Eurographics Symposium on Rendering (2025)
B. Wang and A. Wilkie (Editors)

An evaluation of SVBRDF Prediction from Generative Image Models for Appearance Modeling of 3D Scenes

A. Gauthier[1], V. Deschaintre[2], A. Lanvin[1], F. Durand[3], A. Bousseau[1], and G. Drettakis[1]

[1]Inria & Université Côte d'Azur, France [2]Adobe Research, UK [3]MIT, USA

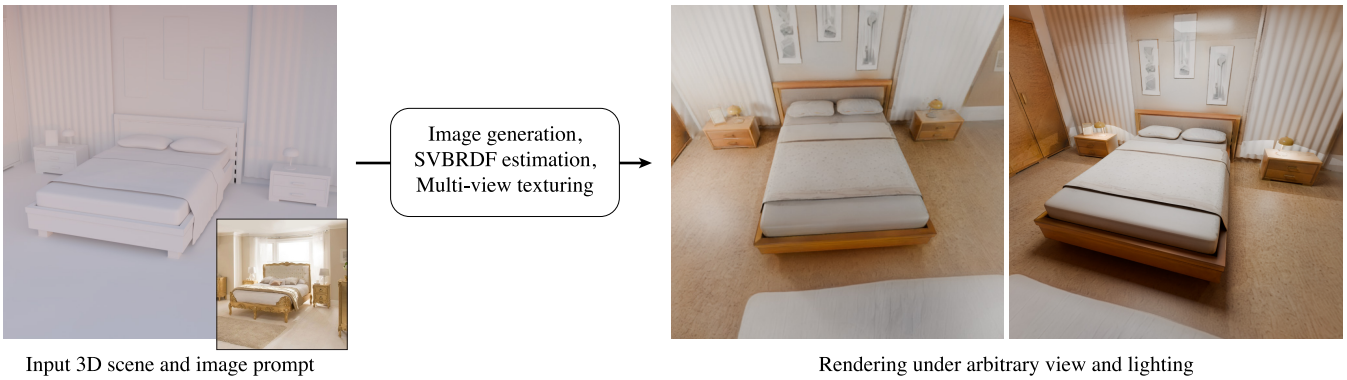

Figure 1: *We evaluate design choices for implementing a fast pipeline to model appearance of 3D scene assets. Given the geometry of the scene and an example image (left), we use generative image diffusion models and SVBRDF predictors to obtain multiview physically-based material maps that are merged into a texture atlas for the scene, enabling rendering under arbitrary view and lighting (right, see accompanying video for animated results). We evaluate how different single-image SVBRDF predictors – originally developed to process photographs – perform in this generative context, both in terms of per-view accuracy and in terms of multiview coherence.*

Abstract

Digital content creation is experiencing a profound change with the advent of deep generative models. For texturing, conditional image generators now allow the synthesis of realistic RGB images of a 3D scene that align with the geometry of that scene. For appearance modeling, SVBRDF prediction networks recover material parameters from RGB images. Combining these technologies allows us to quickly generate SVBRDF maps for multiple views of a 3D scene, which can be merged to form a SVBRDF texture atlas of that scene. In this paper, we analyze the challenges and opportunities for SVBRDF prediction in the context of such a fast appearance modeling pipeline. On the one hand, single-view SVBRDF predictions might suffer from multiview incoherence and yield inconsistent texture atlases. On the other hand, generated RGB images, and the different modalities on which they are conditioned, can provide additional information for SVBRDF estimation compared to photographs. We compare neural architectures and conditions to identify designs that achieve high accuracy and coherence. We find that, surprisingly, a standard UNet is competitive with more complex designs.

CCS Concepts

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

1. Introduction

Modeling appearance for 3D scenes is one of the hardest steps in 3D content creation, requiring painstaking manual selection and tweaking of material maps that encode the spatially-varying parameters of BRDFs (albedo, roughness, metallic). While generative diffusion models have demonstrated potential for the related tasks of image creation [SME21, HJA20] and texturing [RMA*23, CSL*23], the very large datasets of images they require for training specialize them to the RGB domain. In parallel, the field of single-image material estimation has matured to offer solutions for

estimating SVBRDFs at scene scale [LSR*20, KSN24, ZDG*24]. We observe that image generators can be combined with material estimators to form a fast pipeline for appearance modeling of 3D scenes (Figure 1). In this paper, we study the design choices that underpin such a pipeline and compare different single-image SVBRDF predictors when applied to multiple generated images of a scene, seeking for both accurate per-view estimation and coherence over multiple views, since these two properties are critical for merging multiview predictions into an SVBRDF texture atlas.

Our study is motivated by recent work on object texturing using 2D diffusion models [RMA*23, CSL*23]. Combining these methods with single-image SVBRDF estimation, we design a pipeline that takes as input the geometry of a 3D scene and outputs an SVBRDF texture atlas for that scene, which we use to evaluate the various design choices involved. This pipeline is controllable, leveraging conditional image generation [ZRA23a] and image reprojection and inpainting [ALF22] to iteratively generate multiple views of the input scene conditioned on its geometry and user-provided text or image prompts. Complementing multiview image generation with per-view material estimation allows us to obtain SVBRDF maps that we project onto a texture atlas to render the scene from arbitrary viewpoints and under arbitrary lighting.

The main contribution of our work resides in the exploration of various design choices for this material generation pipeline. Specifically, we compare state-of-the-art SVBRDF estimation methods in terms of both single-image accuracy and multiview coherence. Furthermore, we assess whether SVBRDF prediction improves when performed on deep features generated by the diffusion model [LDP*23] rather than on its RGB output. We also study the benefit of providing geometry information (depth and normal buffers) to the material estimation module. Surprisingly, this analysis reveals that a simple UNet that regresses SVBRDF maps is competitive with more complex alternatives.

In summary, this paper introduces two contributions:

- An analysis of the design space of single-image SVBRDF estimation for multiview material generation.
- Based on this analysis, a fast and controllable pipeline to generate SVBRDF textures over indoor scenes, combining generative image models and SVBRDF predictors.

We will release code, model weights and data upon publication.

2. Related Work

Our work leverages recent diffusion models for semi-automated visual content creation. We refer readers to the recent tutorial by Mitra et al. [MCP*24] and the report by Po et al. [PYG*24] for an overview of this very active field, and focus our discussion on the methods most related to our goal of generating SVBRDF textures over 3D scenes.

3D asset texturing. Automatic texturing of existing 3D assets has seen rapid progress with the introduction of high quality image generation methods [RBL*22, RDN*22]. Two main directions have been explored, based on either Score Distillation Sampling (SDS) [PJBM23] or on multiview texturing. SDS-based methods rely on the gradient provided by an image generative model to optimize a 3D representation of an object or scene geometry and/or texture [CLL*24, DOW*24]. Early versions of these methods rely on many diffusion steps of the generative model, often requiring dozens of minutes, while the more recent FlashTex [DOW*24] proposed a hybrid approach leveraging strong initialization and neural hashgrids for faster optimization, requiring only 6 minutes for texturing an object. Still, SDS-based approaches remain slower than multiview texturing methods [ZCQ*24, RMA*23, CSL*23, ZPZ*24, CDG*24, PWMAZ24, ZXS*25, WXM*24, LXLW24] that generate views around an object conditioned on its geometry, and project these images over the object to form a texture atlas. Such methods have so far been mostly demonstrated on isolated objects to create RGB textures. We build on these methods to design a pipeline that works on 3D scenes, and we augment RGB image generation with material prediction to produce SVBRDF atlases.

Several approaches also extended Score Distillation Sampling [ZLX*24, YOPM24, YHK*24] and multiview texturing [ZPX*24, CDG*24, FSW*24] to create materials over 3D objects. The former family of methods relies on expensive inverse rendering to optimize material parameters such that they produce images with a target appearance. In contrast, the latter family relies on retrieval in a library of procedural materials to assign SVBRDFs to segments in the generated RGB images. We share the same goal of augmenting generated images with material information, but we evaluate the performance of SVBRDF prediction rather than material retrieval in the context of multiview texturing. Furthermore, the above methods target isolated objects while we target extended indoor scenes that exhibit significant occlusions across views. Recent concurrent work [VKDN*24, HWLW24] generate multi-view material maps but also target isolated objects.

Finally, several methods have been proposed to automatically generate 3D objects or scenes, including their geometry and textures, from a single image or text prompt. Such methods rely on SDS-based optimization of shape and colors [PJBM23, LGT*23], on retrieval from datasets of 3D objects and materials [YLH*23], on iterative generation of multiview-consistent images/depth/material maps [TZC*23, HCO*23, LWH*23, ZWZ*24], or on volumetric generation [HZG*23, WZB*24]. While we focus on the different scenario of applying materials on an existing scene, we share their motivation for providing high-level controls on content creation, and some of their strategies for iterative generation of multiview-consistent images [HCO*23].

Predicting SVBRDFs. Predicting reflectance of surfaces from photograph(s) has been a long-standing challenge in Computer Graphics and Vision. When many (consistent) views are available, inverse rendering aims at reconstructing material properties through (gradient-based) optimization [JSR*22, LADL18, RH01]. Alternatively, deep neural networks have been used to predict material parameters for acquisition with limited signal, or when the signal is not guaranteed to be consistent across views – as is our case with generated images. Various neural methods were proposed for acquisition on flat surfaces [DAD*18, DAD*19, GSH*20, VMR*23], objects [VBP*24, LXR*18, DLG21, HGY*24] and, more recently, entire scenes from a single view [LSR*20, KSN24, ZDG*24, WTC*25] or multiple views [CLP*23, LGL*25,

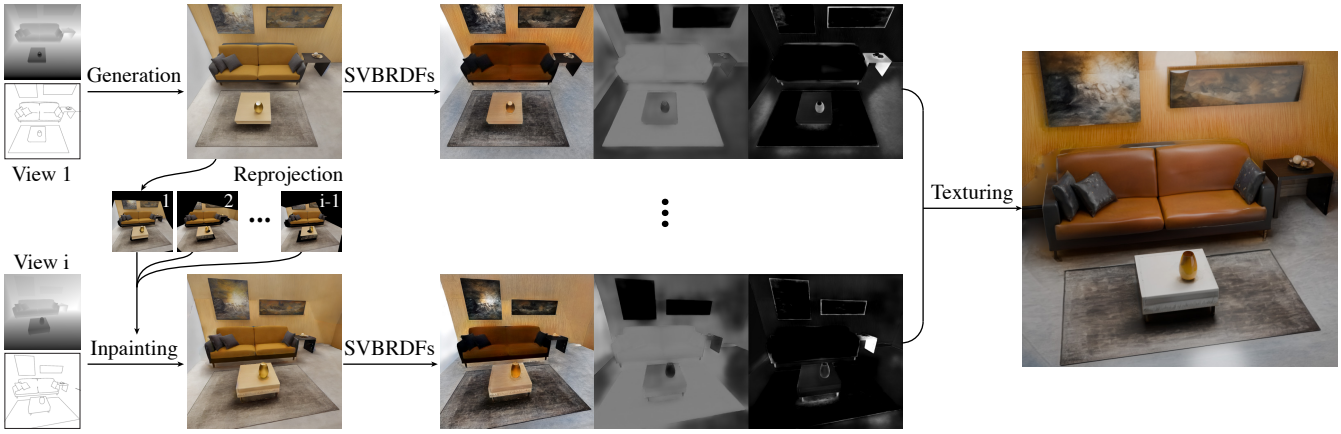

Figure 2: ***Overview of the SVBRDF texturing pipeline.*** *We first generate a sequence of views of the scene using an image diffusion model conditioned on depth and contours (Step 1). While the first view is generated entirely (top left), subsequent views are filled in by re-projecting existing views and inpainting holes revealed by disocclusions (bottom left). We then estimate SVBRDF maps for each view (Step 2), and merge them to form a texture atlas for the scene, enabling physically-based rendering with diverse materials (Step 3).*

HWH*24]. We focus on single-image methods due to the lack of a public dataset with paired renderings, SVBRDF maps, and cameras from multiple views. The decomposition of scene reflectance and geometry has received significant attention in recent years, leveraging the progress of generative models through fine-tuning [KSN24, ZDG*24] or through manipulation and analysis of their internal features [CVW23, BMHF23, ZRL*23, LDW*24]. A key aspect of our work is to evaluate these different approaches in the context of material design for 3D scenes and to compare them in terms of decomposition accuracy and consistency between views.

3. A Multiview Pipeline for Fast Appearance Modeling of 3D Scenes

We first describe a fast pipeline allowing users to apply materials over the 3D geometry of an untextured indoor scene by specifying high-level design goals in the form of text prompts or example images. This pipeline will allow us to investigate the design choices required for this task.

We illustrate this pipeline in Figure 2. Most of its components operate in image space, benefiting from the complementary strengths of diffusion models for controllable generation of multiple views of the scene, SVBRDF predictors for material estimation within each view, and multiview reprojection for texturing the scene with the obtained SVBRDF maps. Importantly, each operation is fast to compute, allowing to generate materials over a scene in minutes. We describe the first and third steps of this pipeline – multiview image generation and multiview texturing – before discussing and evaluating important design choices for SVBRDF prediction in Section 4.

3.1. Multiview Image Generation

The pipeline that underpins our study takes as input the geometry of an indoor 3D scene and a sequence of viewpoints that cover that scene well. For all the results in the paper, we use five viewpoints. We assume the first view is provided by the user. We generate the other four views by an offset of 25 degrees in latitude and longitude. The first step consists in generating multiview-consistent images of that scene observed from the given viewpoints. Inspired by previous work on generative single-object texturing and novel view synthesis [RMA*23, CSL*23, PWMAZ24, KSV*23, HCO*23], we create consistent images of the scene by alternating image generation, image reprojection, and image inpainting (Figure 3).

Image generation. We initiate the process with the first viewpoint, from which we render depth and contour maps that we use to condition an image generative model – ControlNet with a Stable Diffusion backbone [ZRA23a] – to obtain a photorealistic image aligned with the scene geometry. We optionally allow users to further control the content of the generated image by providing a text prompt, or an example image that contains representative materials to be generated. In the case where an example image is provided, we use the IP-Adapter approach [YZL*23] to obtain a text embedding for that image, which can then be used as a condition for Stable Diffusion. We combine all these conditions using the Multi-ControlNet pipeline from the Diffuser library [vPPL*22].

We next project the generated image from the first viewpoint into the second one using their respective cameras and depth maps, which ensures that the portions of the scene observed in both viewpoints are consistent. We hypothesize that while image reprojection only provides approximate placement of highlights, this approximation provides sufficient visual cues to estimate intrinsic appearance maps (metallic, roughness, etc.) because these quantities depend more on the sharpness and contrast of highlights than on their precise position. Our experiments demonstrate that, while potentially counter-intuitive, the combination of simple reprojection and per-view SVBRDF estimation can suffice to produce coherent SVBRDFs, which can in turn be aggregated into a single texture

Figure 3: *We progressively generate multiple coherent views of the scene by iterating over image generation conditioned on geometry, image reprojection into a new viewpoint, and image inpainting.*

atlas for physically-based rendering of specular and glossy reflections.

However, reprojection leaves holes at disocclusions, especially in cluttered indoor scenes where visibility changes significantly between viewpoints. We fill in these holes using image inpainting, as described in the next paragraph. We repeat the reprojection and inpainting steps to progressively populate all viewpoints, using all images generated so far to create the image of the next viewpoint.

Image inpainting. Reprojection provides us with a partially-covered image in the new viewpoint, as well as an occlusion mask indicating the parts to be filled-in. We feed this information to a ControlNet trained to perform inpainting [ZRA23b], along with the same text prompt used to generate the first image, if any. We also provide a text embedding of the first generated image, which helps in generating similar content in inpainted regions. We further improve alignment of the inpainted image to the underlying scene geometry by also conditioning ControlNet with depth and contour maps, as done when generating the image of the first viewpoint. In our experiments, the holes to inpaint cover up to 25% of the image.

3.2. Multiview Texturing

The above iterative procedure generates one image for each input viewpoint while ensuring that the parts seen from several viewpoints are consistent. Feeding each such image to a SVBRDF predictor gives material parameter maps for all viewpoints, as described in the next section. The last step of our pipeline consists in merging this image-space information into a common, scene-space texture atlas.

In contrast to related work on object texturing that populates the texture atlas incrementally [RMA*23, CSL*23, PWMAZ24], we take inspiration from photogrammetry where all photographs are merged at once to select the best observations available for each texel. While several algorithms exist to perform this task, we use the one implemented in MeshLab [CCCS08] for its simplicity and speed. This method takes as input the 3D mesh of the scene, its texture coordinates, and the multiple images of the scene with their respective camera parameters. It then assigns values to each texel by blending its observations according to geometric and color criteria. Since the SVBRDF maps we want to merge contain different quantities (albedo, roughness, metallic), we apply the method on each quantity independently.

4. Evaluating SVBRDF Predictors for Multiview Appearance Modeling

The above pipeline for generative material modeling requires extracting material maps from multiple images created via conditional image diffusion. This material estimation needs to be both accurate and consistent to facilitate the subsequent texture merging process. We now describe the various design choices that emerge from this setup, and analyze the impact of these choices.

4.1. A Design Space of SVBRDF Predictors

Scope of the study. The pipeline we have described in the previous section generates multiple photorealistic images of the 3D scene. A first solution that comes to mind to estimate the materials in that scene is physically-based inverse rendering, i.e. optimizing for material parameters that best reproduce the multiple images [ALKN19, NDDJK21, WZY*23]. However, this solution is not practical because even though the pipeline generates plausible images, the lighting conditions are unknown, and might even be inconsistent due to the simple reprojection employed to iteratively generate the views, preventing the use of physically-based inverse rendering.

These considerations motivate our choice of focusing our study on single-view SVBRDF prediction methods, and assessing whether these methods produce material parameters of sufficient quality and coherence to be combined into a single texture atlas. Specifically, we conducted experiments on two key dimensions of the design space of SVBRDF predictors: the choice of neural architecture, and the choice of data channels provided as input to that architecture, as summarized in Figure 4. For each of these choices, we evaluate the per-view accuracy of the predicted SVBRDFs, as well as their coherence across views.

Choice of architecture. Several neural network architectures have been proposed over the past few years to predict SVBRDF maps from a single image of an indoor scene. Early work relies on convolutional neural networks to predict material maps from RGB images, trained on paired datasets of images and ground-truth material maps to minimize regression (MAE, MSE) and perceptual losses (VGG, LPIPS). We evaluate Zhu et al.'s MGNet [ZLH*22] as a representative architecture of this family, using their model pre-trained on their InteriorVerse dataset, which we use for all comparisons. We also evaluate the performance of a standard UNet trained on the same dataset, which we call the **UNet-RGB** architecture.

More recently, generative models, and in particular diffusion models, have been repurposed for SVBRDF estimation. A key difference with regression-based methods is that generative models treat the input RGB image as guidance within a stochastic generation process that seeks to capture the distribution of SVBRDF maps that best explain the image. As a result, these methods can output multiple plausible interpretations of a given input. We evaluate the methods of Kocsis et al. [KSN24] and Zeng et al. [ZDG*24] as representative works in this category, using their model pre-trained on InteriorVerse. We follow the recommendations of Kocsis et al. and average 10 predictions of their model to form their output, unless specified otherwise. Finally, we also evaluate the single-step

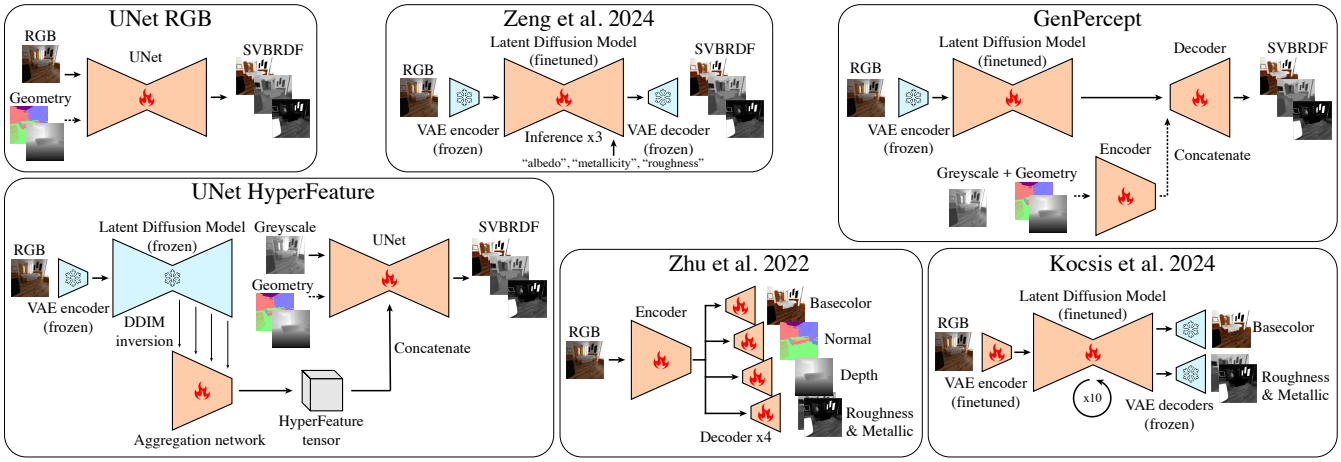

Figure 4: ***SVBRDF predictors.*** *We evaluate several architectures for SVBRDF predictions, taking as input an RGB image and optional geometry buffers (dotted lines).*

approach **GenPercept** by Xu et al. [XGL*25], which repurposes a diffusion model to perform dense perception tasks deterministically. We trained their method on InteriorVerse for our evaluation.

Choice of input channels. The architectures discussed above were developed to take photographs only as input. Yet, our target application – generative material design for 3D scenes – provides us with additional information, such as the scene geometry, that can help the SVBRDF predictor. We have experimented with several choices of additional information for the UNet and the GenPercept architectures that we retrain on InteriorVerse.

First, we can complement the RGB input with depth and normal maps to provide the neural networks with geometric information about the 3D scene, which we expect to help in recovering sharper maps along surface discontinuities, and in distinguishing shading from albedo variations.

Second, since the RGB images we take as input are the result of a generative process, we can complement each image with deep activation features extracted from the neural network that generated that image. Indeed, recent studies suggest that the feature maps produced by generative models of images carry semantic information about the content of the images they generate, allowing the recovery of semantic labels [BRV*21, ZLG*21], keypoint correspondences [LDP*23, PTL*23, HSM*24], depth, normal and albedo [CVW23, BMHF23, ZRL*23]. To evaluate whether such deep features benefit SVBRDF prediction, we follow the approach of Luo et al. [LDW*24, LDP*23] to extract so-called *hyperfeatures* from the denoising UNet of Stable Diffusion used to generate the first view or of the one used to inpaint the other views. We then concatenate these hyperfeatures with the feature maps of the bottleneck of the SVBRDF prediction UNet. We call this the **UNet-HF** (Hyperfeatures) architecture. While the Stable Diffusion activation features (which aggregate into a hyperfeatures tensor) are readily available for the images generated by the pipeline, they are not provided for the images of the InteriorVerse dataset we use for training. We solve this by performing inverse DDIM sampling [SME21] to "invert" each training image using Stable Diffusion, providing us with the UNet activation features corresponding to that image.

Implementation details. We base our implementation of the **UNet-RGB** architecture on the denoising diffusion UNet from the Diffusers library [vPPL*22], and our implementation of the **UNet-HF** architecture on the Readout Guidance [LDW*24] codebase for extracting hyperfeatures from Stable Diffusion's denoising UNet, using 11 timesteps and a projection dimension of 384 (as in [LDP*23]). For single-step diffusion, we follow the implementation of **GenPercept** provided by Xu et al. [XGL*25]. We provide additional details for each architecture in supplementary.

4.2. Accuracy and Coherence of SVBRDF Predictors

We investigate two main dimensions in the design space of SVBRDF generations for 3D scenes – the choice of *neural architecture* and the effect of different *inputs* for SVBRDF estimation. For each dimension, we evaluate the accuracy of SVBRDF predictions for single images as well as the coherence of the predictions over multiple views. We provide both quantitative and qualitative evaluation of both criteria.

Training data and procedure. We use the pre-trained models provided by Zhu et al. [ZLH*22], Kocsis et al. [KSN24] and Zeng et al. [ZDG*24], which have all been trained on the InteriorVerse dataset [ZLH*22]. In addition, we also train the UNet models and the single-step diffusion model [XGL*25] on InteriorVerse.

We conduct our evaluation on the InteriorVerse test dataset because it is the only public dataset containing the BRDF maps we are interested in. The dataset provides viewspace BRDF maps for basecolor (called "albedo"), metallic and roughness. The basecolor is modulated by the metallic, and encodes both diffuse albedo and specular color, as commonly used in physically-based pipelines [BS12]. As noted in RGB↔X [ZDG*24] the rendering images of InteriorVerse contain Monte-Carlo noise, which we denoise using

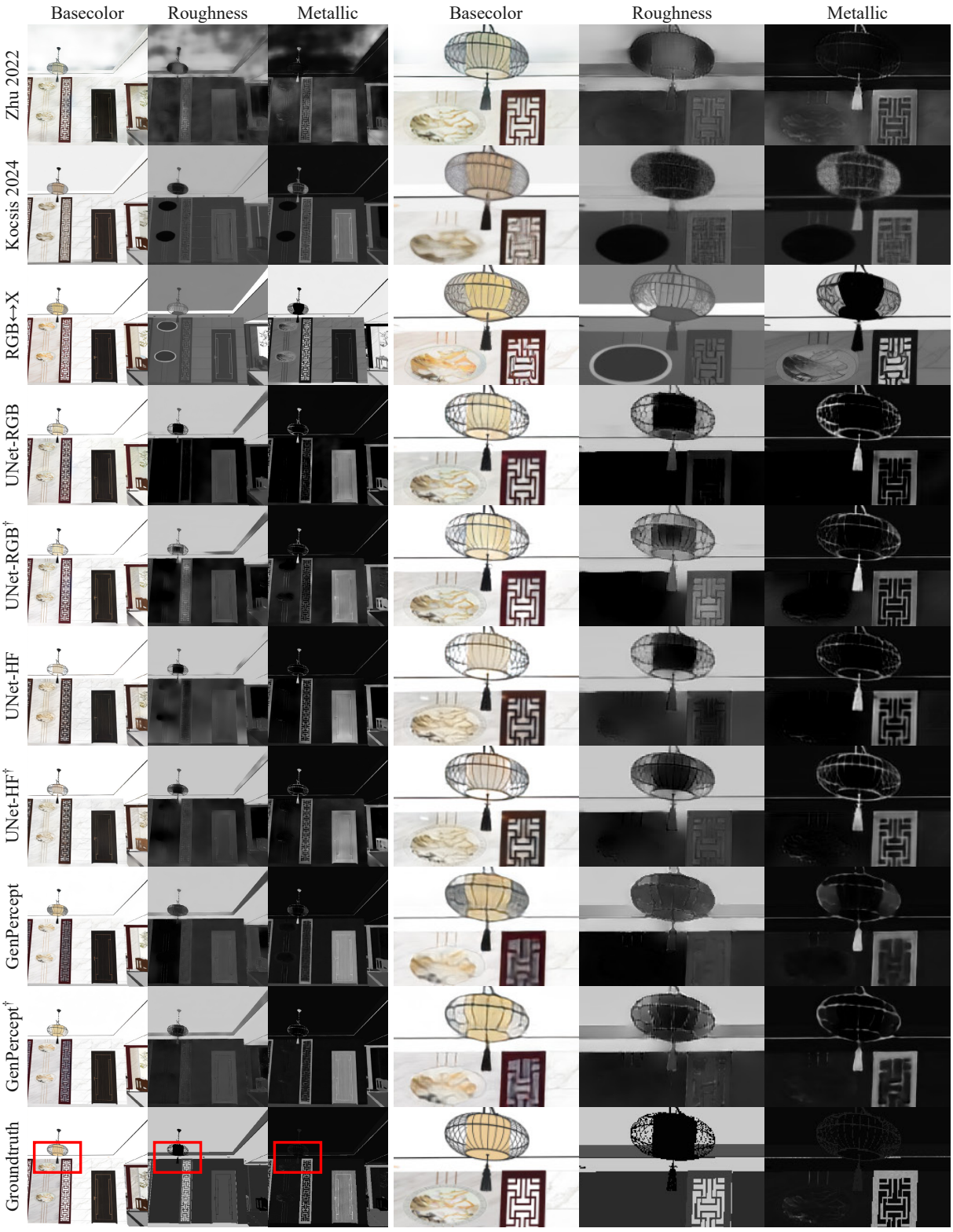

Figure 5: *Qualitative results of SVBRDF estimation for a test set image from Interiorverse, with closeups on the right.* †*Indicates training without normals and depth maps as input. Note how different methods provide a different tradeoff between accuracy, sharpness, and fine details.*

	BASECOLOR			ROUGHNESS			METALLIC		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Zhu2022 [ZLH*22]	17.11	0.668	0.229	13.38	0.348	0.422	14.93	0.676	0.419
Kocsis2024 [KSN24]	19.83	0.743	0.163	14.83	0.469	0.337	17.95	0.805	0.327
RGB↔X [ZDG*24]	20.35	0.763	0.131	11.38	0.220	0.401	6.68	-0.019	0.604
UNet-RGB†	20.26	0.772	0.126	16.92	0.642	0.265	19.52	0.834	0.274
UNet-RGB	21.66	**0.803**	0.109	17.73	0.672	0.241	20.13	0.843	0.256
UNet-HF†	20.74	0.775	0.139	17.20	0.642	0.269	19.82	0.830	0.280
UNet-HF	21.25	0.784	0.134	17.27	0.626	0.265	19.99	0.830	0.276
GenPercept†	**22.94**	0.799	**0.107**	**19.57**	**0.701**	**0.201**	**20.97**	**0.850**	0.256
GenPercept	22.67	0.793	0.113	19.26	0.689	0.202	20.83	0.847	**0.254**

Table 1: ***Quantitative analysis of SVBRDF estimation****. The metrics are computed over the InteriorVerse test dataset (2633 images), color coded between worst and best.* †*Indicates training without normals and depth maps as input.*

Figure 6: *When training using the FLIP loss colorspace [ANA*20], the basecolor converges faster to the right hue and luminance than when using a simple L1 loss in RGB colorspace. Note in particular the hue of the wall and floor. The results are shown after only a few epochs of training to emphasize differences.*

Mitsuba's Optix integration. We crop each image to get a square which we resize to 512^2 during training. Additionally, some images of the InteriorVerse dataset contain very few objects or very little detail, and are sometimes rendered from cameras pointing outside the scene. We remove such data (4% of the dataset) by removing files below 2MB. We will release this curated dataset upon publication, together with the code and network weights.

Regarding the training procedure of these predictors, we use a combination of L-PIPS loss [ZIE*18] $\mathcal{L}^{vgg}$ and L1 loss computed over the perceptual color space proposed by FLIP [ANA*20] $\mathcal{L}_1^{FLIP}$ for the basecolor, which we found to behave better than the L1 loss in RGB space (see Fig. 6). We use a regular L1 loss for both metallic and roughness. The training loss is thus the sum of three groups of terms corresponding to basecolor, metallic and roughness, respectively weighted by α_b, α_m, α_r:

$$\alpha_b(\mathcal{L}_1^{FLIP} + \lambda_b \cdot \mathcal{L}^{vgg}) + \alpha_m \cdot \mathcal{L}_1 + \alpha_r(\mathcal{L}_1 + \lambda_r \cdot \mathcal{L}^{vgg}), \quad (1)$$

where $\alpha_b = 1.0$, $\alpha_m = 2.0$, $\alpha_r = 0.5$, and $\lambda_b = \lambda_r = 0.5$, which we empirically find to work best.

We train each estimator with a maximum budget of 8 days using 4 GPUs (either RTX 6000s or RTX 8000s). We train all networks with an Adam optimizer and a learning rate of 10e-5.

Single-view prediction. We first study how the different architectures and inputs perform on a single-image prediction task.

We run each method on the InteriorVerse test set, composed of 2633 HDR images, which we process with the tone-mapping and normalization algorithms recommended by the authors of each method. As suggested by [KSN24], we lift the luminance ambiguity in the estimation by reporting scale-invariant metrics for the basecolor. This is computed by normalizing the output of each method by the ratio between the ground-truth and the predicted per-channel means. We report standard quantitative error metrics (PSNR, SSIM and LPIPS) in Tab. 1 and we provide representative qualitative comparisons in Fig. 5. We provide additional qualitative results in supplemental materials.

The simple UNet-RGB with geometry cues performs surprisingly well on all metrics, being only outperformed by GenPercept (irrespectively of the use of additional geometric information). The UNet-HF with geometry cues is placed third. From visual inspection (Fig. 5), we see that the method of Kocsis et al. [KSN24] and RGB↔X [ZDG*24] create smooth, piecewise-constant maps, as is the case for ground truth SVBRDFs. However, these piecewise constant maps are not always accurate, in particular for roughness and metallic maps where different objects made of the same materials are assigned different – often erroneous – values, as is the case for the insets in the corresponding rows of Fig. 5 where the roughness and metallic maps should be uniform over the wall rather than being correlated with the basecolor. MGNet [ZLH*22] tends to produce splotchy results in all maps, compromising quality. Finally, GenPercept [XGL*25] is the most accurate according to numerical metrics (Tab. 1), yet struggles to recover fine details, as is the case of the thin structures of the lamp and wall panel in Fig. 5. Overall, each method provides different tradeoffs between sharpness and accuracy, as illustrated by additional examples in the supplemental material. The use of additional geometric inputs tends to improve the overall accuracy of the predictors.

Multi-view consistency. Since our goal is to generate SVBRDF maps to texture a 3D scene, multi-view consistency is important to avoid artifacts in the merged texture atlas. Even if the SVBRDF maps of each image are plausible, if they are not consistent, seams

Methods	Consistency ↓ Basecolor	Roughness	Metallic	Runtime
[KSN24]$^{\times 1}$	0.281	0.227	0.219	$\simeq 6$ s
[KSN24]$^{\times 10}$	0.124	0.100	0.104	60.8 ± 1.3 s
[ZDG*24]	0.155	0.247	0.542	11.1 ± 0.2 s
[ZLH*22]	0.073	0.100	0.105	41 ± 5 ms
UNet-RGB	0.064	0.078	0.060	36 ± 1 ms
UNet-RGB†	0.089	0.117	0.077	36 ± 2 ms
UNet-HF	**0.056**	**0.050**	**0.041**	335 ± 34 ms
UNet-HF†	0.060	0.064	0.048	409 ± 12 ms
GenPercept	0.072	0.104	0.069	69 ± 4 ms
GenPercept†	0.075	0.108	0.069	74 ± 2 ms

Table 2: ***Consistency analysis.*** *Results are color coded between worst and best.* †*indicates training without normals and depth maps as input. The runtime is reported on an RTX 3090. For Kocsis et al.,* $^{\times N}$ *indicates the number of predictions that are averaged, since this stochastic method is capable of sampling multiple predictions for a given input. Note how feeding diffusion hyperfeatures to the UNet architecture greatly improves multi-view consistency compared to RGB input, especially for roughness and metallic.*

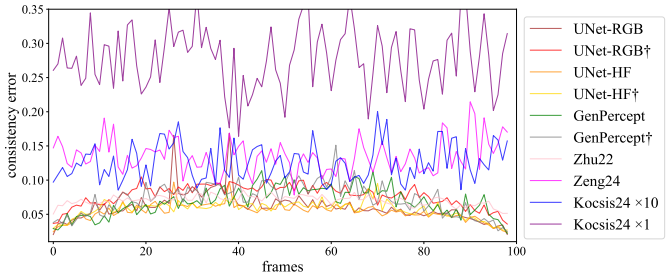

Figure 7: *We show the Stop-the-Pop [RSP*24] metric along a video of the predicted maps for the synthetic scene 'kitchen-005', for the basecolor, on each method. UNet-HF is the most consistent across views (orange and yellow).*

and blur may appear in the texture-space material maps, which may negatively affect rendering quality. Fig. 8 and 9 point to such seams in the merged texture atlas obtained with [KSN24] and [ZDG*24].

For a qualitative evaluation of multi-view consistency, we find that short video segments of a fly through of the maps using different methods are particularly informative. We include such videos in our supplemental materials. These show that UNet-HF results in the lowest level of flickering. We evaluate this quantitatively using the flickering metric proposed by Stop-the-Pop [RSP*24] that warps each frame of the video to its subsequent frame and computes the ꟻLIP [ANA*20] difference between the two images. As we have geometry and camera information, we use pixel-perfect depth maps and cameras for the warping, instead of an optical flow estimation as in Stop-the-Pop. We report in Tab. 2 the sum of ꟻLIP metrics over 5 synthetic scenes that are not part of InteriorVerse, with 100 views each, warping only successive frames to evaluate short range flickering. We visualize the metric over time for a given scene in Fig. 7.

This experiment reveals that the UNet-HF provides the best multi-view consistency overall. A possible interpretation of this result is that the deep features extracted from the image generation models have little dependence on viewpoint, yielding similar material values for objects seen in different views. Note also that averaging 10 predictions increases coherence for the method by Kocsis et al. [KSN24], but in our experiments we noticed that this tends to reduce overall contrast and details. The runtime column of Table 2 illustrates that regression models (bottom section) are much faster than generative models (top section) because they only need a single inference through the architecture (compared to 50 inferences for diffusion models). The UNet-HF architecture accumulates activation features during RGB image generation, and hence requires only a single inference pass through the SVBRDF estimator. As for accuracy, providing additional geometric inputs to the predictors consistently improves the consistency of the estimations.

Runtime analysis. Table 2 additionally provides the runtime of each predictor, generative methods on top and regression methods in the bottom. We see that the former [KSN24, ZDG*24] have runtimes of at least 6 seconds, making them unsuitable for a fast, interactive texturing pipeline. This is due to multiple factors: the sampling strategy which requires many iterative steps, the independent predictions of each channel for the method by Zeng et al. [ZDG*24], and the averaging of multiple outputs for Kocsis et al. [KSN24]. On the other hand, regression-based methods (in the bottom part) only require a single inference making them more suitable for a fast texturing pipeline.

5. Easy and Powerful 3D Scene Appearance Modeling

We now provide results of the multiview texturing pipeline described in Section 3, which leverages image generation and SVBRDF estimation for appearance modeling of 3D scenes. We adopted the **UNet-HF** SVBRDF predictor to produce these results since this is the design that achieves the highest multi-view consistency with competitive single-view accuracy.

Implementation. We based our implementation of multi-view image generation on the Stable Diffusion ControlNet and ControlNet Inpainting pipelines from the Diffusers repository [vPPL*22]. Specifically, we use Stable Diffusion v1.5 and the associated ControlNets (revision v1.1) [ZRA23b] for contour, depth, and inpainting. We generate images at 768^2 resolution to obtain fine details in the textures.

Results. Figures 11 , 12 and 13 show three scenes and different results conditioned either by text or image prompts. The generated SVBRDF atlases exhibit different materials, with a specular mineral floor and a rough carpet in Fig. 11. A golden vase and a dark plastic vase are also displayed. Fig. 12 shows white, blue, and black marble, and thin golden structures, a stone wall, and copper furniture. Finally, Fig. 13 shows a rough rug-like material and a specular white mineral for the floor. Additional texturing results are shown in the supplementary materials (file and video).

Being fast and image-based, this pipeline provides a convenient and easy way to design and iterate over the relightable appearance

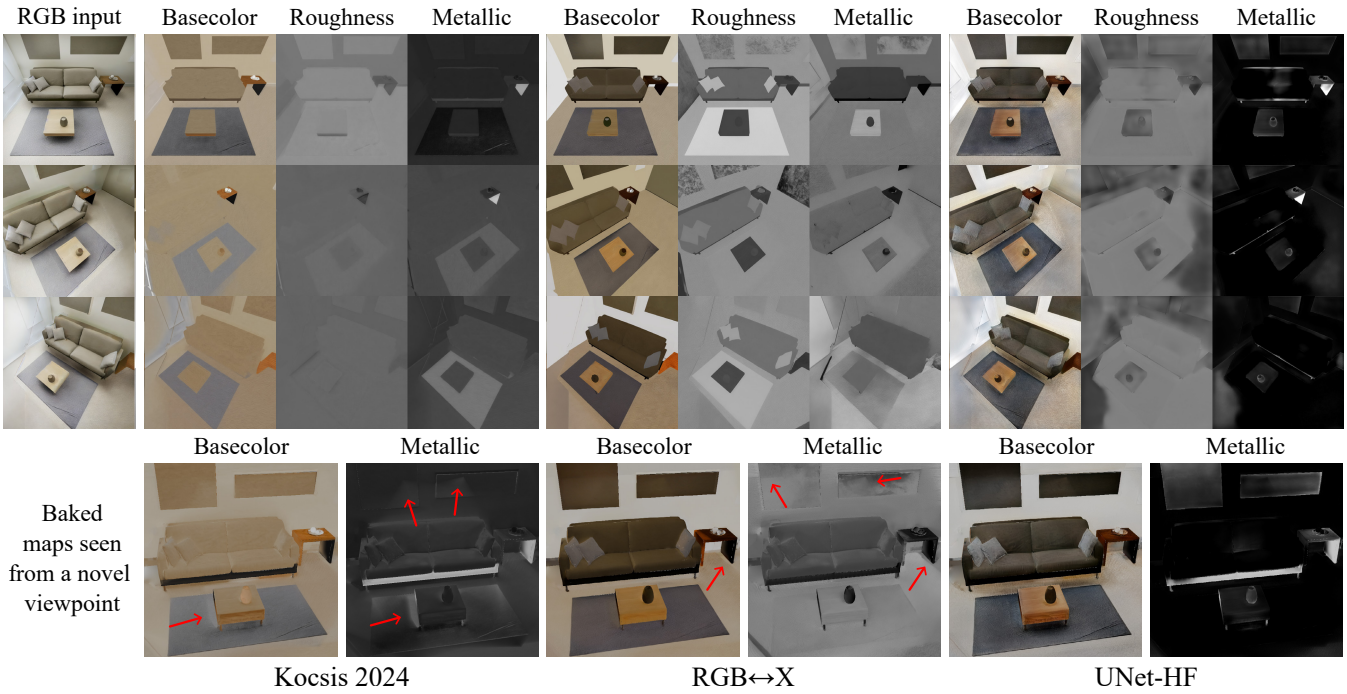

Figure 8: *Qualitative comparison of single-image SVBRDF estimations merged in a texture atlas. On the left we show the input of three SVBRDF estimation techniques (Kocsis 2024, RGB2X, UNet-HF). The top row shows the SVBRDf predictions produced for each image. While Kocsis 2024 and RGB2X produce sharp, piecewise constant maps, these maps are not consistent across views, resulting in visible seams when merged in a texture atlas and rendered from a novel viewpoint (bottom row, red arrows). The UNet-HF architecture achieves higher consistency across views, which reduces the presence of seams (right).*

of an untextured 3D scene geometry. Since the first RGB image is generated in a few seconds, the user can easily iterate over the appearance of the first view. Similarly, if the user is not satisfied by the inpainting in a view, they can regenerate that view (though we did not use this feature in our results). All relit results shown are computed from the final merged atlas, not from a single-view estimate. Note that the viewpoint used for relighting are different from the five views used for generation. We include further results that illustrate the benefits of our appearance modeling pipeline in our supplemental video.

6. Limitations and Future Work

We base our study on the InteriorVerse dataset which, while being the only one available, has limitations in terms of diversity and precision of the ground truth PBR materials. It contains a small number of materials for floor, walls and furniture, and metallic objects are rare. In addition the distributions of roughness and metallic values are uneven (99% of the roughness values are below 0.8). A richer training dataset should result in improved variety of generated materials.

We focused our study on single-image scene material estimation methods, which can still suffer from inconsistencies across viewpoints despite our careful choice of architecture and inputs. Multi-view SVBRDF extraction from photographs has been explored [DAD*19, CLP*23], but the generated images we are dealing with present illumination inconsistencies that might challenge such approaches.

We demonstrated the complementarity of image generation and material prediction by implementing a fast pipeline for appearance modeling of 3D scenes. This pipeline suffers from a few limitations. First, the reprojection may cause artifacts in the presence of thin objects and small geometry, as illustrated in Fig. 10. Second, as we rely on a diffusion model for appearance generation, and deep networks for material estimation, the expressivity and precision of our pipeline is limited by the quality of the models and training datasets. For example, we noticed that image inpainting sometimes produces visible seams, which might remain visible in the merged texture atlas. Finally, using our prototype, a few iterations of prompt engineering can be required to achieve the desired appearance, which could be improved with the better prompt adherence of recent diffusion models [PEL*23, PX22].

7. Conclusion

In this paper we study the complementarity of generative image models and deep material predictors by evaluating the different choices required in a pipeline for designing materials over 3D scenes. This pipeline takes the untextured geometry of a 3D scene as input and lets the designer provide additional conditions (e.g. text or image prompts). From this input, the pipeline quickly generates an SVBRDF texture atlas compatible with physically-based

Figure 9: *Rendered textured scenes using different SVBRDF estimators. [ZLH*22] produces sharp but splotchy maps, [KSN24] generates monotonic maps lacking details, while [ZDG*24] generates blurry maps with visible seams due to the inconsistent estimations. UNet-RGB, UNet-HF and GenPercept produce more contrasted results with reduced seams artifacts, providing tradeoffs in terms of sharpness, colors and consistency. Please see the accompanying video for an animated result.*

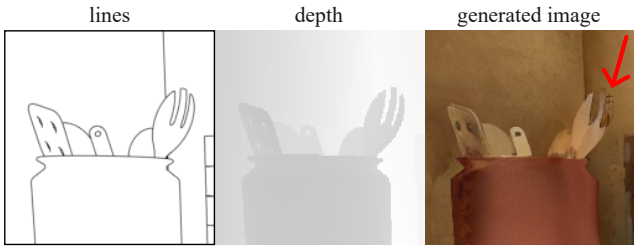

Figure 10: *Thin objects may cause artifacts due to reprojection, as shown in the right part of the wooden fork, which does not generate clean material and geometry.*

rendering. This simple pipeline enables rapid design iterations, and will directly benefit from the rapid progress in image generation control and quality.

We study the impact of the type and input of neural architectures for SVBRDF estimation in the context of this 3D material texturing pipeline. Surprisingly, we find that a simple single-pass UNet architecture can outperform more complex recent solutions in terms of accuracy of prediction, even though recent methods can provide appealing piecewise constant results. We also note that the use of hyperfeatures improves multi-view consistency, possibly because they encode richer semantic information that is more invariant to view changes than RGB values. Finally, we find that providing depth and normals as additional channels provides marginal results improvements.

8. Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work was funded by the European Research Council (ERC) Advanced Grant NERPHYS, number 101141721 https://project.inria.fr/nerphys. The authors are grateful to the OPAL infrastructure of the Université Côte d'Azur for providing resources and support, as well as Adobe and NVIDIA for software and hardware donations. This work was granted access to the HPC resources of IDRIS under the allocation AD011015561 made by GENCI. F. Durand acknowledges funding from from Google, Amazon, and MIT-GIST.

References

[ALF22] AVRAHAMI O., LISCHINSKI D., FRIED O.: Blended diffusion for text-driven editing of natural images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 2

[ALKN19] AZINOVIC D., LI T.-M., KAPLANYAN A., NIESSNER M.: Inverse path tracing for joint material and lighting estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 2447–2456. 4

[ANA*20] ANDERSSON P., NILSSON J., AKENINE-MÖLLER T., OSKARSSON M., ÅSTRÖM K., FAIRCHILD M. D.: ꟻLIP: A Difference Evaluator for Alternating Images. *Proceedings of the ACM on Computer Graphics and Interactive Techniques 3*, 2 (2020), 15:1–15:23. 7, 8

[BMHF23] BHATTAD A., MCKEE D., HOIEM D., FORSYTH D.: Stylegan knows normal, depth, albedo, and more. In *Advances in Neural Information Processing Systems* (2023). 3, 5

[BRV*21] BARANCHUK D., RUBACHEV I., VOYNOV A., KHRULKOV V., BABENKO A.: Label-efficient semantic segmentation with diffusion models, 2021. arXiv:2112.03126. 5

[BS12] BURLEY B., STUDIOS W. D. A.: Physically-based shading at disney. In *ACM Siggraph* (2012), pp. 1–7. 5

[CCCS08] CALLIERI M., CIGNONI P., CORSINI M., SCOPIGNO R.: Masked photo blending: mapping dense photographic dataset on high-resolution 3d models. *Computer & Graphics 32*, 4 (Aug 2008), 464–473. 4

[CDG*24] CEYLAN D., DESCHAINTRE V., GROUEIX T., MARTIN R., HUANG C.-H., ROUFFET R., KIM V., LASSAGNE G.: Matatlas: Text-driven consistent geometry texturing and material assignment, 2024. 2

[CLL*24] CHEN D. Z., LI H., LEE H.-Y., TULYAKOV S., NIESSNER M.: Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 21081–21091. 2

[CLP*23] CHOI J., LEE S., PARK H., JUNG S.-W., KIM I.-J., CHO J.: Mair: multi-view attention inverse rendering with 3d spatially-varying lighting estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), IEEE, pp. 8392–8401. 2, 9

[CSL*23] CHEN D. Z., SIDDIQUI Y., LEE H.-Y., TULYAKOV S., NIESSNER M.: Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396* (2023). 1, 2, 3, 4

[CVW23] CHEN Y., VIÉGAS F., WATTENBERG M.: Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720* (2023). 3, 5

[DAD*18] DESCHAINTRE V., AITTALA M., DURAND F., DRETTAKIS G., BOUSSEAU A.: Single-image svbrdf capture with a rendering-aware deep network. *ACM Trans. Graph. 37*, 4 (jul 2018). 2

[DAD*19] DESCHAINTRE V., AITTALA M., DURAND F., DRETTAKIS

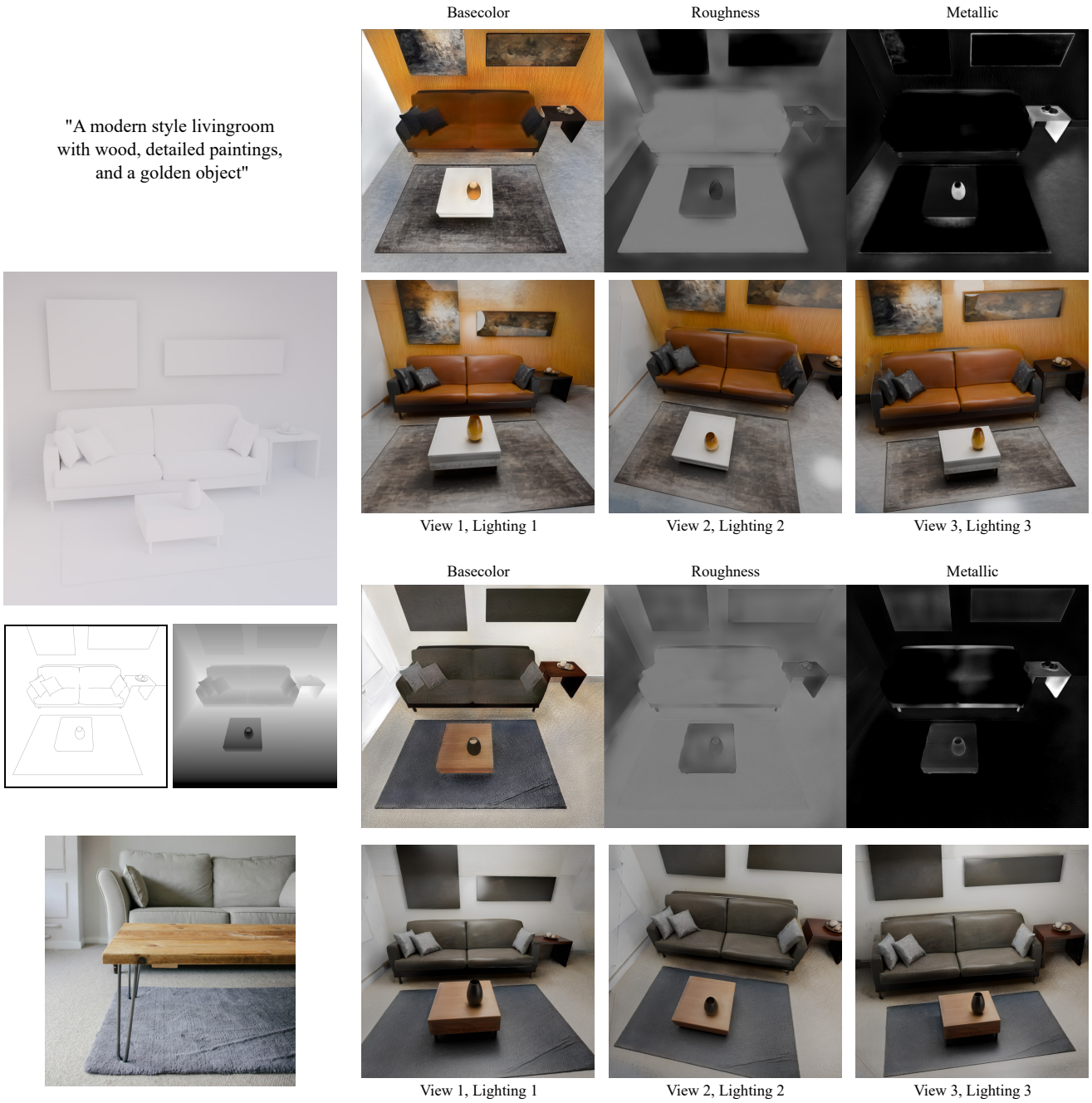

Figure 11: *An example of material design for a living room scene. On the left, we show the two different prompts as well as images used to design the appearance of the scene (the middle image is a shaded version of the input geometry). For each prompt we show the material maps extracted from the generated image, and below the maps three different viewing and lighting conditions (moving light source).*

G., BOUSSEAU A.: Flexible svbrdf capture with a multi-image deep network. *Computer Graphics Forum(Eurographics Symposium on Rendering Conference Proceedings) 38*, 4 (jul 2019), 13. URL: http://www-sop.inria.fr/reves/Basilic/2019/DADDB19. 2, 9

[DLG21] DESCHAINTRE V., LIN Y., GHOSH A.: Deep polarization imaging for 3d shape and svbrdf acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021). 2

[DOW*24] DENG K., OMERNICK T., WEISS A., RAMANAN D., ZHU J.-Y., ZHOU T., AGRAWALA M.: Flashtex: Fast relightable mesh texturing with lightcontrolnet. In *European Conference on Computer Vision*

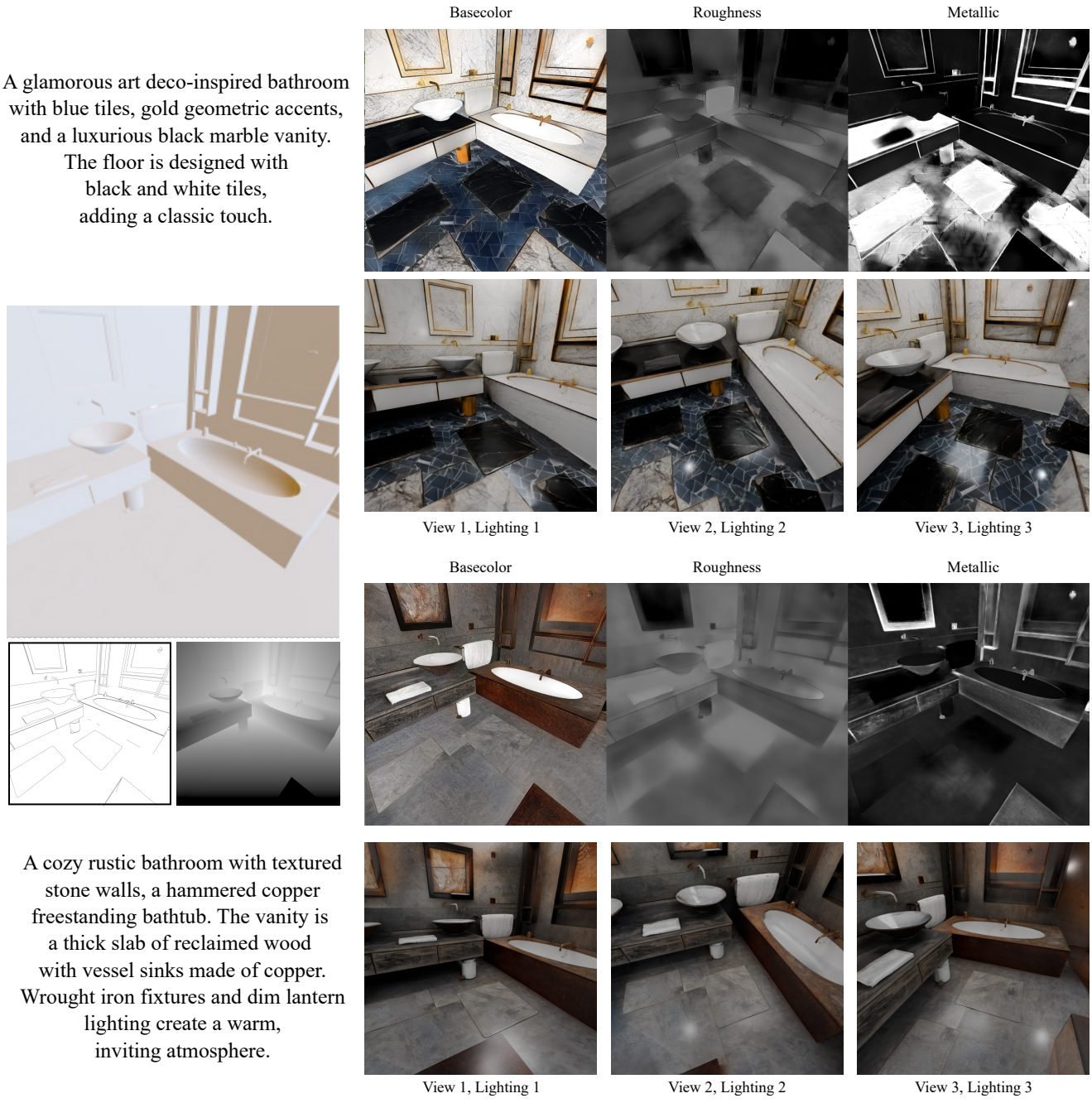

Figure 12: *An example of material design for a bathroom scene. On the left, we show the two different prompts used to design the appearance of the scene. For each prompt we show the material maps extracted from the generated image, and below the maps three different viewing and lighting conditions (moving light source).*

(ECCV) (2024). 2

[FSW*24] FANG Y., SUN Z., WU T., WANG J., LIU Z., WETZSTEIN G., LIN D.: Make-it-real: Unleashing large multimodal model for painting 3d objects with realistic materials, 2024. arXiv:2404.16829. 2

[GSH*20] GUO Y., SMITH C., HAŠAN M., SUNKAVALLI K., ZHAO S.: Materialgan: Reflectance capture using a generative svbrdf model. *ACM Trans. Graph. 39*, 6 (nov 2020). 2

[HCO*23] HÖLLEIN L., CAO A., OWENS A., JOHNSON J., NIESSNER M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023), pp. 7909–7920. 2, 3

Figure 13: *An example of material design for a bedroom scene. On the left, we show the two different image prompts used to design the appearance of the scene. For each prompt we show the material maps extracted from the generated image, and below the maps three different viewing and lighting conditions (moving light source).*

[HGY*24] HONG Y., GUO Y.-C., YI R., CHEN Y., CAO Y.-P., MA L.: Supermat: Physically consistent pbr material estimation at interactive rates. *arXiv preprint arXiv:2411.17515* (2024). 2

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems 33* (2020), 6840–6851. 1

[HSM*24] HEDLIN E., SHARMA G., MAHAJAN S., ISACK H., KAR A., TAGLIASACCHI A., YI K. M.: Unsupervised semantic correspondence using stable diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2024), NIPS '23, Curran Associates Inc. 5

[HWH*24] HE Z., WANG T., HUANG X., PAN X., LIU Z.: Neural

lightrig: Unlocking accurate object normal and material estimation with multi-light diffusion. *arXiv preprint arXiv:2412.09593* (2024). 2

[HWLW24] HUANG X., WANG T., LIU Z., WANG Q.: Material anything: Generating materials for any 3d object via diffusion. *arXiv preprint arXiv:2411.15138* (2024). 2

[HZG*23] HONG Y., ZHANG K., GU J., BI S., ZHOU Y., LIU D., LIU F., SUNKAVALLI K., BUI T., TAN H.: Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400* (2023). 2

[JSR*22] JAKOB W., SPEIERER S., ROUSSEL N., NIMIER-DAVID M., VICINI D., ZELTNER T., NICOLET B., CRESPO M., LEROY V., ZHANG Z.: Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 2

[KSN24] KOCSIS P., SITZMANN V., NIESSNER M.: Intrinsic image diffusion for indoor single-view material estimation. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 2, 3, 4, 5, 7, 8, 10

[KSV*23] KANT Y., SIAROHIN A., VASILKOVSKY M., GULER R. A., REN J., TULYAKOV S., GILITSCHENSKI I.: invs : Repurposing diffusion inpainters for novel view synthesis. In *SIGGRAPH Asia 2023 Conference Papers* (2023). 3

[LADL18] LI T.-M., AITTALA M., DURAND F., LEHTINEN J.: Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 37*, 6 (2018), 222:1–222:11. 2

[LDP*23] LUO G., DUNLAP L., PARK D. H., HOLYNSKI A., DARRELL T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems* (2023). 2, 5

[LDW*24] LUO G., DARRELL T., WANG O., GOLDMAN D. B., HOLYNSKI A.: Readout guidance: Learning control from diffusion features. 3, 5

[LGL*25] LIANG R., GOJCIC Z., LING H., MUNKBERG J., HASSELGREN J., LIN Z.-H., GAO J., KELLER A., VIJAYKUMAR N., FIDLER S., WANG Z.: Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. *arXiv preprint arXiv: 2501.18590* (2025). 2

[LGT*23] LIN C.-H., GAO J., TANG L., TAKIKAWA T., ZENG X., HUANG X., KREIS K., FIDLER S., LIU M.-Y., LIN T.-Y.: Magic3d: High-resolution text-to-3d content creation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 2

[LSR*20] LI Z., SHAFIEI M., RAMAMOORTHI R., SUNKAVALLI K., CHANDRAKER M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2020), pp. 2475–2484. 2

[LWH*23] LIU R., WU R., HOORICK B. V., TOKMAKOV P., ZAKHAROV S., VONDRICK C.: Zero-1-to-3: Zero-shot one image to 3d object. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 9264–9275. 2

[LXLW24] LIU Y., XIE M., LIU H., WONG T.-T.: Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia 2024 Conference Papers* (2024), pp. 1–11. 2

[LXR*18] LI Z., XU Z., RAMAMOORTHI R., SUNKAVALLI K., CHANDRAKER M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG) 37*, 6 (2018), 1–11. 2

[MCP*24] MITRA N. J., CEYLAN D., PATASHNIK O., COHENOR D., GUERRERO P., HUANG C.-H., SUNG M.: Diffusion models for visual content creation. In *Siggraph Tutorial* (2024). 2

[NDDJK21] NIMIER-DAVID M., DONG Z., JAKOB W., KAPLANYAN A.: Material and Lighting Reconstruction for Complex Indoor Scenes with Texture-space Differentiable Rendering. In *Eurographics Symposium on Rendering - DL-only Track* (2021), The Eurographics Association. 4

[PEL*23] PODELL D., ENGLISH Z., LACEY K., BLATTMANN A., DOCKHORN T., MÜLLER J., PENNA J., ROMBACH R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023). 9

[PJBM23] POOLE B., JAIN A., BARRON J. T., MILDENHALL B.: Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations* (2023). 2

[PTL*23] PAN X., TEWARI A., LEIMKÜHLER T., LIU L., MEKA A., THEOBALT C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings* (New York, NY, USA, 2023), SIGGRAPH '23, Association for Computing Machinery. 5

[PWMAZ24] PERLA S. R. K., WANG Y., MAHDAVI-AMIRI A., ZHANG H.: Easi-tex: Edge-aware mesh texturing from single image. *ACM Transactions on Graphics (Proceedings of SIGGRAPH) 43*, 4 (2024). 2, 3, 4

[PX22] PEEBLES W., XIE S.: Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748* (2022). 9

[PYG*24] PO R., YIFAN W., GOLYANIK V., ABERMAN K., BARRON J. T., BERMANO A., CHAN E., DEKEL T., HOLYNSKI A., KANAZAWA A., LIU C., LIU L., MILDENHALL B., NIESSNER M., OMMER B., THEOBALT C., WONKA P., WETZSTEIN G.: State of the art on diffusion models for visual computing. *Computer Graphics Forum 43*, 2 (2024). 2

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022). 2

[RDN*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125 1*, 2 (2022), 3. 2

[RH01] RAMAMOORTHI R., HANRAHAN P.: A signal-processing framework for inverse rendering. In *SIGGRAPH* (2001). 2

[RMA*23] RICHARDSON E., METZER G., ALALUF Y., GIRYES R., COHEN-OR D.: Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 Conference Proceedings* (New York, NY, USA, 2023), SIGGRAPH '23, Association for Computing Machinery. 1, 2, 3, 4

[RSP*24] RADL L., STEINER M., PARGER M., WEINRAUCH A., KERBL B., STEINBERGER M.: StopThePop: Sorted Gaussian Splatting for View-Consistent Real-time Rendering. *ACM Transactions on Graphics 43*, 4 (2024). 8

[SME21] SONG J., MENG C., ERMON S.: Denoising diffusion implicit models. In *International Conference on Learning Representations* (2021). 1, 5

[TZC*23] TANG S., ZHANG F., CHEN J., WANG P., FURUKAWA Y.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv* (2023). 2

[VBP*24] VAINER S., BOSS M., PARGER M., KUTSY K., DE NIGRIS D., ROWLES C., PERONY N., DONNÉ S.: Collaborative control for geometry-conditioned pbr image generation. *arXiv preprint arXiv:2402.05919* (2024). 2

[VKDN*24] VAINER S., KUTSY K., DE NIGRIS D., ROWLES C., ELIZAROV S., DONNÉ S.: Jointly generating multi-view consistent pbr textures using collaborative control. *arXiv preprint arXiv:2410.06985* (2024). 2

[VMR*23] VECCHIO G., MARTIN R., ROULLIER A., KAISER A., ROUFFET R., DESCHAINTRE V., BOUBEKEUR T.: Controlmat: Controlled generative approach to material capture. *arXiv preprint arXiv:2309.01700* (2023). 2

[vPPL*22] VON PLATEN P., PATIL S., LOZHKOV A., CUENCA P., LAMBERT N., RASUL K., DAVAADORJ M., NAIR D., PAUL S., BERMAN W., XU Y., LIU S., WOLF T.: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 3, 5, 8

[WTC*25] WANG L., TRAN D. M., CUI R., TG T., CHANDRAKER M., FRISVAD J. R.: Materialist: Physically based editing using single-image inverse rendering. *arXiv preprint arXiv:2501.03717* (2025). 2

[WXM*24] WANG Y., XU X., MA L., WANG H., DAI B.: Boosting 3d object generation through pbr materials. In *SIGGRAPH Asia 2024 Conference Papers* (2024), pp. 1–11. 2

[WZB*24] WEI X., ZHANG K., BI S., TAN H., LUAN F., DESCHAINTRE V., SUNKAVALLI K., SU H., XU Z.: Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385* (2024). 2

[WZY*23] WU L., ZHU R., YALDIZ M. B., ZHU Y., CAI H., MATAI J., PORIKLI F., LI T.-M., CHANDRAKER M., RAMAMOORTHI R.: Factorized inverse path tracing for efficient and accurate material-lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 3848–3858. 4

[XGL*25] XU G., GE Y., LIU M., FAN C., XIE K., ZHAO Z., CHEN H., SHEN C.: What matters when repurposing diffusion models for general dense perception tasks? In *Proc. of the IEEE International Conf. on Learning Representations* (2025). 5, 7

[YHK*24] YEH Y.-Y., HUANG J.-B., KIM C., XIAO L., NGUYEN-PHUOC T., KHAN N., ZHANG C., CHANDRAKER M., MARSHALL C. S., DONG Z., LI Z.: Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 4304–4314. 2

[YLH*23] YAN K., LUAN F., HAŠAN M., GROUEIX T., DESCHAINTRE V., ZHAO S.: Psdr-room: Single photo to scene using differentiable rendering. In *ACM SIGGRAPH Asia 2023 Conference Proceedings* (2023). 2

[YOPM24] YOUWANG K., OH T.-H., PONS-MOLL G.: Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 2

[YZL*23] YE H., ZHANG J., LIU S., HAN X., YANG W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 3

[ZCQ*24] ZENG X., CHEN X., QI Z., LIU W., ZHAO Z., WANG Z., FU B., LIU Y., YU G.: Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2024), pp. 4252–4262. 2

[ZDG*24] ZENG Z., DESCHAINTRE V., GEORGIEV I., HOLD-GEOFFROY Y., HU Y., LUAN F., YAN L.-Q., HAŠAN M.: Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers* (New York, NY, USA, 2024), SIGGRAPH '24, Association for Computing Machinery. 2, 3, 4, 5, 7, 8, 10

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (2018). 7

[ZLG*21] ZHANG Y., LING H., GAO J., YIN K., LAFLECHE J.-F., BARRIUSO A., TORRALBA A., FIDLER S.: Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10145–10155. 5

[ZLH*22] ZHU J., LUAN F., HUO Y., LIN Z., ZHONG Z., XI D., WANG R., BAO H., ZHENG J., TANG R.: Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers* (2022), ACM. 4, 5, 7, 8, 10

[ZLX*24] ZHANG Y., LIU Y., XIE Z., YANG L., LIU Z., YANG M., ZHANG R., KOU Q., LIN C., WANG W., JIN X.: Dreammat: High-quality pbr material generation with geometry- and light-aware diffusion models. *ACM Transactions on Graphics (Proc. SIGGRAPH) 43*, 4 (jul 2024). 2

[ZPX*24] ZHANG S., PENG S., XU T., YANG Y., CHEN T., XUE N., SHEN Y., BAO H., HU R., ZHOU X.: Mapa: Text-driven photorealistic material painting for 3d shapes. In *ACM SIGGRAPH 2024 Conference Papers* (2024), pp. 1–12. 2

[ZPZ*24] ZHANG H., PAN Z., ZHANG C., ZHU L., GAO X.: Texpainter: Generative mesh texturing with multi-view consistency. In *ACM SIGGRAPH 2024 Conference Papers* (New York, NY, USA, 2024), SIGGRAPH '24, Association for Computing Machinery. 2

[ZRA23a] ZHANG L., RAO A., AGRAWALA M.: Adding conditional control to text-to-image diffusion models, 2023. 2, 3

[ZRA23b] ZHANG L., RAO A., AGRAWALA M.: Controlnet-v1-1. https://github.com/lllyasviel/ControlNet-v1-1-nightly, 2023. 4, 8

[ZRL*23] ZHAO W., RAO Y., LIU Z., LIU B., ZHOU J., LU J.: Unleashing text-to-image diffusion models for visual perception. *ICCV* (2023). 3, 5

[ZWZ*24] ZHANG L., WANG Z., ZHANG Q., QIU Q., PANG A., JIANG H., YANG W., XU L., YU J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG) 43*, 4 (2024), 1–20. 2

[ZXS*25] ZHANG Y., XIONG Z., SHEN Z., LIN G., WANG H., VUN N.: Instex: Indoor scenes stylized texture synthesis. *arXiv preprint arXiv:2501.13969* (2025). 2