, 2017. 2

[39] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: Svbrdf acquisition with a single mobile phone image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 72–87, 2018. 2

[40] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2, 4

[41] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W Jacobs. Shape and material capture at home. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6123–6133, 2021. 4

[42] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *Computer Graphics Forum*, volume 40, pages 101–113. Wiley Online Library, 2021. 4

[43] Morteza Mardani, Guilin Liu, Aysegul Dundar, Shiqiu Liu, Andrew Tao, and Bryan Catanzaro. Neural ffts for universal texture image synthesis. *Advances in Neural Information Processing Systems*, 33:14081–14092, 2020. 4

[44] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 6

[45] Sachin Mehta and Mohammad Rastegari. Mobilevit: lightweight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 2, 4, 6

[46] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pages 2391–2400. PMLR, 2017. 3

[47] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Mueller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. *arXiv:2111.12503*, 2021. 4

[48] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 3

[49] Jannik Boll Nielsen, Henrik Wann Jensen, and Ravi Ramamoorthi. On optimal, minimal brdf sampling for reflectance acquisition. *ACM Transactions on Graphics (TOG)*, 34(6):1–11, 2015. 5

[50] Carlos Rodriguez-Pardo and Elena Garces. Neural photometry-guided visual attribute transfer. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 4, 5

[51] Carlos Rodriguez-Pardo and Elena Garces. SeamlessGAN: Self-Supervised Synthesis of Tileable Texture Maps. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 3, 4

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 4

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 2, 3, 5

[54] Daniel B Russakoff, Carlo Tomasi, Torsten Rohlfing, and Calvin R Maurer. Image similarity using mutual information of regions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–607. Springer, 2004. 5

[55] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 4

[56] Shen Sang and Manmohan Chandraker. Single-shot neural relighting and svbrdf estimation. In *Proceedings of the Eu-*

ropean Conference on Computer Vision (ECCV), pages 85–101. Springer, 2020. 2

[57] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2020. 2, 4, 5

[58] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018. 3, 4

[59] Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik. Match: differentiable material graphs for procedural material capture. *ACM Transactions on Graphics (TOG)*, 2020. 3, 7, 8

[60] Ava P Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N Bhatia, and Connor W Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 7(8):1356–1367, 2021. 3, 7

[61] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 3

[62] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sỳkora. Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics (TOG)*, 39(4):73–1, 2020. 4

[63] Giuseppe Vecchio, Simone Palazzo, and Concetto Spampinato. Surfacenet: Adversarial svbrdf estimation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12840–12848, 2021. 1, 2, 4, 5

[64] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 195–206, 2007. 3

[65] Chen Wang, Xiang Wang, Jiawei Zhang, Liang Zhang, Xiao Bai, Xin Ning, Jun Zhou, and Edwin Hancock. Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition*, 124:108498, 2022. 4

[66] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2, 4

[67] Tao Wen, Beibei Wang, Lei Zhang, Jie Guo, and Nicolas Holzschuch. Svbrdf recovery from a single image with highlights using a pre-trained generative adversarial network. In *Computer Graphics Forum*. Wiley Online Library, 2022. 2

[68] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 4

[69] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 3

[70] Wenjie Ye, Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. Single image surface appearance modeling with self-augmented cnns and inexact supervision. In *Computer Graphics Forum*, volume 37, pages 201–211. Wiley Online Library, 2018. 2

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[72] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 4

[73] Xilong Zhou, Miloš Hašan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Kalantari. Tilegen: Tileable, controllable material generation and capture. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2022. 1

[74] Xilong Zhou and Nima Khademi Kalantari. Adversarial single-image svbrdf estimation with hybrid training. In *Computer Graphics Forum*, volume 40, pages 315–325. Wiley Online Library, 2021. 2, 3, 7, 8