MaterialPicker: Multi-Modal Material Generation with Diffusion Transformers

Anonymous submission

Figure 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, height, roughness and metallicity, shown in a column next to the input crops) from smartphone-captured photos, then applying these materials in a 3D scene[1] for photo-realistic rendering results.

Abstract

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.

[1]The 3D model in Fig. 1 was downloaded from Open3DLab [33], licensed under CC BY-NC-ND 4.0 [7]. We gratefully acknowledge Open3DLab for providing the model. For the purpose of visualization in our work, we made adjustments to the positions of certain objects, modified the material maps, and updated the lighting settings in Blender.

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.

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 [8, 9, 17, 30, 42, 47, 55], strong restrictions on the capture conditions are imposed. These methods typically require a flash light co-located with the camera as the only light source, and/or a fronto-parallel view of a flat sample. Even methods designed for capture using non-planar photographs [28] 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 [1, 47], as compared to general-purpose text-to-image diffusion models [31, 38, 39].

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 can train our model for the "material picking task", with various observation angles and distortion. We additionally use a text-to-material dataset [47] 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 [6]. However, our target domain is materials, which we represent as a set of 2D maps (albedo, normal, height, roughness, 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 [19], material picking [28] and text-to-material generation [47, 48]. 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 distortion.
- 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 [15], 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 [8–10, 17, 42, 55]. However, these methods rely on specific capture condition using a flash light co-located with the camera location. Martin et al. [30] 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 [18, 56] 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 [39], enabled more stable, diffusion based, material generators [48, 50]. Such diffusion models can also be used to support material acquisition tasks [47], for example when paired with ControlNet [54]. All these diffusion-based approaches either attempt to train the model from scratch, using solely synthetic material data [47, 48] or significantly alter the architecture of the original text-to-image model [50], preventing the use of the pre-existing priors in large scale image generation models [39], 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 [16, 25] 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 [51, 52] 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 [43, 53]. Closest to our work is Material palette [28], targeting material extraction from a single photo, not restricted to fronto parallel images. The method leverages Dreambooth [40] optimized through a LoRA [24] on Stable Diffusion [39] 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. [49] present an automatic texture exemplar extraction based on Trimmed Texture CNN. VQ-

GAN [12] 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. [19] propose to rectify occlusions and distortions in texture images via a conditional denoising UNet 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 [23, 44–46] 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 UNets, incorporating self-attention and enhanced upscaling layers [11], prior-based text-to-image model [31, 38], a VAE [26] for latent diffusion models (LDM) [39] and temporal attention layers for video generations [2, 3]. 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 [35], benefiting from the scalability of Transformer models, removing the convolutions inductive bias. PixArt-α presents a DiT-based text-to-image that can synthesize high resolution images with low training cost. Stable Diffusion 3 [13] 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 [6], 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 Transformer

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_{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 ϵ_t is the noise predicted at each step by a neural network f_θ. The neural network f_θ is conditioned on the denoising step t to predict the noise ϵ_t, which is then used to reconstruct x_{t-1} from x_t in each reverse step [23]:

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

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

We use a Diffusion Transformer [35] architecture as a backbone to model f_θ. The visual data $x \in \mathbb{R}^{T\times 3\times H\times W}$ is tokenized patch-wise, resulting in visual tokens $\hat{x} \in \mathbb{R}^{V\times D}$ where H, W, T 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 an pre-trained encoder [37] 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 $x_0 \in \mathbb{R}^{T\times 3\times H\times W}$ after decoding through linear layers. Since the number of tokens grows quickly with resolution, we use a variational autoencoder (VAE) model [35, 39] before the tokenizing process, producing a latent representation of $y \in \mathbb{R}^{T'\times D'\times H'\times W'}$ 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 [34]. Each object is randomly assigned a unique material from around 3,000 stationary (i.e., approximately shift invariant) materials. 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.

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 [30], 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, with an architecture similar to the one described in Brooks et al. [6], as a base model. We retarget it into a multi-channel material generator.

To retarget this 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 [2], 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. Our proposed use of a video model is in contrast to image diffusion models which typically generate 3 channels (RGB) and need to be non-trivially adjusted to generate more channels [27].

To enable material generation from an image input, 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. Recall that the input image can be captured with arbitrary camera pose, and may include perspective and distortions. Using this approach, the self-attention mechanism of the transformer ensures that the generated material parameters are aligned with each other, and allows for non-aligned pixels between the input image condition and the generated material maps. The model simply learns that all frames except the first need to be aligned. This property is key for texture rectification, which is challenging for convolution-based architectures, in which (approximate) pixel alignment between input and output is built in due to the convolution inductive bias.

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 [41].

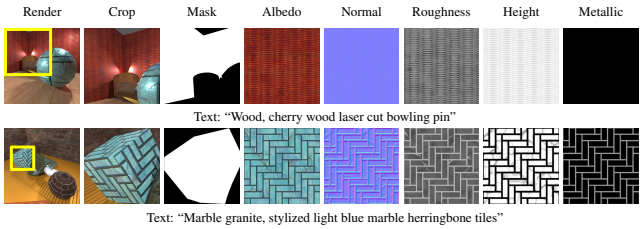

Figure 2. Our *Scenes* dataset. We build random scenes and render paired text/image-to-material dataset with 3K randomly sampled materials. In each row we show a 2K synthetic rendering, a crop with a dominant material, the material mask and corresponding material maps.

Instead, our model automatically identifies the dominant material [29] in the image. We add a mask V to be inferred from the input image as the second frame. Our training data x can thus be represented as $x = \text{stack}(I, V, M)$, where $x \in \mathbb{R}^{7\times3\times H\times W}$; we have 7 RGB frames: input, mask, and five material maps. Noise ϵ_t is applied only to the last six frames occupied by V and M, resulting in $x_t = \text{stack}(I, V_t, M_t)$, with the first frame (input image) remaining free of noise. Our objective from Eq. 1 is

$$\mathbb{E}_{x\sim p_{\text{data}},t\sim U(0,1)}\left[\|\epsilon_t - f_\theta(x_t; c, t)[-6:]\|^2\right], \quad (2)$$

where c denotes the text input (material description). This process can be seen as a completion of the frames (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, $x = \text{stack}(V, M)$ where V is a uniformly white RGB image. The computation of the loss remains unchanged.

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 0.99×10^{-4} with an effective batch size of 64. The model is finetuned on 256×256 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. We also randomly drop the conditions to retain the capacity to use classifier-free guidance (CFG) [22]. 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 50 diffusion steps. The model natively outputs a resolution of 256 due to limited computational resources. We apply an upsampler [32] to increase the resolution of each material map to 512×512.

4. Results

We evaluate the performance of our MaterialPicker across multiple dimensions. First, we perform qualitative and quantitative comparisons with Material Palette [28] on material extraction using both synthetic and real-world images (Sec. 4.2). Next, we compare with a texture rectification method on real-world images (Sec. 4.2) and with MatGen [47] and MatFuse [48] on text-to-material generation (Sec. 4.3). Finally, we conduct ablation studies on multi-modality, dataset design and evaluate the impact of the input image scale(Sec. 4.4). We also ablate the usage of a mask and evaluate robustness to distortion and lighting/shadowing in the supplemental materials.

4.1. Evaluation dataset and metrics

Synthetic Evaluation dataset. For systematic evaluation, we build a synthetic evaluation dataset by gathering a diverse set of 531 materials from PolyHaven [36], applied to three interior scenes from the Archinteriors collection [14] (which are 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 [4] with the scene's default illumination setup. We generate a total of 1,593 synthetic renderings, and crop a square around the location of the object with replaced materials.

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.

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 [40], we leverage CLIP-I, which is the average pairwise cosine similarity between ViT-L-14 CLIP [37] embeddings of two sets of images. We also use the DINO metric [40] to measure the average pairwise cosine similarity between ViT-L-16 DINO embeddings. Additionally, we report the FID score [21] to measure the statistical visual similarity between two image distributions. We compute the FID score for each of the 531 output material map sets against the corresponding ground truth material map sets, and average the FID scores over the three scenes in our synthetic evaluation dataset.

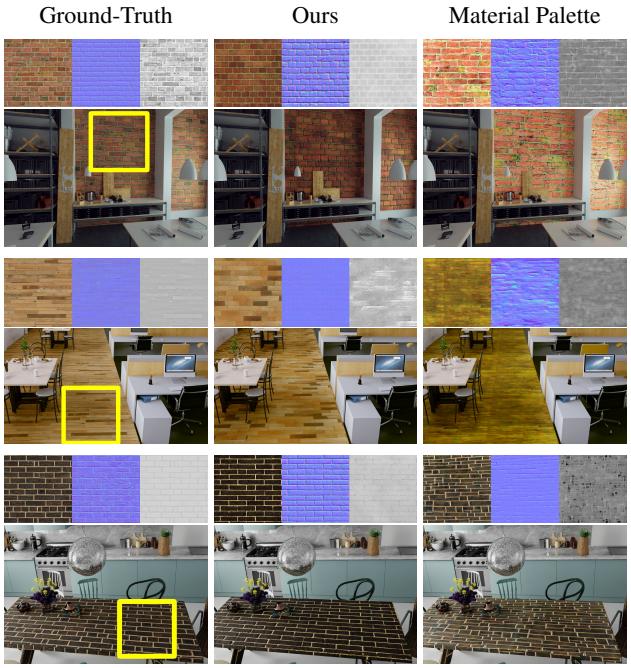

Figure 3. Comparisons with Material palette [28] on synthetic dataset for material extraction. The first column shows the ground truth material maps from PolyHaven, with the rendered scene below. The yellow square area indicates the crop used as the input for both models. The second and third columns 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.

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 [28] on our synthetic evaluation dataset (Sec. 4.1). 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 with Material Palette on the entire synthetic dataset in Tab. 1. We find that our proposed model performs better on all three met-

Table 1. Quantitative results of material extraction. We compare with Material Palette [28] and report the average CLIP-I metric ↑ and DINO metric ↑ between the output material maps and ground truth. We also report the FID metric ↓ between the generated and ground-truth material maps. The 95% confidence interval can be found in the supplemental materials.

CLIP↑	Albedo	Normal	Roughness	Render
Mat-Palette	0.816	0.867	0.791	0.955
Ours	**0.857**	**0.874**	**0.866**	**0.967**
DINO↑	**Albedo**	**Normal**	**Roughness**	**Render**
Mat-Palette	**0.503**	0.631	0.502	0.797
Ours	0.494	**0.672**	**0.566**	**0.863**
FID↓	**Albedo**	**Normal**	**Roughness**	**Render**
Mat-Palette	112.249	92.620	155.354	26.828
Ours	**107.310**	**91.057**	**79.785**	**12.224**

rics for the vast majority of the generated material channels and the corresponding re-rendered images.

We show qualitative evaluation on real photographs in Fig. 4 where we can 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 [41], our model operates out-of-box with an input image only, showcasing its potential as a lightweight *MaterialPicker*.

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 [19]. In Fig. 5, 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 environment maps. We find that the compared method doesn't 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.

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 [48] and MatGen [47]. 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. 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.

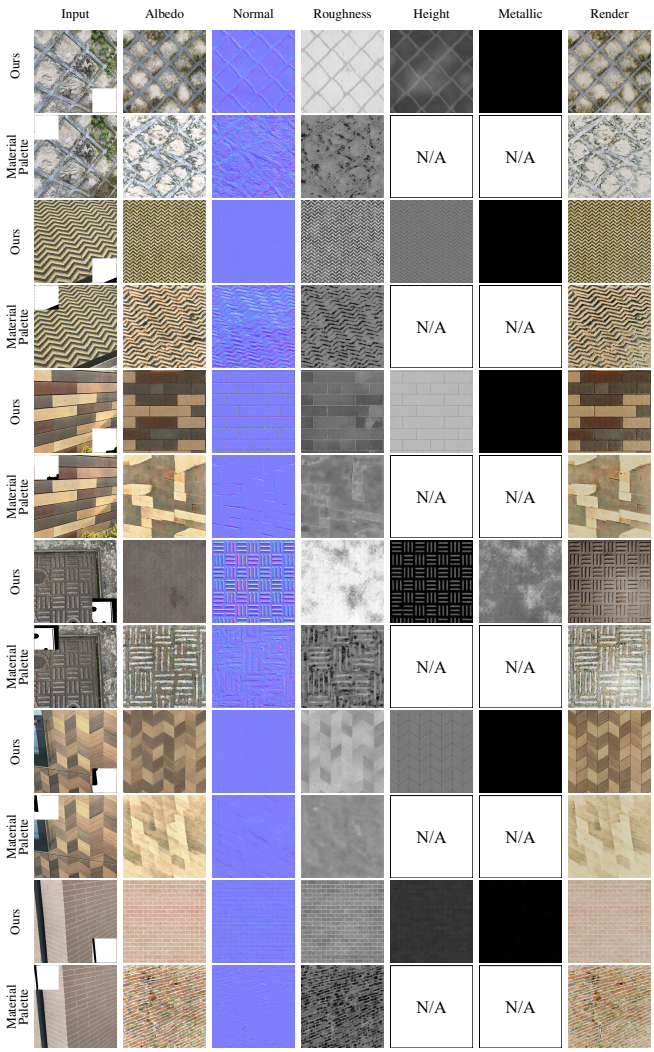

Figure 4. Comparison between our method and Material Palette [28] on material extraction for real photographs. First column shows the input images and the generated (ours bottom-right)/provided (Material Palette top-left) mask for each method. Second to the sixth columns show the generated material maps and rendering under an environment map. We see that our approach better corrects for distortion and match the original appearance.

4.4. Ablation Study

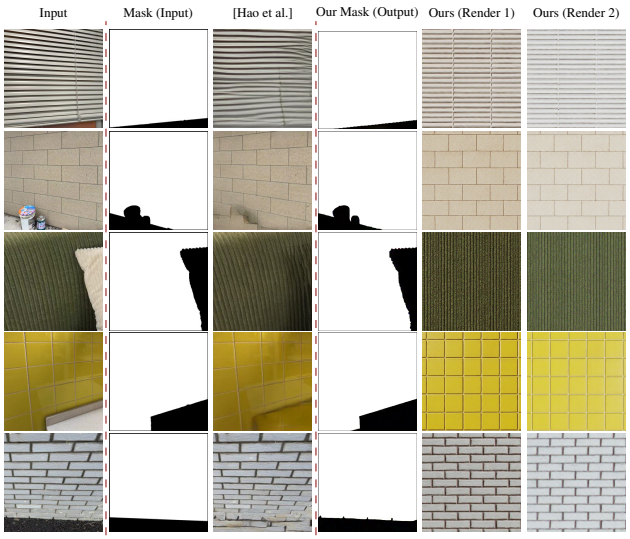

Figure 5. Comparison with Hao et al. [19] on texture rectification for real photos. The first column shows the input photos. The second and third columns are the required input masks and output textures of Hao et al. The fourth column shows masks generated by our model, followed by two renderings (Render 1 & Render 2) of our generated material maps 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.

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 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

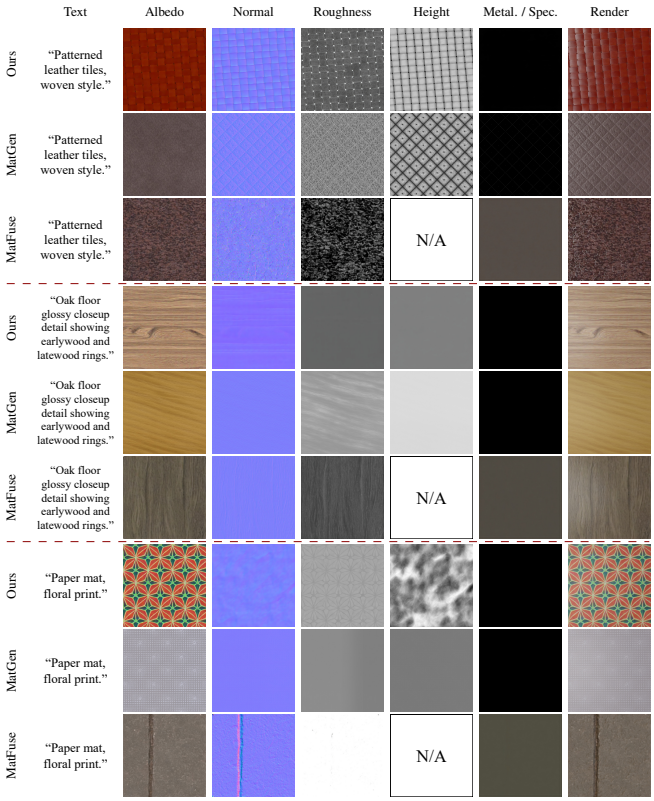

Figure 6. Comparison of text-to-material generation between our model, MatGen [47], and MatFuse [48]. The "Text" column contains the input text conditions. The second to last column show the generated material maps, along with a rendering under environment lighting. We note that MatFuse generates a specular map rather than a metallic map.

patterns commonly found in real-world scenarios as shown in Fig. 8. By mixing additional training data, our model synthesizes more diverse texture patterns and features such as woven pattern or the texture of a manhole cover.

4.4.3 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. 9. We see that our result follow the scale of the input as it increases from top to bottom.

5. Limitations

Despite strong generation capacity, our model may still encounter challenging inputs, as shown in Fig. 10. 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

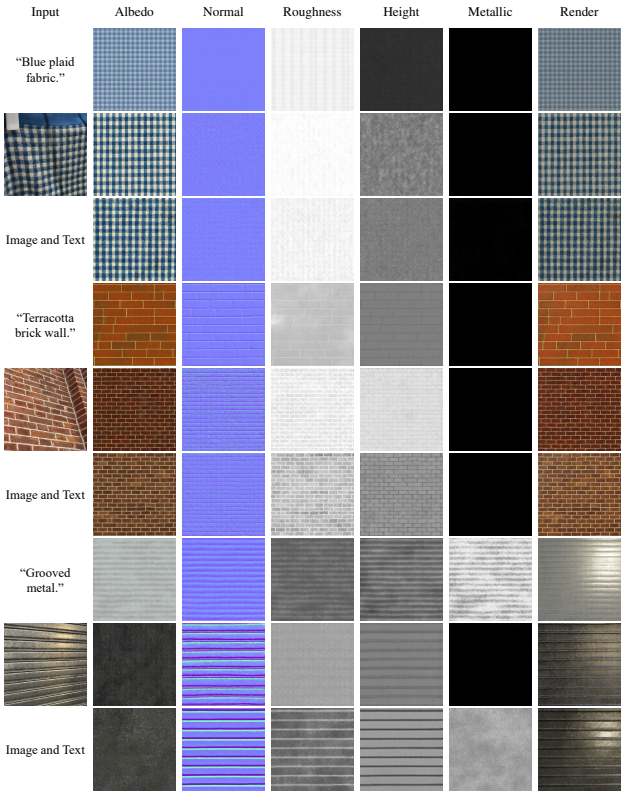

Figure 7. Comparison of different input combinations. The first column shows the input condition. Text conditioning provides only coarse guidance for material generation, while image inputs offer explicit cues on material appearance. However, image inputs remain ambiguous with respect to material properties,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

not produce opacity maps as outputs. Finally, preserving semantically meaningful patterns, such as text, is a remaining challenge in our approach.

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 repurposing 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 [47].

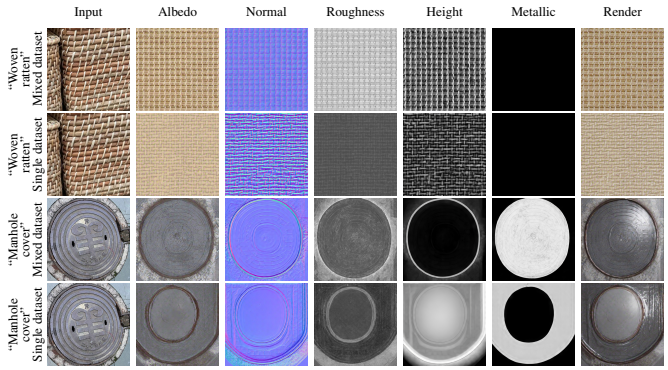

Figure 8. Impact of our text-to-material synthetic dataset on generation and generalization. For each sample, the first row shows the generation results from our baseline model, trained on mixed datasets (Sec. 3.2), and the second row shows results from a model trained only on the *Synthetic scenes*. 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.

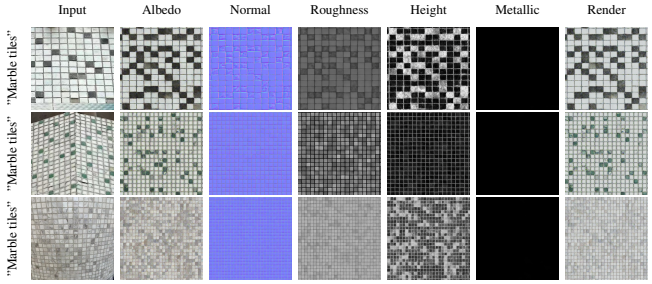

Figure 9. Evaluation of our model's adaptability to different input texture scales on real photographs. We can see that our results are generated with a scale matching that of the input.

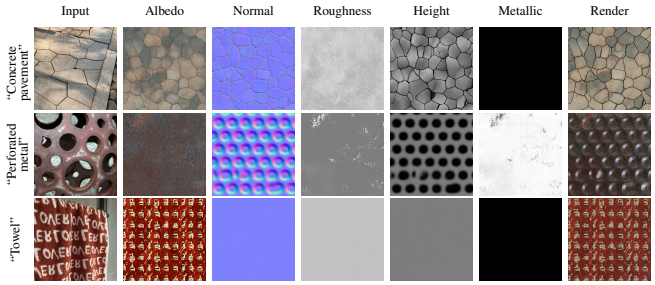

Figure 10. Limitations. We show limitations of our model, such as complex lighting and shadows in the first row, materials with perforations in the second row, and structurally significant elements like text in the third row.

References

[1] Adobe. Adobe substance3d asset, 2024. https://substance3d.adobe.com/assets. 1

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel

Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3, 4

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[4] Blender Online Community. Blender - a 3d modelling and rendering package. http://www.blender.org, 2018. Accessed: 2024-11-05. 5

[5] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 1

[6] 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. Video generation models as world simulators. 2024. 2, 3, 4

[7] Creative Commons. Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0), 2013. Accessed: 2024-12-04. 1

[8] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Trans. Graph.*, 37(4):128:1–128:15, 2018. 1, 2

[9] Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. Flexible svbrdf capture with a multi-image deep network. In *Computer graphics forum*, pages 1–13. Wiley Online Library, 2019. 1

[10] Valentin Deschaintre, George Drettakis, and Adrien Bousseau. Guided fine-tuning for large-scale material transfer. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)*, 39(4), 2020. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3

[14] Evermotion. Archinteriors collection, 2021. Accessed: 2024-11-05. 5

[15] D. Guarnera, G.C. Guarnera, A. Ghosh, C. Denk, and M. Glencross. Brdf representation and acquisition. *Computer Graphics Forum*, 35(2):625–650, 2016. 2

[16] Paul Guerrero, Milos Hasan, Kalyan Sunkavalli, Radomir Mech, Tamy Boubekeur, and Niloy Mitra. Matformer: A generative model for procedural materials. *ACM Trans. Graph.*, 41(4), 2022. 2

[17] Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. Highlight-aware two-stream network for single-image svbrdf acquisition. *ACM Trans. Graph.*, 40(4), 2021. 1, 2

[18] Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. Materialgan: Reflectance capture using a generative svbrdf model. *ACM Trans. Graph.*, 39(6):254:1–254:13, 2020. 2

[19] Guoqing Hao, Satoshi Iizuka, Kensho Hara, Edgar Simo-Serra, Hirokatsu Kataoka, and Kazuhiro Fukui. Diffusion-based holistic texture rectification and synthesis. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2, 3, 6, 7, 1

[20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[25] Yiwei Hu, Paul Guerrero, Milos Hasan, Holly Rushmeier, and Valentin Deschaintre. Generating procedural materials from text or image prompts. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 3

[27] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579*, 2023. 4

[28] Ivan Lopes, Fabio Pizzati, and Raoul de Charette. Material palette: Extraction of materials from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4379–4388, 2024. 1, 2, 5, 6, 3, 4

[29] Jianye Lu, Julie Dorsey, and Holly Rushmeier. Dominant texture and diffusion distance manifolds. *Computer Graphics Forum*, 28(2):667–676, 2009. 4

[30] Rosalie Martin, Arthur Roullier, Romain Rouffet, Adrien Kaiser, and Tamy Boubekeur. Materia: Single image high-resolution material capture in the wild. *Computer Graphics Forum*, 41(2):163–177, 2022. 1, 2, 4

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and

Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 3

[32] Hui Niu. waifu2x-ncnn-vulkan, 2023. GitHub repository. 4

[33] Open3D Lab. Open3d lab, 2024. Accessed: 2024-12-04. 1

[34] C. C. Patterson, D. P. Greenberg, J. F. Hughes, and J. Arvo. An experimental investigation of computer graphics realism. In *Proceedings of the 16th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 41–50. ACM, 1984. 3

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[36] Poly Haven. Poly haven - free public 3d asset library. https://polyhaven.com/. Accessed: 2024-11-11. 5

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 3

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 2, 5

[41] Prafull Sharma, Julien Philip, Michaël Gharbi, Bill Freeman, Fredo Durand, and Valentin Deschaintre. Materialistic: Selecting similar materials in images. *ACM Trans. Graph.*, 42 (4), 2023. 4, 6

[42] Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik. Match: Differentiable material graphs for procedural material capture. *ACM Trans. Graph.*, 39(6):1–15, 2020. 1, 2

[43] Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik. Match: Differentiable material graphs for procedural material capture. *ACM Trans. Graph.*, 39(6):1–15, 2020. 2

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[45] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[46] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[47] Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. Controlmat: A controlled generative approach to material capture. *ACM Trans. Graph.*, 43(5), 2024. 1, 2, 5, 6, 7, 8

[48] Giuseppe Vecchio, Renato Sortino, Simone Palazzo, and Concetto Spampinato. Matfuse: controllable material generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4429–4438, 2024. 2, 5, 6, 7

[49] Huisi Wu, Wei Yan, Ping Li, and Zhenkun Wen. Deep texture exemplar extraction based on trimmed t-cnn. *IEEE Transactions on Multimedia*, 23:4502–4514, 2020. 2

[50] Bowen Xue, Claudio Guarnera, Shuang Zhao, and Zahra Montazeri. Reflectancefusion: Diffusion-based text to svbrdf generation. In *Eurographics Symposium on Rendering*. Eurographics Association, 2024. 2

[51] K. Yan, F. Luan, M. Hašan, T. Groueix, V. Deschaintre, and S. Zhao. Psdr-room: Single photo to scene using differentiable rendering. In *ACM SIGGRAPH Asia 2023 Conference Proceedings*, 2023. 2

[52] Yu-Ying Yeh, Zhengqin Li, Yannick Hold-Geoffroy, Rui Zhu, Zexiang Xu, Miloš Hašan, Kalyan Sunkavalli, and Manmohan Chandraker. Photoscene: Photorealistic material and lighting transfer for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18562–18571, 2022. 2

[53] Cheng Zhang, Bailey Miller, Kai Yan, Ioannis Gkioulekas, and Shuang Zhao. Path-space differentiable rendering. *ACM Trans. Graph.*, 39(4):143:1–143:19, 2020. 2

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[55] Xilong Zhou and Nima Khademi Kalantari. Look-ahead training with learned reflectance loss for single-image svbrdf estimation. *ACM Transactions on Graphics (TOG)*, 41(6):1–12, 2022. 1, 2

[56] Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. Tilegen: Tileable, controllable material generation and capture. In *SIGGRAPH Asia 2022 Conference Papers*, New York, NY, USA, 2022. Association for Computing Machinery. 2

MaterialPicker: Multi-Modal Material Generation with Diffusion Transformers

Supplementary Material

A. More Results

A.1. Image Conditioned Generation

We show qualitative and quantitative comparisons with Material Palette [28] on material extraction in Sec. 4.2 of the main paper. We provide more visual comparisons in Fig. 15 and Fig. 16. Since Material Palette generates only three material maps (albedo, normal, and roughness), we present results for these channels, along with the re-rendered images. We report the average CLIP-I metric ↑ and DINO metric ↑ between the output material maps and ground truth in Tab. 1 in the main paper. We further report the 95% confidence interval in Tab. 2 of the Supplemental Material. Our method achieves higher 95% confidence intervals for the vast majority of the generated material 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.

A.2. Ablation Study

A.2.1 Mask as Input or Output

As opposed to existing material generation models, our model doesn't require the target material to cover the entire input image [47] or manually-created masks [28] 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 ϵ_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. 11, we find that our proposed model, which automatically predicts a mask, performs comparably well to this variant requiring the mask as input.

A.3. Evaluations on the Robustness.

We further test our model's robustness under various factors, including different texture scales (in the main paper), varying levels of distortion, and diverse lighting conditions.

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 [28] and follow the texture processing steps outlined by Hao et al. [19]. We apply homography transformations [20] and thin plate spline transformations [5] to the textures. Our results in Fig. 12 show that the model 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 Fig. 17, Fig. 18, Fig. 19, Fig. 20, Fig. 21 and Fig. 22.

We further evaluate the model's performance when the input image contains specular highlights and shadows in Fig. 13. We see that these highlights and shadows in real photos do not "leak" into material maps, highlighting the

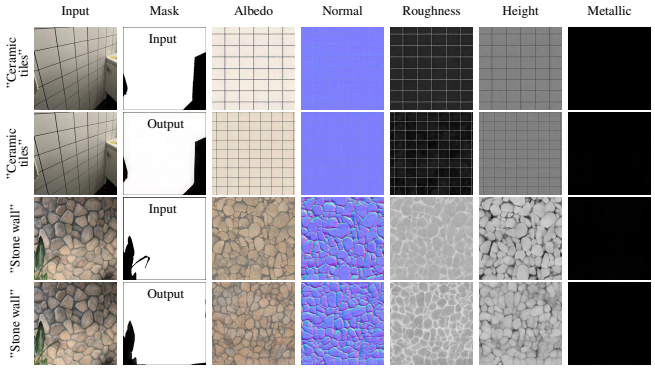

Figure 11. Impact of mask on material generation quality. Here shows the effect of using the mask as an input versus as an output on the quality of generated materials. Each pair of consecutive rows represents the results from the model with the mask as input (top row) and the model with the mask as 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.

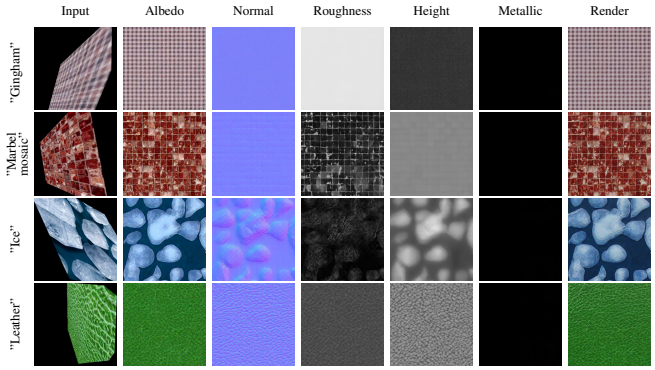

Figure 12. Evaluation of our model's robustness to varying levels of distortion. The first column shows textures transformed with homography and thin plate spline transformations. The following columns present the material maps and the rendering images. The results demonstrate that our model effectively rectifies textures with various patterns and different types of distortion, maintaining high-quality outputs.

Table 2. Quantitative results of material extraction. We compare with Material Palette [28] and present the 95% confidence intervals for the CLIP-I metric ↑ and the DINO metric ↑, calculated between the output material maps and the ground truths.

CLIP↑	Albedo	Normal	Roughness	Render
Mat-Palette	(0.814, 0.820)	(0.864, 0.870)	(0.788, 0.794)	(0.954, 0.956)
Ours	**(0.855, 0.860)**	**(0.872, 0.876)**	**(0.863, 0.869)**	**(0.966, 0.968)**
DINO↑	**Albedo**	**Normal**	**Roughness**	**Render**
Mat-Palette	**(0.494, 0.513)**	(0.622, 0.640)	(0.493, 0.512)	(0.792, 0.802)
Ours	(0.484, 0.503)	**(0.663, 0.680)**	**(0.556, 0.576)**	**(0.859, 0.867)**

model's robustness to various lighting conditions.

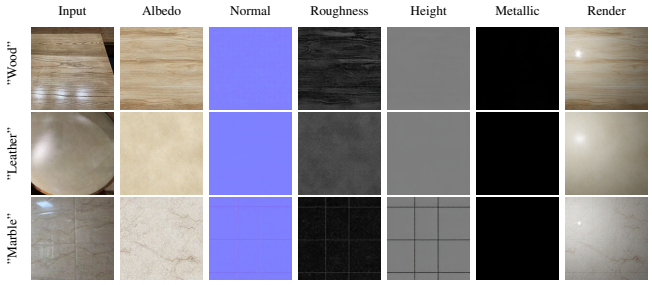

Figure 13. Evaluation of the robustness of our model to lighting and shadow interference. We test scenarios where the input photographs contain point light sources, shadows, or environmental reflections. The generated material maps and rendered images demonstrate the ability of our model to handle these interferences, preserving material quality and accurately representing the input photos. The leftmost side of each row is labeled with the text conditioning input used.

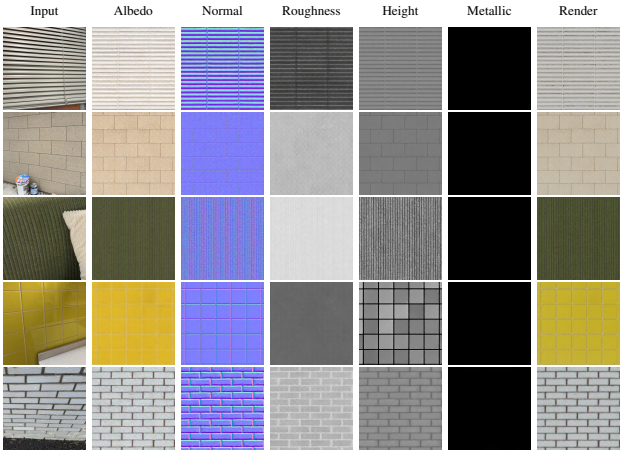

Figure 14. The material maps of the five samples used for texture rectification comparison in Fig. 5 of the main paper.

B. Additional Real Examples

In Fig. 14, we provide the material maps of the five examples used for texture synthesis in Fig. 5 of the main paper. In Fig. 17, Fig. 18, Fig. 19, Fig. 20, Fig. 21 and Fig. 22, we show additional examples of material extraction using our model, along with the uncropped original images from our real photographs evaluation dataset (Sec. 3.2). These examples include various indoor and outdoor materials captured under complex real-world lighting conditions. Our model generalizes well to real photos, producing renderings that are visually similar to the photographs and providing accurate masks, demonstrating our model's generalization capabilities.

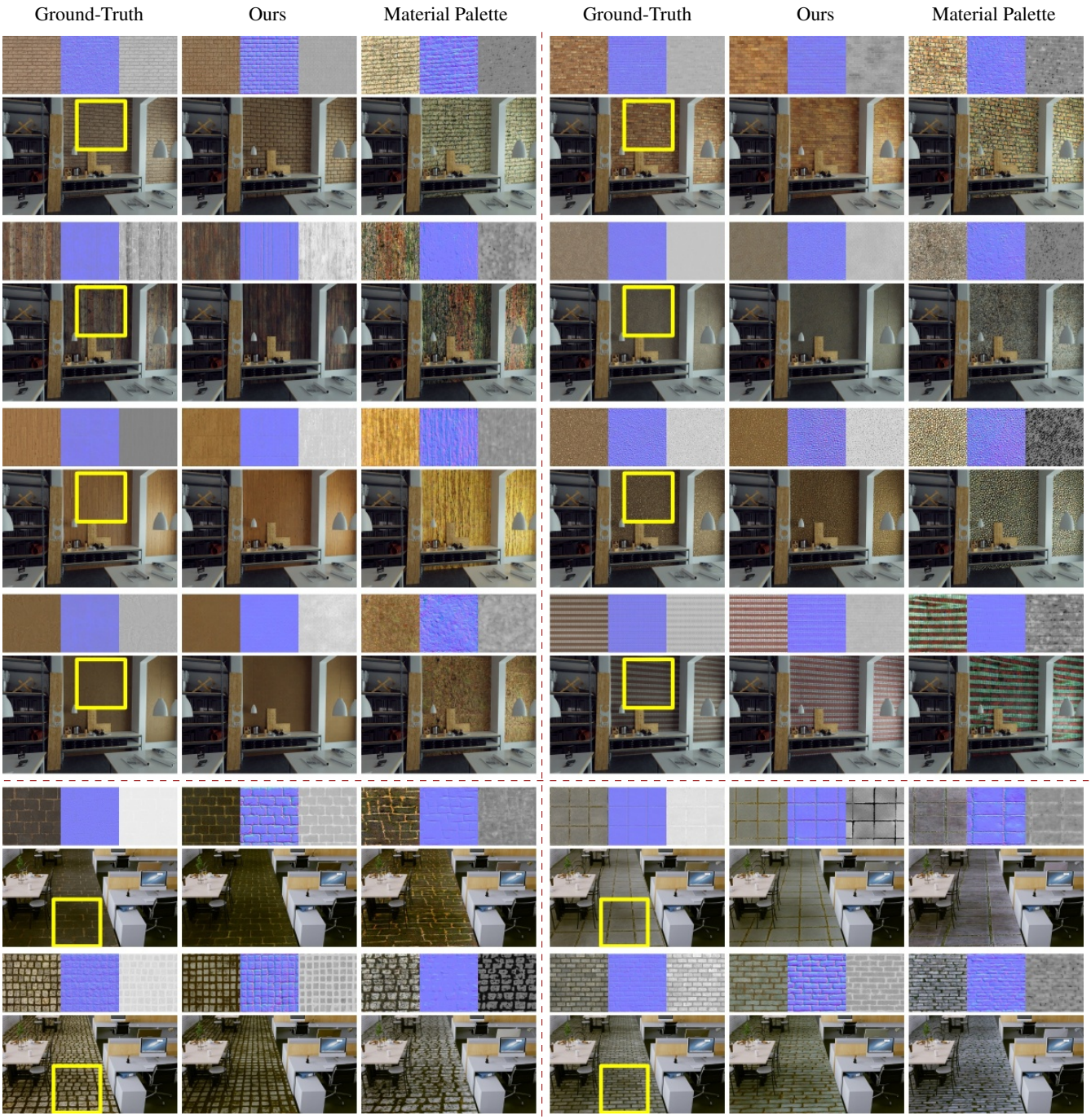

Figure 15. More results of comparisons with Material palette [28] on synthetic dataset (Sec. 4.1) for material extraction. Each row contains two sets of comparisons. In each set, the first column shows the ground truth material maps from PolyHaven, with the rendered scene below. The yellow square area 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.

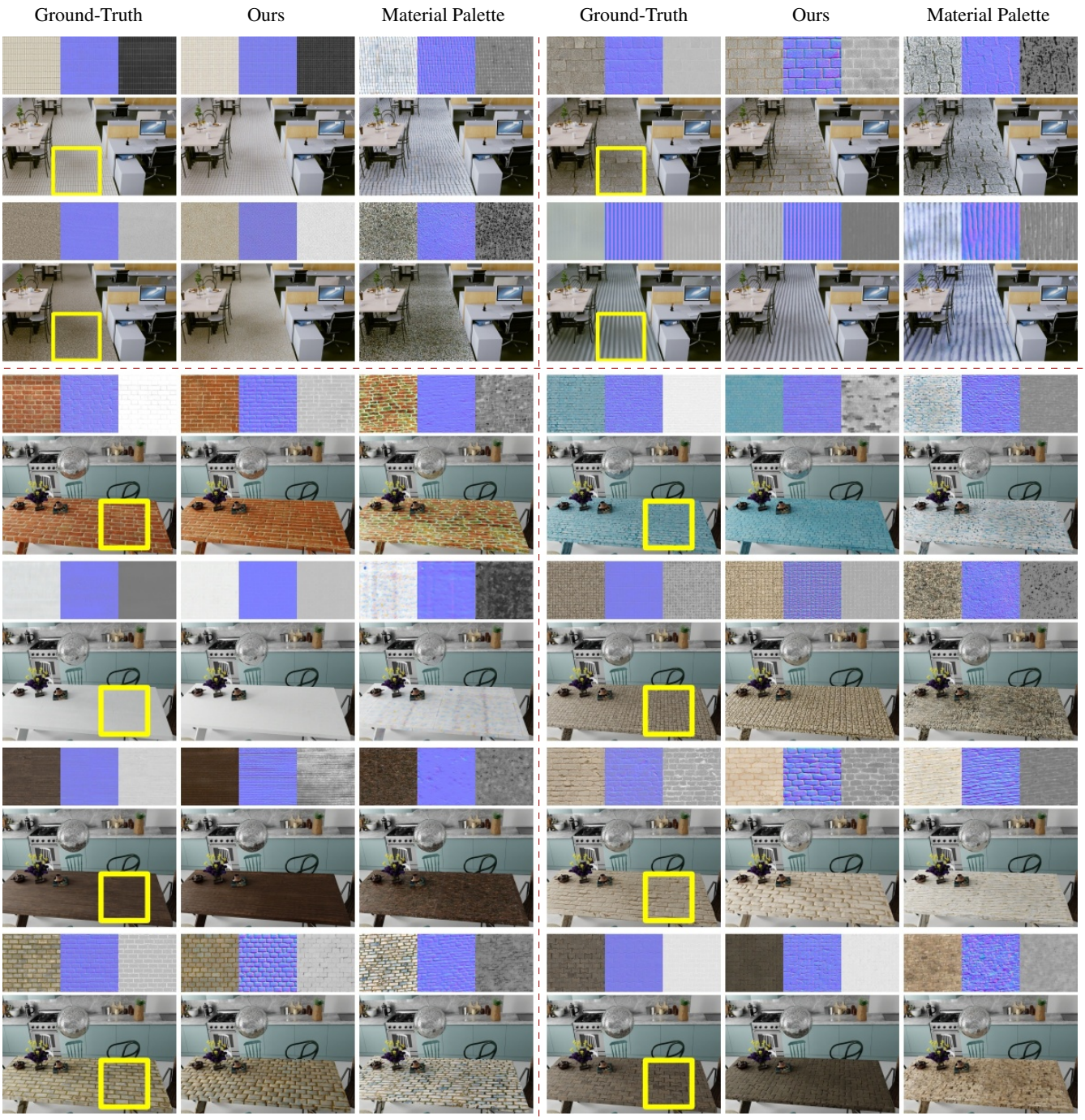

Figure 16. More results of comparisons with Material palette [28] on synthetic dataset (Sec. 4.1) for material extraction. Each row contains two sets of comparisons. In each set, the first column shows the ground truth material maps from PolyHaven, with the rendered scene below. The yellow square area 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.

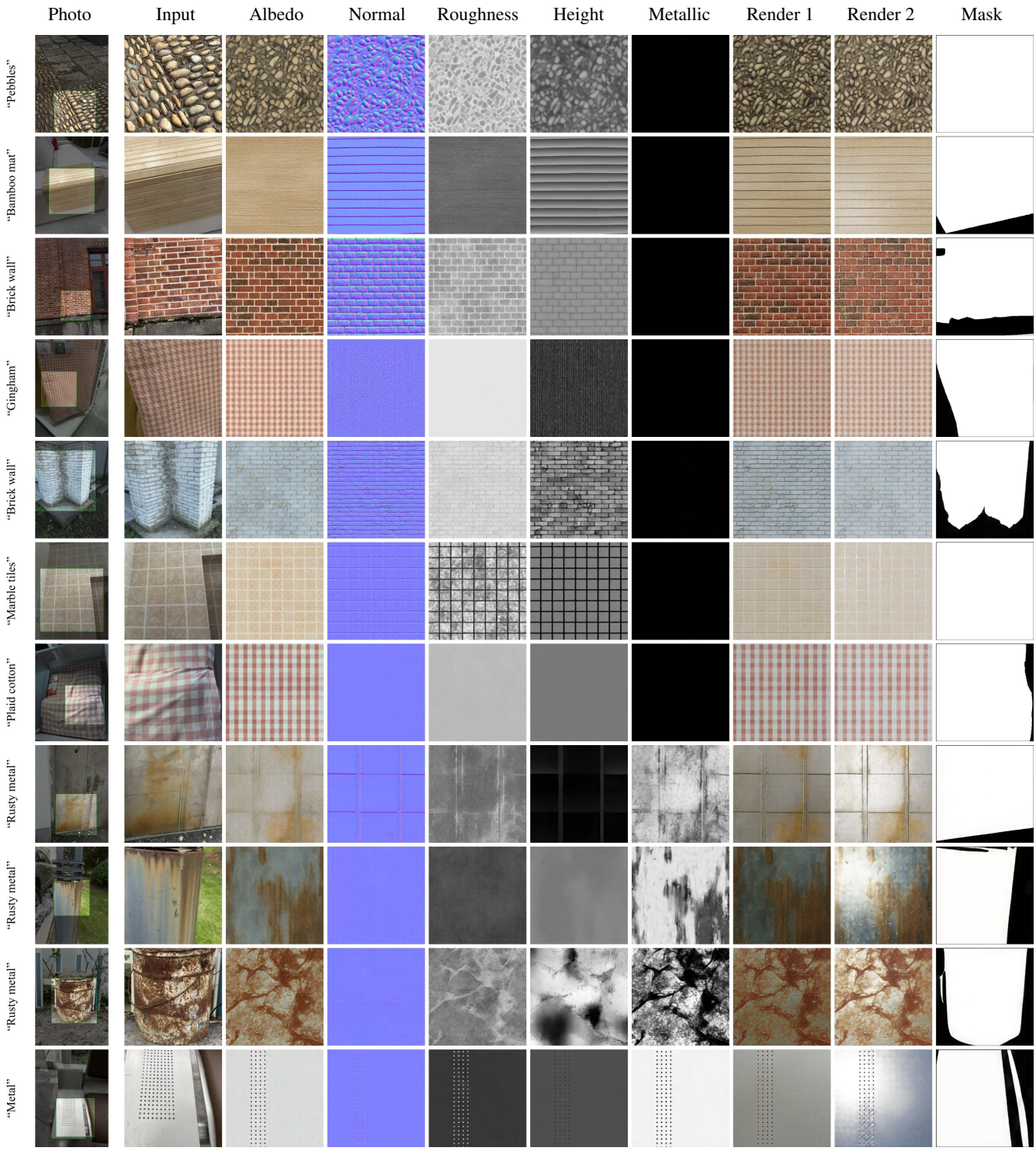

Figure 17. More results of our method on material extraction for real photographs. The first column shows real photographs captured by smartphones, before cropping (as described in Sec. 3.2). The box indicates the cropped area, which is the image actually used as input for the model in the second column. Third to the ninth columns show the generated material maps and rendering under two environment maps. The last column shows the mask of the dominant material location automatically predicted by our model. We conduct tests on diverse material types in both indoor and outdoor scenes, demonstrating the generalization capability of our model. The leftmost side of each row is labeled with the text conditioning input used.

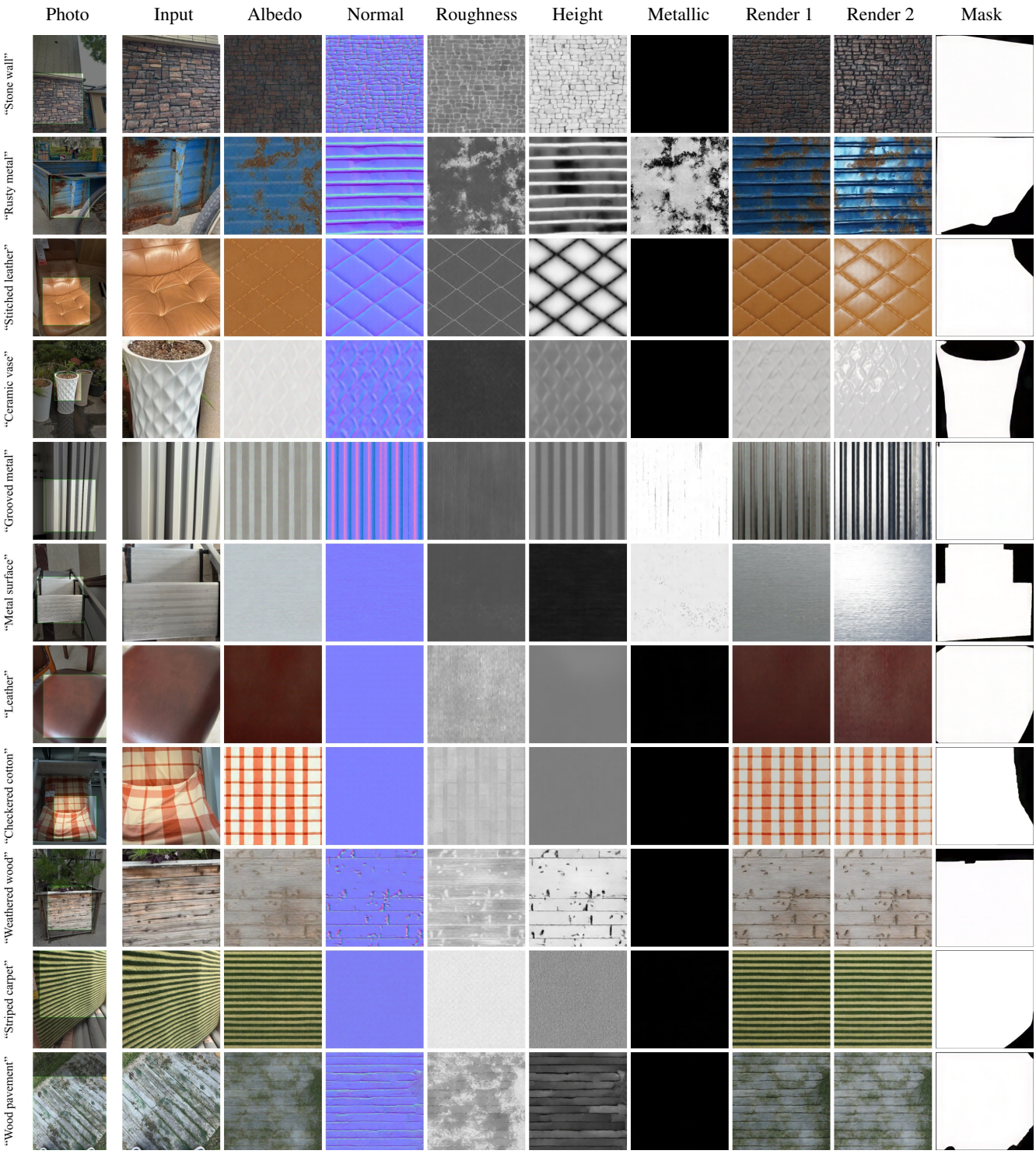

Figure 18. More results of our method on material extraction for real photographs. The first column shows real photographs captured by smartphones, before cropping (as described in Sec. 3.2). The box indicates the cropped area, which is the image actually used as input for the model in the second column. Third to the ninth columns show the generated material maps and rendering under two environment maps. The last column shows the mask of the dominant material location automatically predicted by our model. We conduct tests on diverse material types in both indoor and outdoor scenes, demonstrating the generalization capability of our model. The leftmost side of each row is labeled with the text conditioning input used.

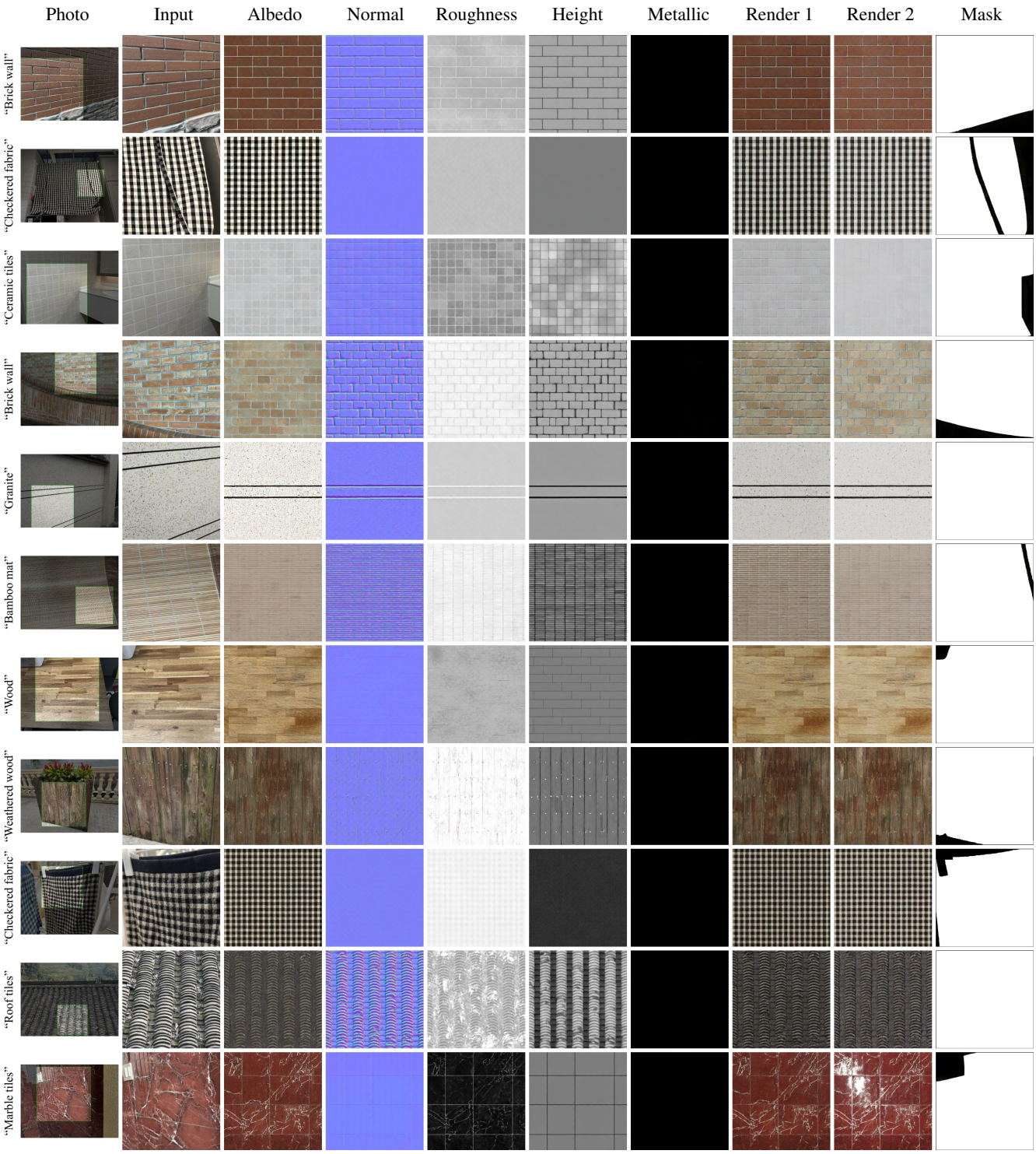

Figure 19. More results of our method on material extraction for real photographs. The first column shows real photographs captured by smartphones, before cropping (as described in Sec. 3.2). The box indicates the cropped area, which is the image actually used as input for the model in the second column. Third to the ninth columns show the generated material maps and rendering under two environment maps. The last column shows the mask of the dominant material location automatically predicted by our model. We conduct tests on diverse material types in both indoor and outdoor scenes, demonstrating the generalization capability of our model. The leftmost side of each row is labeled with the text conditioning input used.

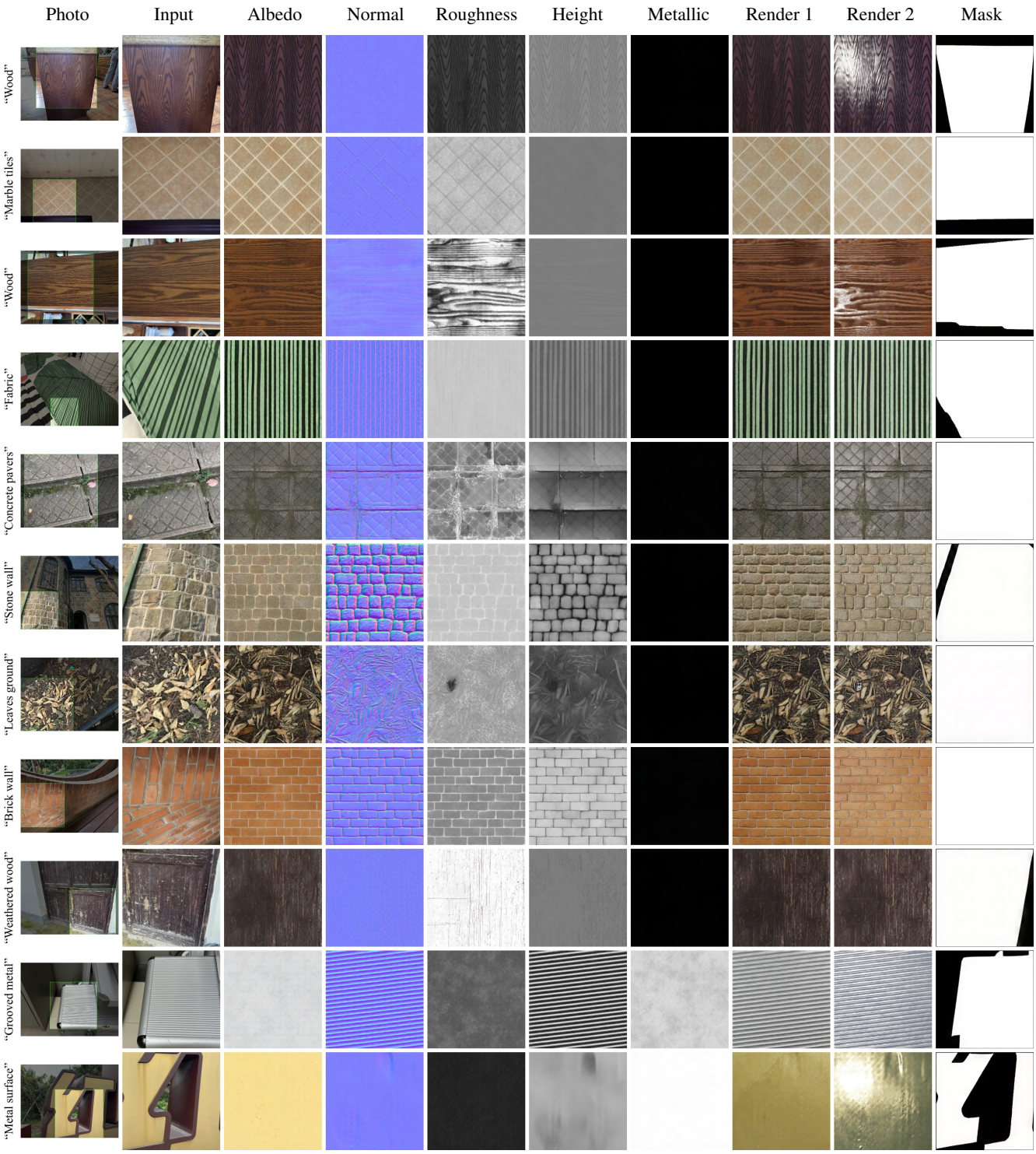

Figure 20. More results of our method on material extraction for real photographs. The first column shows real photographs captured by smartphones, before cropping (as described in Sec. 3.2). The box indicates the cropped area, which is the image actually used as input for the model in the second column. Third to the ninth columns show the generated material maps and rendering under two environment maps. The last column shows the mask of the dominant material location automatically predicted by our model. We conduct tests on diverse material types in both indoor and outdoor scenes, demonstrating the generalization capability of our model. The leftmost side of each row is labeled with the text conditioning input used.

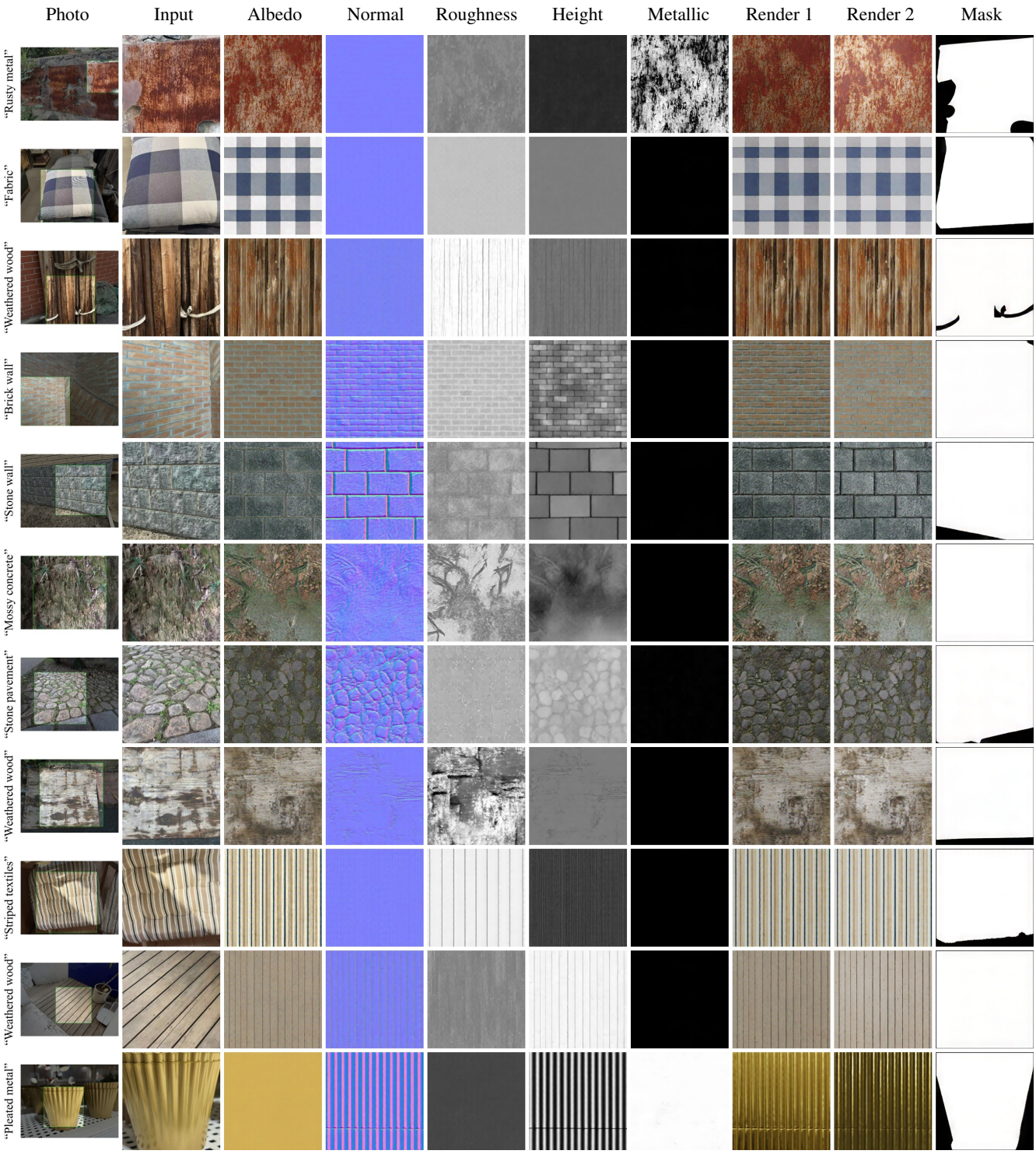

Figure 21. More results of our method on material extraction for real photographs. The first column shows real photographs captured by smartphones, before cropping (as described in Sec. 3.2). The box indicates the cropped area, which is the image actually used as input for the model in the second column. Third to the ninth columns show the generated material maps and rendering under two environment maps. The last column shows the mask of the dominant material location automatically predicted by our model. We conduct tests on diverse material types in both indoor and outdoor scenes, demonstrating the generalization capability of our model. The leftmost side of each row is labeled with the text conditioning input used.

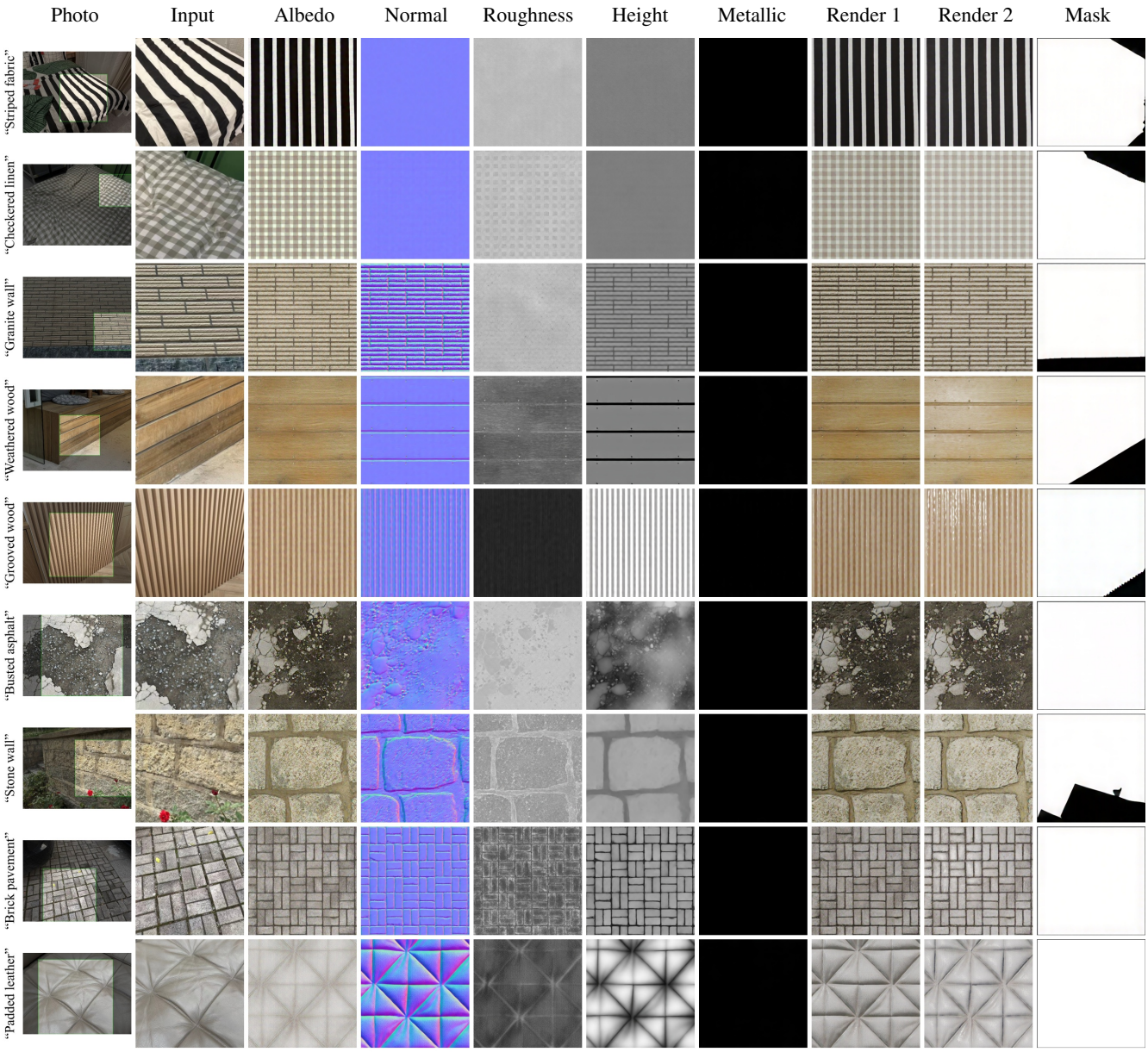

Figure 22. More results of our method on material extraction for real photographs. The first column shows real photographs captured by smartphones, before cropping (as described in Sec. 3.2). The box indicates the cropped area, which is the image actually used as input for the model in the second column. Third to the ninth columns show the generated material maps and rendering under two environment maps. The last column shows the mask of the dominant material location automatically predicted by our model. We conduct tests on diverse material types in both indoor and outdoor scenes, demonstrating the generalization capability of our model. The leftmost side of each row is labeled with the text conditioning input used.