

# MaterialPicker: Multi-Modal DiT-Based Material Generation

XIAOHE MA, State Key Lab of CAD&CG, Zhejiang University, China

VALENTIN DESCHAINTE, Adobe Research, United Kingdom

MILOŠ HAŠAN, Adobe Research, USA

FUJUN LUAN, Adobe Research, USA

KUN ZHOU, State Key Lab of CAD&CG, Zhejiang University and ZJU-FaceUnity Joint Lab of Intelligent Graphics, China

HONGZHI WU, State Key Lab of CAD&CG, Zhejiang University, China

YIWEI HU, Adobe Research, USA



Fig. 1. We introduce MaterialPicker, a DiT-based model that generates high-quality materials, conditioned on image crops and/or text prompts. Our model accurately captures textures and material properties even from photographs of distorted or partially obscured surfaces. We demonstrate MaterialPicker by extracting material properties (albedo, normal, roughness, height and metallicity, shown in a column next to the input crops) from smartphone-captured photos, then applying these materials (as indicated by green arrows) in a 3D scene for photo-realistic rendering results.

High-quality material generation is key for virtual environment authoring and inverse rendering. We propose MaterialPicker, a multi-modal material generator leveraging a Diffusion Transformer (DiT) architecture, improving and simplifying the creation of high-quality materials from text prompts and/or photographs. Our method can generate a material based on an image crop of a material sample, even if the captured surface is distorted, viewed at an angle or partially occluded, as is often the case in photographs of natural scenes. We further allow the user to specify a text prompt to provide additional guidance for the generation. We finetune a pre-trained DiT-based

video generator into a material generator, where each material map is treated as a frame in a video sequence. We evaluate our approach both quantitatively and qualitatively and show that it enables more diverse material generation and better distortion correction than previous work.

CCS Concepts: • **Computing methodologies** → **Appearance and texture representations**.

Additional Key Words and Phrases: Material appearance, capture, generative models

## ACM Reference Format:

Xiaohe Ma, Valentin Deschaintre, Miloš Hašan, Fujun Luan, Kun Zhou, Hongzhi Wu, and Yiwei Hu. 0. MaterialPicker: Multi-Modal DiT-Based Material Generation. *ACM Trans. Graph.* 0, 0, Article 0 ( 0), 12 pages. <https://doi.org/10.1145/3731199>

## 1 INTRODUCTION

High-quality materials are a core requirement for photorealistic image synthesis. We present a multi-modal material generator, conditioned on a text prompt and/or an image. The image can be a photograph containing a material sample captured at any angle, potentially distorted or partially occluded. Our model lets users “pick” materials from any photograph just by outlining a rough crop square around the material sample of interest.

\*Corresponding authors: Xiaohe Ma (xiaohema1998@gmail.com).

Authors' addresses: Xiaohe Ma, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China, xiaohema1998@gmail.com; Valentin Deschaintre, Adobe Research, London, United Kingdom, deschain@adobe.com; Miloš Hašan, Adobe Research, San Jose, USA, mihasan@adobe.com; Fujun Luan, Adobe Research, San Jose, USA, fluan@adobe.com; Kun Zhou, State Key Lab of CAD&CG, Zhejiang University and ZJU-FaceUnity Joint Lab of Intelligent Graphics, Hangzhou, China, kunzhou@acm.org; Hongzhi Wu, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China, hwu@acm.org; Yiwei Hu, Adobe Research, San Jose, USA, yiwu@adobe.com.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 0 Copyright held by the owner/author(s).

0730-0301/0/0-ART0

<https://doi.org/10.1145/3731199>

Traditional material acquisition often requires tens or hundreds of photo samples under known light conditions and camera poses. Even with recent advances in material acquisition allowing single or few image(s) capture [Deschaintre et al. 2018, 2019; Guo et al. 2021; Martin et al. 2022; Shi et al. 2020a; Vecchio et al. 2024a; Zhou and Kalantari 2022], restrictions on the capture conditions are imposed. These methods typically require a camera flash as the only light source, and/or a fronto-parallel view of a flat sample. Even methods designed for capture using non-planar photographs [Lopes et al. 2024] cannot handle significant texture distortion in the input photographs. Many recent material generation methods are trained from scratch on synthetic materials, limiting the generation diversity due to limited datasets [Adobe 2024; Vecchio et al. 2024a], as compared to general-purpose text-to-image diffusion models [Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2022].

We propose to tackle these challenges with two new ideas. First, we create a dataset which contains 800K crops of synthetic scene renderings, textured with randomly assigned materials, with each crop paired with its ground truth material. Using this data, we train our model for the “material picking” task. We additionally use a text-to-material dataset [Vecchio et al. 2024a] containing 800K pairs of text descriptions and associated ground truth material maps, encouraging material generation diversity and resulting in a multi-modal generator that can accept images, text or both.

Second, we re-purpose a text-to-video generation model to generate material maps instead. We use a Diffusion-Transformer (DiT) based architecture, which has been shown to be effective for high-quality video synthesis [Brooks et al. 2024]. However, our target domain is materials, which we represent as a set of 2D maps (albedo, normal, roughness, height, metallicity). To adapt our base DiT model, trained on videos, to materials, we finetune it by considering each material map as a “frame” in a video sequence. This approach preserves the strong prior information in the video model, improving our method’s generalization and robustness.

We evaluate our model on both real and synthetic input images and compare it against the state-of-the-art methods for texture rectification [Hao et al. 2023], material picking [Lopes et al. 2024], material acquisition [Vecchio et al. 2024a] and text-to-material generation [Vecchio et al. 2024a,b]. We show that our approach generates materials that follow the input text prompt and/or match the appearance of the material sample in the input image, while correcting its distortions. In summary, we make the following contributions:

- We propose a material generation model which uses text and/or image prompts as inputs, while being robust to distortion, occlusion and perspective in the input image.
- We design a large-scale dataset of crops of material samples paired with the corresponding ground truth material maps, enabling our model to handle a range of viewing angles and distortions.
- We adapt a Diffusion Transformer text-to-video model for material generation by treating material maps as video frames, preserving the original prior knowledge embedded in the model to generate diverse materials.

## 2 RELATED WORK

### 2.1 Material Acquisition and Generation

While material acquisition has been a long standing challenge [Guarnera et al. 2016], lightweight material acquisition and generation have seen significant progress using machine learning. Various methods were proposed to infer PBR material maps from only a single or few photographs [Deschaintre et al. 2018, 2019, 2020; Guo et al. 2021; Shi et al. 2020a; Zhou and Kalantari 2022]. However, these methods rely on specific capture condition using a flash light co-located with the camera location. Martin et al. [2022] propose to use natural illumination but doesn’t support direct metallic and roughness map estimations. Further, these methods rely on the camera being fronto-parallel or very close to it. This kind of photographs require specific captures, making the use of in the wild photos for material creation challenging.

As an alternative to create materials, generative model for materials were proposed. GAN-based approaches [Guo et al. 2020; Zhou et al. 2022] show that unconditional generation of materials is possible and can be used for material acquisition via optimization of their noise and latent spaces. Recent progress in generative model, and more specifically diffusion models [Rombach et al. 2022], enabled more stable, diffusion based, material generators [Vecchio et al. 2024b; Xue et al. 2024]. Such diffusion models can also be used to support material acquisition tasks [Vecchio et al. 2024a], for example when paired with ControlNet [Zhang et al. 2023]. All these diffusion-based approaches either attempt to train the model from scratch, using solely synthetic material data [Vecchio et al. 2024a,b] or significantly alter the architecture of the original text-to-image model [Xue et al. 2024], preventing the use of the pre-existing priors in large scale image generation models [Rombach et al. 2022], limiting their generalization and diversity. Further, image prompts are limited to fronto-parallel photographs, which requires a specific capture.

Other methods leveraged transformers as a model for material generation [Guerrero et al. 2022; Hu et al. 2023] but focused on procedural material, which relies on generating functional graph generation, a very different modality. These procedural representations have resolution and editability benefits, but cannot easily model materials with complex texture patterns in the wild. In contrast, our model supports generating materials from any image or text prompt and produces varied, high-quality material samples.

### 2.2 Material Extraction and Rectification

Different methods were proposed to rectify textures or generally enable non-fronto-parallel textures as input. Some approaches [Yan et al. 2023; Yeh et al. 2022] aim to evaluate the materials in an image through a retrieval and optimization method. Given an image, they retrieve the geometries and procedural materials in databases to optimize their position and appearance via differentiable rendering [Shi et al. 2020b; Zhang et al. 2020]. Closest to our work is Material Palette [Lopes et al. 2024], targeting material extraction from a single photo, not restricted to fronto parallel images. The method leverages Dreambooth [Ruiz et al. 2023] optimized through a LoRA [Hu et al. 2021] on Stable Diffusion [Rombach et al. 2022] to learn a “concept” for each material. This lets them generate a texture with a similar appearance to the target material and use a

separate material estimation network to decompose the texture into material maps. However, this LoRA optimization step takes up to 3 minutes for each image, and we find that our approach reproduces better the target appearance.

A related field is that of texture synthesis from real-world images. Wu et al. [2020] present an automatic texture exemplar extraction based on Trimmed Texture CNN. VQGAN [Esser et al. 2021] achieves high resolution image-to-image synthesis with a transformer-based architecture. These methods however do not support the common occlusions and deformations that occur in natural images. To tackle this limitation, Hao et al. [2023] propose to rectify occlusions and distortions in texture images via a conditional denoising U-Net with an occlusion-aware latent transformer. We show that our approach yields better texture rectification and simultaneously generates material parameters.

### 2.3 Diffusion Models and Diffusion Transformers

Diffusion models [Ho et al. 2020; Sohl-Dickstein et al. 2015; Song and Ermon 2019; Song et al. 2021] are state-of-the-art generative models, showing great results across various visual applications such as image synthesis and video generation. The core architecture of diffusion models progressed from simple U-Nets, incorporating self-attention and enhanced upscaling layers [Dhariwal and Nichol 2021], prior-based text-to-image model [Nichol et al. 2021; Ramesh et al. 2022], a VAE [Kingma and Welling 2022] for latent diffusion models (LDM) [Rombach et al. 2022] and temporal attention layers for video generations [Blattmann et al. 2023a,b]. These image generation methods all rely on a U-Net backbone, a convolutional-based encoder-decoder architecture.

Recently, transformer-based diffusion models, Diffusion Transformers (DiT) were proposed [Peebles and Xie 2023], benefiting from the scalability of Transformer models, removing the convolutions inductive bias. PixArt- $\alpha$  presents a DiT-based text-to-image that can synthesize high resolution images with low training cost. Stable Diffusion 3 [Esser et al. 2024] demonstrates that a multi-modal DiT model trained with Rectified Flow can achieve superior image synthesis quality. Compared to the U-Net architecture, the DiT shows greater flexibility in the representation on the visual data, which is particularly important to video synthesis tasks. Sora [Brooks et al. 2024], a DiT-based video diffusion model, encodes video sequences as tokens and uses transformers to denoise these visual tokens, demonstrating the ability to generate minute-long, high-resolution high-quality videos. We adapt a DiT-based video generation model for our purpose and show that it can be flexibly transformed into a multi-channel material generator.

## 3 METHOD

### 3.1 Diffusion Transformers

Diffusion models are generative models that iteratively transform an initial noise distribution (e.g. Gaussian noise) into a complex real-world data distribution (e.g., images, or their encodings). The diffusion process relies on a *forward* process that progressively transforms the original data distribution into a noise distribution. For example, this can be achieved by iteratively adding Gaussian

noise to the data sample. Given data samples  $x \sim p_{\text{data}}$ , corrupted data  $p(x_T|x_0) = \prod_{t=1}^T p(x_t|x_{t-1}, \epsilon)$ ,  $\epsilon \sim \mathcal{N}(0, I)$  are constructed in  $T$  diffusion steps.

To sample the original data distribution  $p_{\text{data}}$  from the noise distribution, a *reverse* mapping  $p(x_0) = p(x_T) \prod_{t=1}^T q(x_{t-1}|x_t, \epsilon_t)$  needs to be modeled where  $\epsilon_t$  is the noise sampled at each step. A neural network  $f_\theta$  is conditioned on the denoising step  $t$  to predict the noise  $\epsilon_t$ , which is then used to reconstruct  $x_{t-1}$  from  $x_t$  in each reverse step [Ho et al. 2020]:

$$\mathbb{E}_{x \sim p_{\text{data}}, t \sim U(0, T)} [\|\epsilon_t - f_\theta(x_t; c, t)\|^2], \quad (1)$$

where  $c$  is conditional inputs (e.g., text prompts or images).

We use a Diffusion Transformer [Peebles and Xie 2023] architecture as a backbone to model  $f_\theta$ . The visual data  $x \in \mathbb{R}^{F \times 3 \times H \times W}$  is tokenized patch-wise, resulting in visual tokens  $\hat{x} \in \mathbb{R}^{V \times D}$  where  $H, W, F$  are the spatial and temporal dimensions of the video,  $V$  is the number of tokens and  $D$  is the feature dimension. Positional encoding is also added to  $\hat{x}$  to specify spatial and temporal order. Any condition  $c$  is also embedded as tokens  $\hat{c} \in \mathbb{R}^{V' \times D}$  where  $V'$  is the number of the tokens for conditional inputs. For example, when  $c$  is a text, it is encoded by a pre-trained encoder [Radford et al. 2021] with additional embedding layers to map it into the same feature dimension  $D$ . The transformer  $f_\theta(\hat{x}_t; \hat{c}, t)$  is trained to denoise each patch at timestep  $t$ . The final denoised patches  $\hat{x}_0 \in \mathbb{R}^{V \times D}$  are reassembled as visual data  $x_0 \in \mathbb{R}^{F \times 3 \times H \times W}$  after decoding through linear layers. Since the number of tokens grows quickly with resolution, we use a variational autoencoder (VAE) model [Peebles and Xie 2023; Rombach et al. 2022] before the tokenizing process, producing a latent representation of  $y \in \mathbb{R}^{F' \times D' \times H' \times W'}$  of the original data  $x$  for the transformer to process.

### 3.2 Datasets

To train our material generative model, we propose two datasets, *Scenes* and *Materials*. Together, these datasets enable joint training for both surface rectification and high quality material generation.

For the *Scenes* dataset, we build a set of synthetic indoor scenes with planar floors, walls, and randomly placed 3D objects, such as cubes, spheres, cylinders, cones, and toruses, similar to random Cornell boxes [Cohen and Greenberg 1985]. Each object is randomly assigned a unique material from around 3,000 stationary (i.e., approximately shift invariant) materials. We use the Blender implementation of the Disney Principled BSDF model [McAuley et al. 2012] for rendering the dataset and other visualizations in the paper, using base color (albedo), normal, roughness, metallic, and height maps, leaving other parameters as default. Using this approach we create a dataset of 100,000 high-resolution rendered images, with different kinds of light sources, including point lights and area lights, to simulate complex real-world illumination (see Fig. 2). We randomly place cameras to capture a wide variety of view points and maximize coverage.

We further crop the rendered images to construct training data, including input images, corresponding material maps, binary material mask, and the material name as an optional text prompt. During cropping, we ensure that the dominant material occupies at least 70% of the region. Importantly, we rescale the material maps based

on UV coordinates to ensure that the rendered crops and target material maps share a matching texture scale. After cropping, this dataset contains 800,000 text-image-mask-material tuples. We will share out dataset creation script, facilitating reproduction using public materials datasets [Ma et al. 2023; Vecchio and Deschaintre 2024].

As our *Scenes* dataset only contains stationary materials, it may fail to represent the full diversity of textures in the wild. To enhance the generalization capability, we use an additional *Materials* dataset [Martin et al. 2022], which we augment to 800,000 cropped material maps. We use the name of the materials as the text prompts for text-to-material generation. These data items can be thought of as text-material pairs. This additional data diversity leads to significant improvement for non-stationary textures in input photographs as discussed in Sec. 4.4.2.

### 3.3 Generative Material Model

We employ a pre-trained DiT-based text-to-video generative model as our base model, with an architecture similar to that of the publicly available HunyuanVideo model [Weijie et al. 2024]. It follows a decoder-only Transformer structure with stacked self-attention blocks. The model takes both text and visual conditions as input: the input frames are first encoded by a 3DVAE encoder, producing a latent representation which is then corrupted by noise during training. Simultaneously, text prompts are processed by a T5-based encoder, producing text embeddings that are appended to the noisy latents. Additionally, timestep embeddings and spatial-temporal positional embeddings are added to the latent sequences to provide temporal and spatial context (i.e., each token’s frame number and position within the frame). The DiT backbone denoises the latent sequence, which is decoded into video frames using the 3DVAE decoder. We retarget this architecture into a multi-channel material generator.

To retarget the model while preserving its learned prior knowledge, we stack the material maps  $M$  (albedo map, normal map, height map, roughness map and metallicity map) into a “video” of 5 frames, and compute the temporal positional embedding assuming their time stamp interval is 1 e.g., fps=1. Since DiT flexibly generates tokenized data, as opposed to a U-Net architecture [Blattmann et al. 2023a], the number of frames it is able to produce is not fixed, allowing us to adapt the original video generator to generate the right number of “frames” to meet our requirement. Note that we treat all material maps as “keyframes” (images) in the 3DVAE, with no motion prediction between the frames.

For image-conditioned material generation, we consider the input image  $I$  as the first frame, with the model generating the stacked material maps  $M$  as the subsequent frames, similar to a video extension model. This setup allows the transformer’s self-attention mechanism to jointly reason over both the input image and the predicted material maps, while tolerating pixel misalignment due to perspective distortion or varying camera poses.

Our decision to use a video-generation backbone is further motivated by its inherent capacity to enforce temporal consistency,

which, in this context, translates to spatial alignment across material maps. The DiT model implicitly learns that all frames beyond the first should remain temporally coherent, a property that aligns well with the goal of generating consistent texture channels. Convolution-based architectures such as SDXL [Podell et al. 2023] led to worse results in our initial experiments; we hypothesize this is because they prefer (approximate) pixel alignment between input and output.

This design also avoids architectural modifications that disrupt pretrained knowledge. Image diffusion models are typically trained to generate 3 channels (RGB) and need to be non-trivially adjusted to generate more channels [Liu et al. 2023], or generate a single material map at a time and repurpose the input text prompt as a “switch” [Zeng et al. 2024]. Recent work [Vecchio et al. 2024a] modified the architecture to generate multiple maps, but had to train the model from random weights, missing the rich prior provided by large scale image datasets about material appearance. Our use of a video model enables the generation of multiple maps with minimally invasive architecture modifications, inheriting strong priors from pretrained video diffusion models.

Our solution preserves compatibility with inference-time techniques such as noise rolling [Vecchio et al. 2024a], TexSliders [Guerrero-Viu et al. 2024], SDEdit [Meng et al. 2021], etc., which could broaden its range of applications. Furthermore, since the DiT backbone operates on tokenized representations rather than fixed-size tensors (as in U-Net-based architectures), it is extensible, as additional maps (e.g., opacity maps) can be incorporated by appending new frames at the end of the token sequence. It also remains computationally efficient, since only a small number of frames is generated.

We additionally train our material generator to produce a segmentation mask for the dominant material in the crop. Typically, the user-provided crop is not entirely covered by a single material (see Fig. 4). Performing conservative cropping on an image may reduce the number of usable pixels, while using an additional segmentation mask requires additional user input or a separate segmentation model [Sharma et al. 2023]. Instead, our model automatically identifies the dominant material [Lu et al. 2009] in the image. We add a mask  $S$  to be inferred from the input image as the second frame. Our training data  $x$  can thus be represented as  $x = \text{stack}(I, S, M)$ , where  $x \in \mathbb{R}^{7 \times 3 \times H \times W}$ ; we have 7 RGB frames: input, mask, and five material maps. Since mask, height, roughness, and metallic maps are single-channel, we convert them into RGB images before concatenating them with other frames. Noise  $\epsilon_t$  is applied only to the last six frames occupied by  $S$  and  $M$ , resulting in  $x_t = \text{stack}(I, S_t, M_t)$ , with the first frame (input image) remaining free of noise. Our objective from Eq. 1 is

$$\mathbb{E}_{x \sim p_{\text{data}}, t \sim U(0, T)} \left[ \|\epsilon_t - f_{\theta}(x_t; c, t)[-6 :]\|^2 \right], \quad (2)$$

where  $c$  denotes the text material description. This process can be seen as frames completion (mask and material channels) given the input image and text condition. The notation  $[-6 :]$  refers to the last 6 frames generated by the Transformer. When the input consists solely of  $c$  without  $I$ ,  $x = \text{stack}(S, M)$  where  $S$  is a uniformly white RGB image. The computation of the loss remains unchanged.

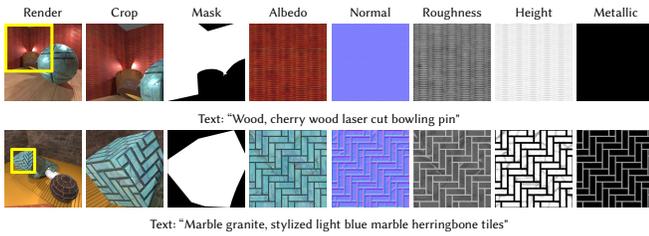


Fig. 2. Our *Scenes* dataset. We build random scenes and render paired text/image-to-material dataset with 3K randomly sampled materials. In each row we show a 2K synthetic rendering, a crop with a dominant material, the material mask and corresponding material maps.

### 3.4 Training and Inference

We finetune the pre-trained DiT model using the AdamW optimizer on 8 Nvidia A100 GPUs. The learning rate is set at  $0.99 \times 10^{-4}$  with an effective batch size of 64. The model is finetuned on  $256 \times 256$  resolution for about 70K steps, which takes 90 hours. During training, we feed data from our two training datasets *Scenes* and *Materials* in a 5:3 ratio, prioritizing the task of image-conditioned material generation. For text-only or unconditional generation, the mask is replaced by a completely white image placeholder.

Our model completes a generation in 12 seconds using DDIM [Song et al. 2020] with 50 diffusion steps on an Nvidia A100. The model natively outputs a resolution of 256 due to limited computational resources. We apply an upsampler [Niu 2023] to increase the resolution of each material map to  $512 \times 512$ .

## 4 RESULTS

We evaluate the performance of our MaterialPicker across multiple dimensions. First, we perform qualitative and quantitative comparisons with Material Palette [Lopes et al. 2024] on material extraction using both synthetic and real-world images (Sec. 4.2). Next, we compare with a material acquisition method [Vecchio et al. 2024a] and a texture rectification method on real-world images (Sec. 4.2), and with MatGen [Vecchio et al. 2024a] and MatFuse [Vecchio et al. 2024b] on text-to-material generation (Sec. 4.3). Finally, we conduct ablation studies on multi-modality, dataset design, the usage of a mask, and evaluate the impact of the input image scale and the robustness to distortion and lighting/shadowing. We also demonstrate our generalizability to complex patterns and the ability to generate seamless results (Sec. 4.4).

### 4.1 Evaluation dataset and metrics

**4.1.1 Synthetic evaluation dataset.** For systematic evaluation, we build a synthetic evaluation dataset by gathering a diverse set of 531 materials from PolyHaven<sup>1</sup>, applied to three interior scenes from the Archinteriors collection [Evermotion 2021] (completely independent from our training set). For each scene, we sequentially apply the 531 collected materials to a designated object inside the scene, and render 2D images using Blender Cycles [Blender Community 2018] with the scene's default illumination setup. We generate 1,593

<sup>1</sup><https://polyhaven.com/>

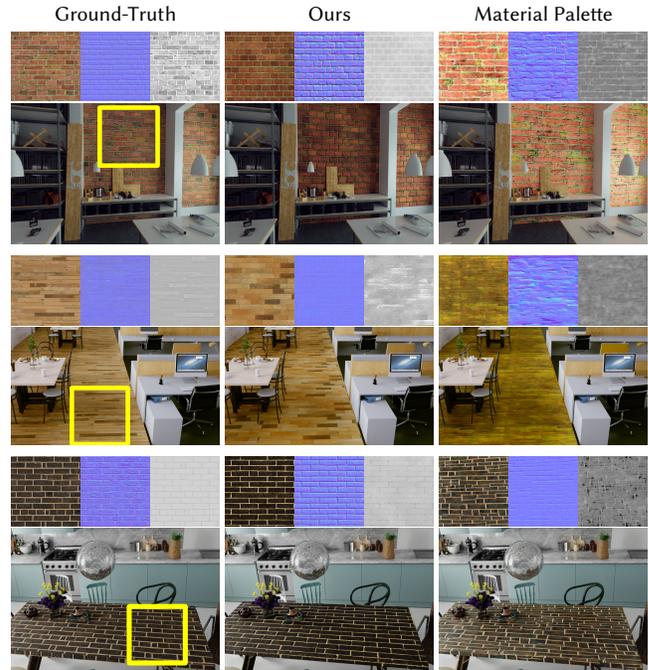


Fig. 3. Comparisons with Material Palette [Lopes et al. 2024] on synthetic dataset for material extraction. The first column shows the ground truth material maps from PolyHaven, with the rendered scene below. The yellow square area indicates the crop used as the input for both models. The second and third columns show the material maps extracted by our model and Material Palette, along with the re-rendered images. We can see that our approach better matches the ground truth appearance.

synthetic renderings, and crop a square around the location of the object with replaced materials.

**4.1.2 Real photographs evaluation dataset.** To validate the generalization of our models, we curate an evaluation dataset containing real photographs captured by smartphones. This dataset covers a comprehensive set of real-world materials observed under both natural outdoor lighting and complex indoor illumination. We crop the photographs with a primary focus on our target material, without strictly limiting the cropping boundaries.

**4.1.3 Evaluation metrics.** Since we do not target pixel-aligned material capture, per-pixel metrics cannot be used for our results. Instead, we focus on the *appearance similarity* of the materials extracted from the photo inputs. Following related work on high-fidelity image synthesis such as DreamBooth [Ruiz et al. 2023], we leverage CLIP-I, which is the average pairwise cosine similarity between ViT-L-14 CLIP [Radford et al. 2021] embeddings of two sets of images. We also use the DINO metric [Ruiz et al. 2023] to measure the average pairwise cosine similarity between ViT-L-16 DINO embeddings.

### 4.2 Image Conditioned Generation

We evaluate the performance of our model on both synthetic images and real photographs. We first show a visual comparison with the state-of-the-art method Material Palette [Lopes et al. 2024] on our

Table 1. Quantitative results of material extraction. We compare with Material Palette [Lopes et al. 2024] and report the average CLIP-I metric  $\uparrow$  and DINO metric  $\uparrow$  between the output material maps and ground truth alongside the 95% confidence interval.

	CLIP $\uparrow$	Albedo	Normal	Roughness	Render
<b>Mat-Palette</b>	0.816 $\pm$ 0.03	0.867 $\pm$ 0.03	0.791 $\pm$ 0.03	0.955 $\pm$ 0.01	
<b>Ours</b>	<b>0.857<math>\pm</math>0.02</b>	<b>0.874<math>\pm</math>0.02</b>	<b>0.866<math>\pm</math>0.03</b>	<b>0.967<math>\pm</math>0.01</b>	
	DINO $\uparrow$	Albedo	Normal	Roughness	Render
<b>Mat-Palette</b>	0.503 $\pm$ 0.1	0.631 $\pm$ 0.09	0.502 $\pm$ 0.09	0.797 $\pm$ 0.05	
<b>Ours</b>	<b>0.494<math>\pm</math>0.1</b>	<b>0.672<math>\pm</math>0.08</b>	<b>0.566<math>\pm</math>0.1</b>	<b>0.863<math>\pm</math>0.04</b>	

synthetic evaluation dataset (Sec. 4.1.1). Since Material Palette generates only three material maps (albedo, normal, and roughness), we present both qualitative and quantitative results for these channels, along with the re-rendered images using these generated material maps. Our method takes 12 seconds to generate a material while Material Palette takes 3 minutes, on the same Nvidia A100 GPU, a 15 times speedup. Furthermore, our model can generate materials in batches. In Fig. 3 we show that our model produces material maps with a closer texture appearance and better matching the ground-truth material maps. In contrast, Material Palette struggles to reconstruct structured textures often resulting in distorted lines. We also observe that in the rendered images, our generated materials better matches the original input images.

We include a quantitative comparison and 95% confidence interval with Material Palette on the entire synthetic dataset in Tab. 1. We find that our model performs better on two metrics for the vast majority of generated materials channels, with the exception of the Albedo for which the intervals overlap. Our re-rendered images also show consistently higher alignment with the ground truth.

We show qualitative evaluation on real photographs in Fig. 4 where we see that our model generalizes well to photographs of materials from various angles. We render the generated materials on a planar surface under environment lighting, showing strong visual similarity to the original input images. Unlike Material Palette, which requires input masks from a separate segmentation step [Sharma et al. 2023], our model operates out-of-box with an input image only, showcasing its potential as a lightweight *MaterialPicker*.

We show a comparison with ControlMat [Vecchio et al. 2024a] in Fig. 5. Since ControlMat uses local features of the images extracted by ControlNet [Zhang et al. 2023] to guide the diffusion process, it is trained to generate results that align with the input, therefore fail to handle imperfect perspectives and distortions. Further, since the generation of multichannel materials in ControlMat depends on modifications to the VAE, it must be trained from scratch without leveraging any image or video priors from which our method benefits. This limits ControlMat generalization to complex textures captured under unseen lighting conditions (e.g. 4th row).

Since our model automatically performs perspective rectification on the generated materials, we further compare against another state-of-the-art texture rectification and synthesis method [Hao et al. 2023]. In Fig. 17, we evaluate both methods using real photographs. Since our model directly outputs material maps, instead of textures, we present our results by rendering them under different

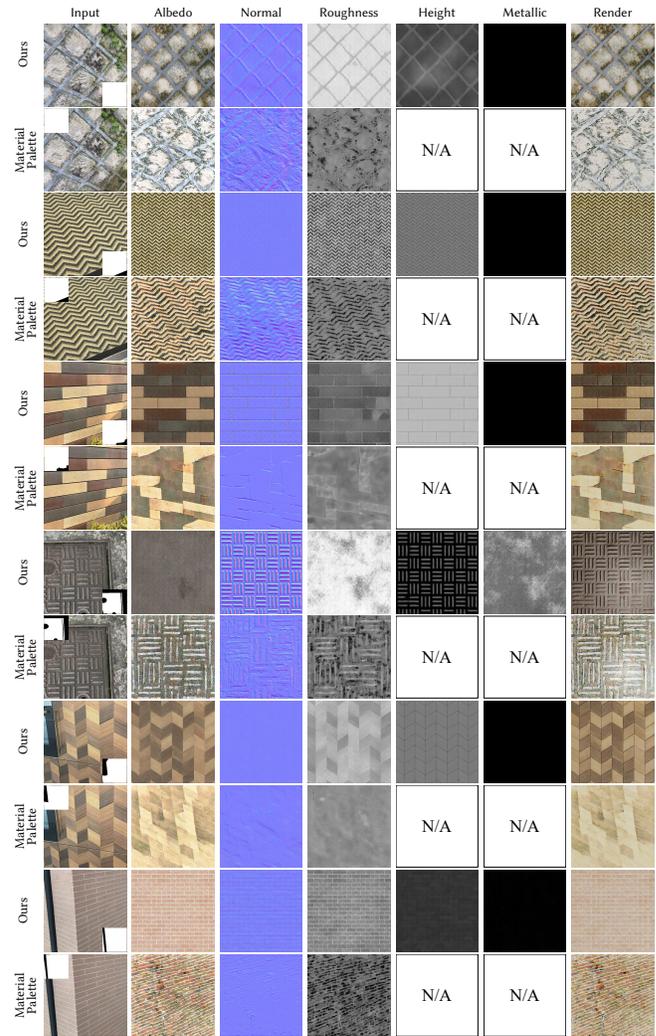


Fig. 4. Comparison between our method and Material Palette [Lopes et al. 2024] on material extraction for real photographs. The first column shows the input images and the generated (ours bottom-right)/provided (Material Palette top-left) masks. The second to sixth columns show the generated material maps and rendering under an environment map. We see that our approach better corrects for distortion and match the original appearance.

environment maps. We find that the compared method does not generalize well to real-world photographs, taken from non-frontal and/or non-parallel setups and fails to correct distortion in these cases. In contrast, our approach synthesizes a fronto-parallel view and remains robust across various real-world lighting conditions and viewing angles. Finally, as previously, our model does not require detailed masks as input, directly rectifying the dominant texture in the input image.

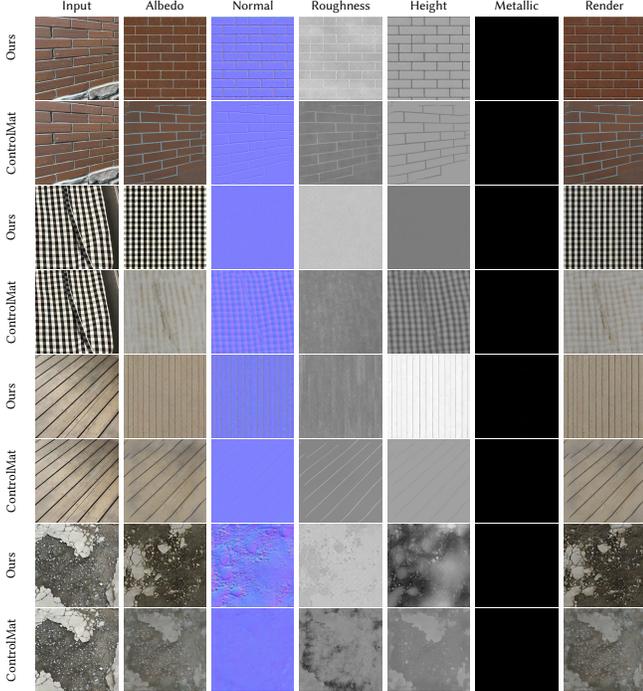


Fig. 5. Comparison between our method and ControlMat [Vecchio et al. 2024a] on material extraction for real photographs with distortions. The first column shows the input images. The second to the sixth columns show the generated material maps and rendering under an environment map. As ControlMat relies on well aligned conditions through ControlNet, our approach demonstrates superior performance in correcting perspectives and distortions. Further, as ControlMat is trained only on material data, it does not benefit from the learned priors from a pre-trained model like our fine-tuning, making their method less effective at generalizing to unseen patterns and lighting conditions (e.g. in the 4th row).

### 4.3 Text Conditioned Generation

Although the primary focus of our method is the generation of materials from photos, our multi-modal model also supports text-conditioned generation without image inputs. We evaluate its performance on the text-to-material task, comparing it with two state-of-the-art diffusion-based generative models for material synthesis: MatFuse [Vecchio et al. 2024b] and MatGen [Vecchio et al. 2024a]. As shown in Fig. 6, our model demonstrates strong text-to-material synthesis capability, producing high-quality material samples, comparable to other state-of-the-art approaches. We report the cosine similarity between each rendered image and the text condition, calculated using ViT-L-14 CLIP [Radford et al. 2021] embeddings. Leveraging a pretrained text-to-video model as a prior, our model can interpret complex semantics beyond the material-only training set, such as "wood rings" and "floral" patterns.

### 4.4 Ablation Study

**4.4.1 Multi-modality.** Our generative material model takes advantages of its multi-modality. Though it is designed to create material maps from input photographs, it can benefit from additional signal

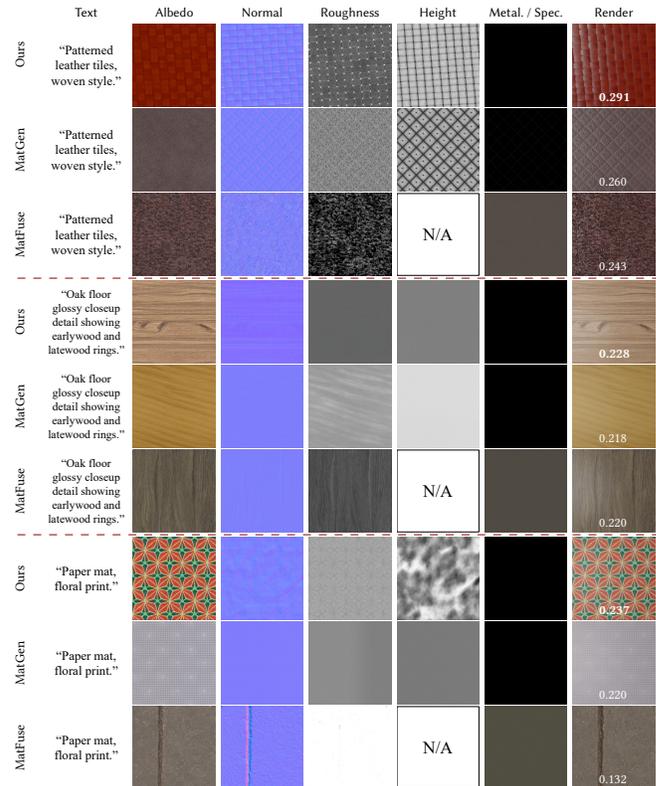


Fig. 6. Comparison of text-to-material generation between our model, MatGen [Vecchio et al. 2024a], and MatFuse [Vecchio et al. 2024b]. The "Text" column contains the input text conditions. The second to last columns show the generated material maps, along with a rendering under environment lighting. The CLIP score between the rendering and the text is reported at the bottom of each rendered image (higher is better). We note that MatFuse generates a specular map rather than a metallic map.

to reduce the ambiguity of a single in-the-wild photograph. We present different combinations of input conditions in Fig. 7 including 1) text condition only; 2) image condition only and 3) text+image dual conditions. We found that text conditioning provides high level guidance for material generation. On the other hand, image conditioning contains ambiguities, as lighting and camera poses are uncontrolled. Combining both options enables text prompts to guide the model in identifying the reflective properties of a material. For instance, by prompting the model with appropriate text, it can better differentiate between metallic and non-metallic materials, as shown in the third example in Fig. 7.

**4.4.2 Mixed Dataset.** In Sec. 3.2, we introduce two datasets used to train our model. To confirm that using both datasets help, we train a variant using only the *Scene* dataset. Since this dataset primarily contains stationary materials, training exclusively on it reduces our model's generalization for complex texture patterns commonly found in real-world scenarios as shown in Fig. 8. By mixing additional training data, our model synthesizes more diverse texture

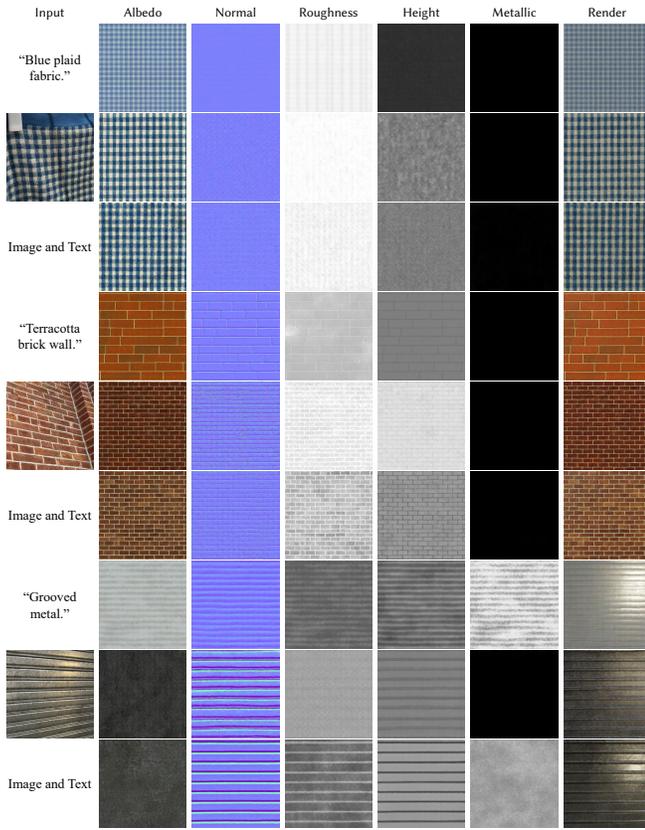


Fig. 7. Comparison of different input combinations. The first column shows the input condition. Text conditioning provides only coarse guidance for material generation, while image inputs offer explicit cues on material appearance. However, image inputs remain ambiguous with respect to material properties, as seen in the third example. Using both text and image conditioning simultaneously reduces this ambiguity, enhancing control and quality.

patterns and features such as woven pattern or the texture of a manhole cover.

**4.4.3 Mask as Input or Output.** As opposed to existing material generation models, our model does not require the target material to cover the entire input image [Vecchio et al. 2024a] or manually-created masks [Lopes et al. 2024] to identify the sample of interest. Our model instead outputs a mask along with the generated materials. To assess the impact of generating this mask, we train an alternative model using our two datasets, with a slight modification to the model configuration. We add noise  $\epsilon_t$  to the material maps  $M$  only, with  $x_t = \text{stack}(I, V, M_t)$ , leaving the image and mask as non-noised inputs (or  $x_t = \text{stack}(V, M_t)$  without  $I$ ), using our adaptation of a video model (as described in Sec. 3.3). The loss is then computed on the material maps  $M$  only. As shown in Fig. 9, we find that our proposed model, which automatically predicts a mask, performs comparably well to this variant requiring the mask as input.

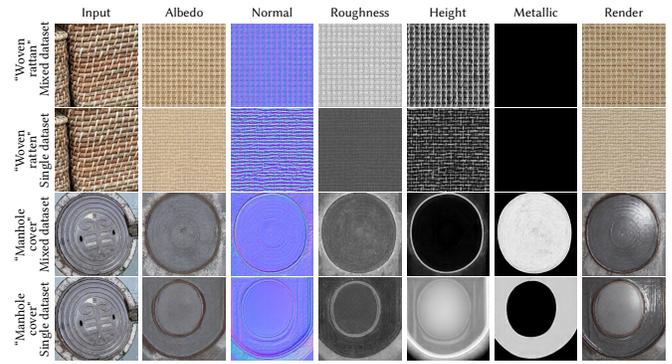


Fig. 8. Impact of our text-to-material synthetic dataset on generation and generalization. For each sample, the first row shows the generation results from our baseline model, trained on mixed datasets (Sec. 3.2), and the second row shows results from a model trained only on the *Synthetic scenes*. The model trained with mixed dataset is able to synthesize better non-stationary, realistic textures. The leftmost side of each row is labeled with the text conditioning input used.

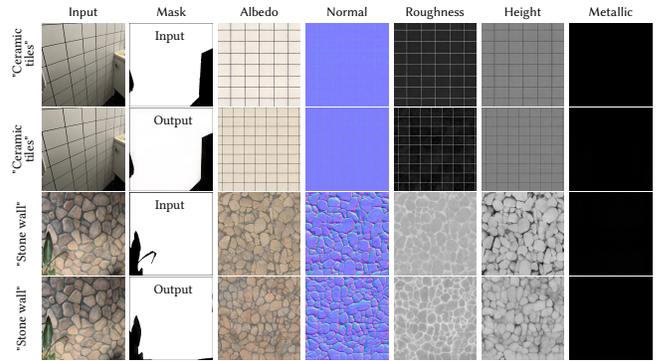


Fig. 9. Impact of mask on material generation quality. Here shows the effect of using the mask as an input versus as an output on the quality of generated materials. Each pair of consecutive rows represents the results from the model with the mask as input (top row) and the model with the mask as output (bottom row). The results show that our model can accurately predict masks without a decrease in material quality. The leftmost side of each row is labeled with the text conditioning input used.

**4.4.4 Input scale.** Reproducing the texture scale in the input photos is critical for material generation. As we process our training data to align the scales of input images and output material maps (Sec. 3.2), our model generates scale-matched materials, as shown in Fig. 10. We see that our result follows the scale of the input as it increases from top to bottom.

**4.4.5 Evaluations on the Robustness.** To examine the robustness of our model to strong, real-world, distortions, we generate a synthetic test set that use textures from the texture datasets TexSD [Lopes et al. 2024] and follow the texture processing steps outlined by Hao et al. [2023]. We apply homography transformations [Hartley and Zisserman 2003] and thin plate spline transformations [Bookstein 1989] to the textures. Our results in Fig. 11 show that the model

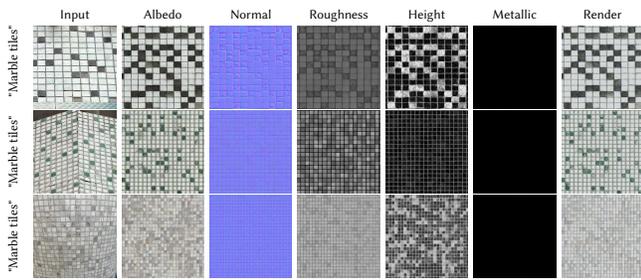


Fig. 10. Evaluation of our model’s adaptability to different input texture scales on real photographs. We can see that our results are generated with a scale matching that of the input.

is robust to severe distortions, stretching, and the blurring effects introduced by these transformations. More examples of real photos with distortion or surface geometry diversity can be found in supplemental materials.

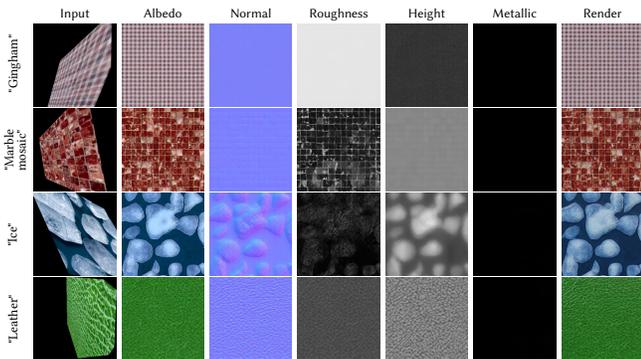


Fig. 11. Evaluation of our model’s robustness to varying levels of distortion. The first column shows textures transformed with homography and thin plate spline transformations. The following columns present the material maps and the rendering images. The results demonstrate that our model effectively rectifies textures with various patterns and different types of distortion, maintaining high-quality outputs.

We further evaluate the model’s performance when the input image contains specular highlights and shadows in Fig. 12. We see that these highlights and shadows in real photos do not “leak” into material maps, highlighting the model’s robustness to various lighting conditions.

To further demonstrate the generalization ability of our model beyond common indoor and outdoor scenes, we test several images sourced from Pixabay<sup>2</sup> (royalty-free). The results are shown in Fig. 13. Notably, for challenging appearances such as animal skin and fur, and plant surfaces, our model can generate visually plausible results.

**4.4.6 Tileable Generation.** Although our model is not explicitly trained to produce tileable outputs, we can easily incorporate noise rolling, a test-time technique proposed in ControlMat [Vecchio et al. 2024a], to generate seamless material maps without any re-training.

<sup>2</sup><https://pixabay.com/>

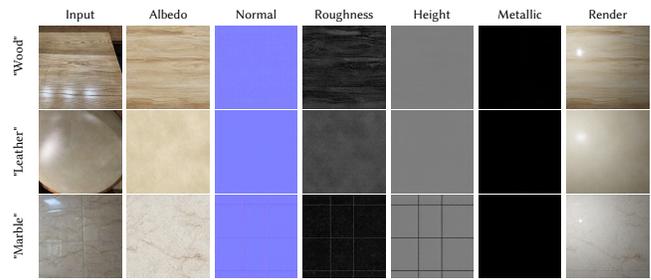


Fig. 12. Evaluation of the robustness of our model to lighting and shadow interference. We test scenarios where the input photographs contain point light sources, shadows, or environmental reflections. The generated material maps and rendered images demonstrate the ability of our model to handle these interferences, preserving material quality and accurately representing the input photos. The leftmost side of each row is labeled with the text conditioning input used.

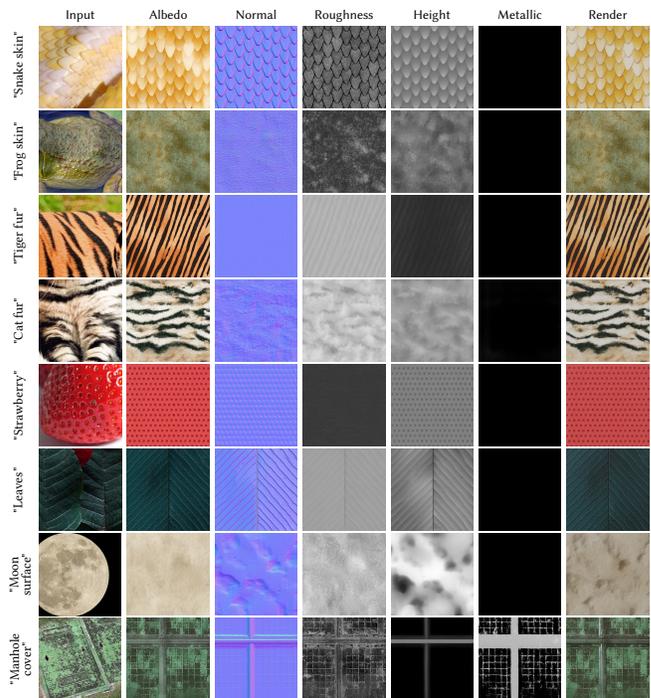


Fig. 13. Evaluation of the generalization ability of our model to complex patterns. We test on several royalty-free images from Pixabay, including animal and plant appearances as well as abstract textures like the moon surface. The results demonstrate that our model generalizes well beyond typical indoor and outdoor scenes, producing visually plausible outputs for a diverse range of complex surface patterns.

We present several successful examples of applying this technique in Fig. 14.

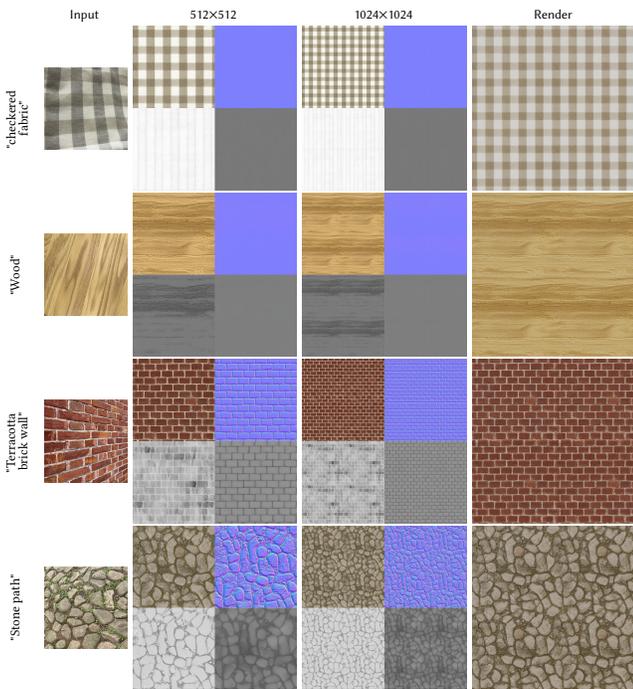


Fig. 14. Evaluation of tileable material generation. During inference, we can employ the noise rolling strategy [Vecchio et al. 2024a] to enhance the seamlessness of the generated results. The first column shows the input images, followed by the 512×512 material maps. The third column displays 1024×1024 tiling results obtained by stitching the 512-resolution outputs, and the final column presents the rendering results using the 1024-resolution maps. Maps are ordered as albedo, normal, roughness, and height, from top to bottom and left to right. Metallic maps are omitted as they are black in these cases.

## 5 LIMITATIONS

Despite strong generation capacity, our model may still encounter challenging inputs, as shown in Fig. 15. In the first row we show an example where our model confuses shading and albedo variation. Our model may also have difficulty handling materials with cutouts or holes, since it does not produce opacity maps as outputs. Also, preserving semantically meaningful patterns, such as text, is a remaining challenge in our approach. Furthermore, our current model is not capable of handling transparency and refraction effects. In the third row of Fig. 11, we present an example of generating "ice", where the model tries to approximate the appearance using the albedo map. The fourth and fifth rows of Fig. 15 illustrate more challenging cases, in which the model fails to reproduce the visual characteristics of transparent and translucent materials due to the lack of relevant training data.

To improve the model's capacity for handling complex appearance effects such as transparency, translucency, and detailed BRDF components, one promising direction is to incorporate more advanced material maps (e.g., specular, coat, and subsurface scattering) into the training data, illustrated in Fig. 16. As outlined in Sec. 3.3, our architecture is flexible and can be extended to generate additional

output channels corresponding to these effects. We leave this extension for future work, once sufficiently large-scale datasets with such complex annotations become available.

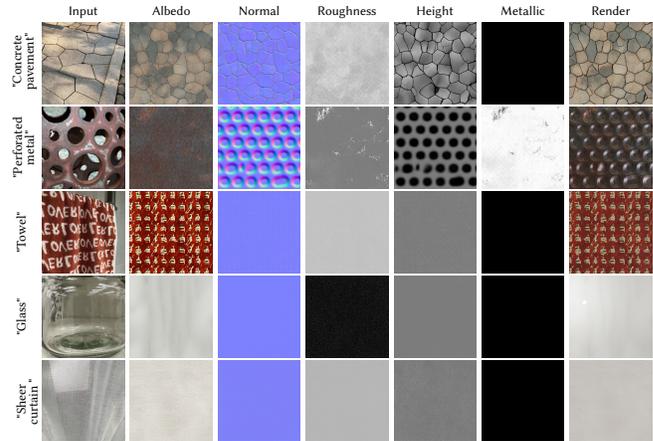


Fig. 15. Limitations. We show limitations of our model, such as complex lighting and shadows in the first row, materials with perforations in the second row, structurally significant elements like text in the third row, and effects involving transparency and translucency in the last two rows.



Fig. 16. Future work. These images are obtained using the method described in Sec. 3.2, with the addition of a translucency map during rendering. We consider such complex scenarios as future work.

## 6 CONCLUSION

We present a generative model for high-quality material synthesis from text prompts and/or crops of natural images by finetuning a pretrained text-to-video generative model, which provides strong prior knowledge. The flexible video DiT architecture lets us adjust the model for multi-channel material generation. We show extensive evaluation on both synthetic and real examples and conduct systematic ablation studies and test on robustness. We believe that our re-purposing of a video model for multi-channel generation opens an interesting avenue for other domain which require the generation of additional channels, such as intrinsic decomposition [Vecchio et al. 2024a].

## ACKNOWLEDGMENTS

We would like to thank Prafull Sharma for his early implementation of the data rendering code [Sharma et al. 2023]. This work is partially supported by NSF China (62332015, 62227806 & 62421003), the XPLOER PRIZE and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

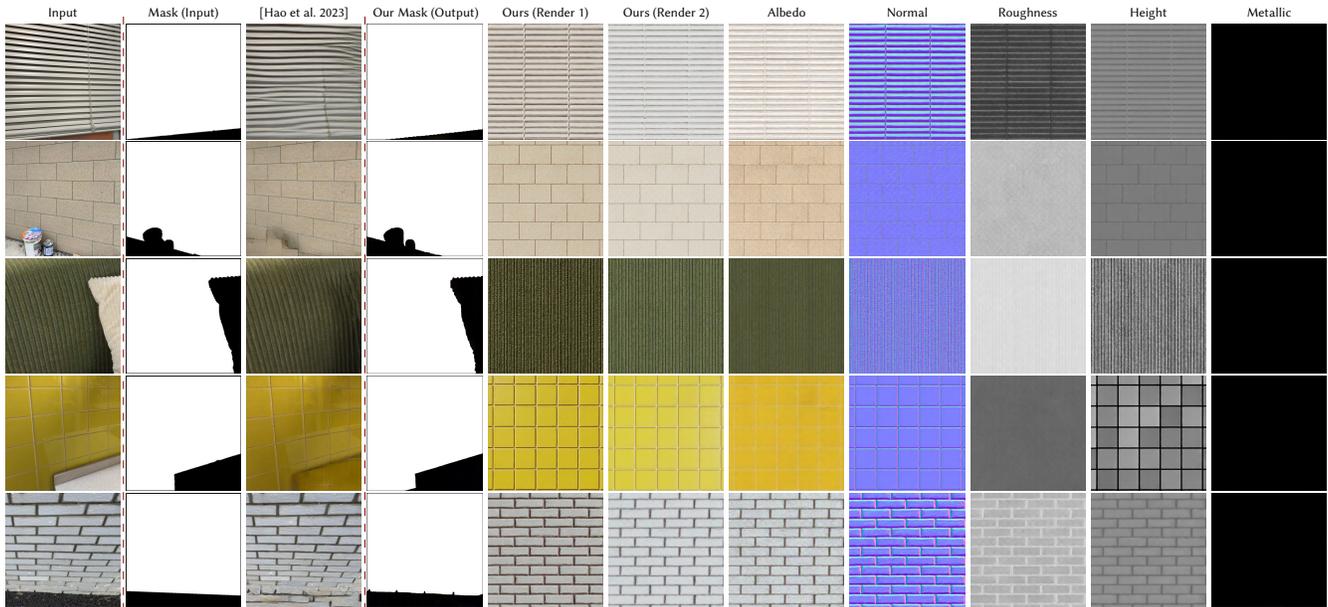


Fig. 17. Comparison with Hao et al. [2023] on texture rectification for real photos. The first column shows the input photos. The second and third columns are the required input masks and output textures of Hao et al. The fourth column shows masks generated by our model, followed by two renderings (Render 1 & Render 2) of our generated material maps (in the last five columns) under different environment maps. Despite not requiring an input mask, our method better rectifies perspective and distortions. Further, as we support material extraction, our result does not contain shading from the input image.

## REFERENCES

- Adobe. 2024. Adobe Substance3D Asset. <https://substance3d.adobe.com/assets>.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).
- Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22563–22575.
- Blender Community. 2018. Blender - A 3D Modelling and Rendering Package. <http://www.blender.org>. Accessed: 2024-11-05.
- Fred L. Bookstein. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence* 11, 6 (1989), 567–585.
- Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). <https://openai.com/research/video-generation-models-as-world-simulators>
- Michael F Cohen and Donald P Greenberg. 1985. The hemi-cube: A radiosity solution for complex environments. *ACM Siggraph Computer Graphics* 19, 3 (1985), 31–40.
- Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-image SVBRDF Capture with a Rendering-aware Deep Network. *ACM Trans. Graph.* 37, 4 (2018), 128:1–128:15.
- Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2019. Flexible svbrdf capture with a multi-image deep network. In *Computer graphics forum*, Vol. 38. Wiley Online Library, 1–13.
- Valentin Deschaintre, George Drettakis, and Adrien Bousseau. 2020. Guided Fine-Tuning for Large-Scale Material Transfer. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)* 39, 4 (2020).
- Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.
- Evermotion. 2021. Archinteriors Collection. <https://evermotion.org/shop/cat/397/archinteriors> Accessed: 2024-11-05.
- D. Guarnera, G.C. Guarnera, A. Ghosh, C. Denk, and M. Glencross. 2016. BRDF Representation and Acquisition. *Computer Graphics Forum* 35, 2 (2016), 625–650. <https://doi.org/10.1111/cgf.12867>
- Paul Guerrero, Milos Hasan, Kalyan Sunkavalli, Radomir Mech, Tamy Boubekeur, and Niloy Mitra. 2022. MatFormer: A Generative Model for Procedural Materials. *ACM Trans. Graph.* 41, 4, Article 46 (2022). <https://doi.org/10.1145/3528223.3530173>
- Julia Guerrero-Viu, Milos Hasan, Arthur Roullier, Midhun Harikumar, Yiwei Hu, Paul Guerrero, Diego Gutierrez, Belen Masia, and Valentin Deschaintre. 2024. Texsliders: Diffusion-based texture editing in clip space. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.
- Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. 2021. Highlight-Aware Two-Stream Network for Single-Image SVBRDF Acquisition. *ACM Trans. Graph.* 40, 4, Article 123 (jul 2021), 14 pages. <https://doi.org/10.1145/3450626.3459854>
- Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020. MaterialGAN: Reflectance Capture using a Generative SVBRDF Model. *ACM Trans. Graph.* 39, 6 (2020), 254:1–254:13.
- Guoqing Hao, Satoshi Iizuka, Kensho Hara, Edgar Simo-Serra, Hirokatsu Kataoka, and Kazuhiro Fukui. 2023. Diffusion-based Holistic Texture Rectification and Synthesis. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.
- Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- Yiwei Hu, Paul Guerrero, Milos Hasan, Holly Rushmeier, and Valentin Deschaintre. 2023. Generating Procedural Materials from Text or Image Prompts. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.
- Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [stat.ML]* <https://arxiv.org/abs/1312.6114>

- Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. 2023. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579* (2023).
- Ivan Lopes, Fabio Pizzati, and Raoul de Charette. 2024. Material palette: Extraction of materials from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4379–4388.
- Jianye Lu, Julie Dorsey, and Holly Rushmeier. 2009. Dominant Texture and Diffusion Distance Manifolds. *Computer Graphics Forum* 28, 2 (2009), 667–676. <https://doi.org/10.1111/j.1467-8659.2009.01407.x> <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2009.01407.x>
- Xiaohe Ma, Xianmin Xu, Leyao Zhang, Kun Zhou, and Hongzhi Wu. 2023. OpenSVBRDF: a database of measured spatially-varying reflectance. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–14.
- Rosalie Martin, Arthur Roullier, Romain Rouffet, Adrien Kaiser, and Tamy Boubekeur. 2022. MaterIA: Single Image High-Resolution Material Capture in the Wild. *Computer Graphics Forum* 41, 2 (2022), 163–177. <https://doi.org/10.1111/cgf.14466> <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14466>
- Stephen McAuley, Stephen Hill, Naty Hoffman, Yoshiharu Gotanda, Brian Smits, Brent Burley, and Adam Martinez. 2012. Practical physically-based shading in film and game production. In *ACM SIGGRAPH 2012 Courses* (Los Angeles, California) (SIGGRAPH '12). Association for Computing Machinery, New York, NY, USA, Article 10, 7 pages. <https://doi.org/10.1145/2343483.2343493>
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- Hui Niu. 2023. waifu2x-ncnn-vulkan. <https://github.com/nihui/waifu2x-ncnn-vulkan> GitHub repository.
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22500–22510.
- Prafull Sharma, Julien Philip, Michaël Gharbi, Bill Freeman, Fredo Durand, and Valentin Deschaintre. 2023. Materialistic: Selecting Similar Materials in Images. *ACM Trans. Graph.* 42, 4, Article 154 (jul 2023), 14 pages. <https://doi.org/10.1145/3592390>
- Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik. 2020a. MATch: Differentiable Material Graphs for Procedural Material Capture. *ACM Trans. Graph.* 39, 6 (Dec. 2020), 1–15.
- Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik. 2020b. MATch: Differentiable Material Graphs for Procedural Material Capture. *ACM Trans. Graph.* 39, 6 (Dec. 2020), 1–15.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502* (October 2020). <https://arxiv.org/abs/2010.02502>
- Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=PXTIG12RRHS>
- Giuseppe Vecchio and Valentin Deschaintre. 2024. MatSynth: A Modern PBR Materials Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22109–22118.
- Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. 2024a. ControlMat: A Controlled Generative Approach to Material Capture. *ACM Trans. Graph.* 43, 5, Article 164 (sep 2024), 17 pages. <https://doi.org/10.1145/3688830>
- Giuseppe Vecchio, Renato Sortino, Simone Palazzo, and Concetto Spampinato. 2024b. Matfuse: controllable material generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4429–4438.
- Kong Weijie, Tian Qi, Zhang Zijian, Min Rox, Dai Zuoohuo, Zhou Jin, Xiong Jiangfeng, Li Xin, Wu Bo, Zhang Jianwei, Wu Kathrina, Lin Qin, Wang Aladdin, Wang Andong, Li Changlin, Huang Duoju, Yang Fang, Tan Hao, Wang Hongmei, Song Jacob, Bai Jiawang, Wu Jianbing, Xue Jinbao, Wang Joey, Yuan Junkun, Wang Kai, Liu Mengyang, Li Pengyu, Li Shuai, Wang Weiyan, Yu Wenqing, Deng Xinchu, Li Yang, Long Yanxin, Chen Yi, Cui Yutao, Peng Yuanbo, Yu Zhentao, He Zhiyu, Xu Zhiyong, Zhou Zixiang, Xu Zunnan, Tao Yangyu, Lu Qinglin, Liu Songtao, Zhou Dax, Wang Hongfa, Yang Yong, Wang Di, Liu Yuhong, Jiang Jie, and Zhong Caesar. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. <https://arxiv.org/abs/2412.03603>
- Huisi Wu, Wei Yan, Ping Li, and Zhenkun Wen. 2020. Deep texture exemplar extraction based on trimmed T-CNN. *IEEE Transactions on Multimedia* 23 (2020), 4502–4514.
- Bowen Xue, Claudio Guarnera, Shuang Zhao, and Zahra Montazeri. 2024. Reflectance-Fusion: Diffusion-based text to SVBRDF Generation. In *Eurographics Symposium on Rendering*. Eurographics Association.
- K. Yan, F. Luan, M. Hašan, T. Groueix, V. Deschaintre, and S. Zhao. 2023. PSDR-Room: Single Photo to Scene using Differentiable Rendering. In *ACM SIGGRAPH Asia 2023 Conference Proceedings*.
- Yu-Ying Yeh, Zhengqin Li, Yannick Hold-Geoffroy, Rui Zhu, Zexiang Xu, Miloš Hašan, Kalyan Sunkavalli, and Manmohan Chandraker. 2022. Photoscene: Photorealistic material and lighting transfer for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18562–18571.
- Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024. Rgb-->x: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.
- Cheng Zhang, Bailey Miller, Kai Yan, Ioannis Gkioulekas, and Shuang Zhao. 2020. Path-Space Differentiable Rendering. *ACM Trans. Graph.* 39, 4 (2020), 143:1–143:19.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.
- Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2022. TileGen: Tileable, Controllable Material Generation and Capture. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) (SA '22). Association for Computing Machinery, New York, NY, USA, Article 34, 9 pages. <https://doi.org/10.1145/3550469.3555403>
- Xilong Zhou and Nima Khademi Kalantari. 2022. Look-ahead training with learned reflectance loss for single-image svbrdf estimation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–12.