TextureDreamer: Image-guided Texture Synthesis through Geometry-aware Diffusion

Supplementary Material

Figure 10. **Diversity** of synthesized textures.

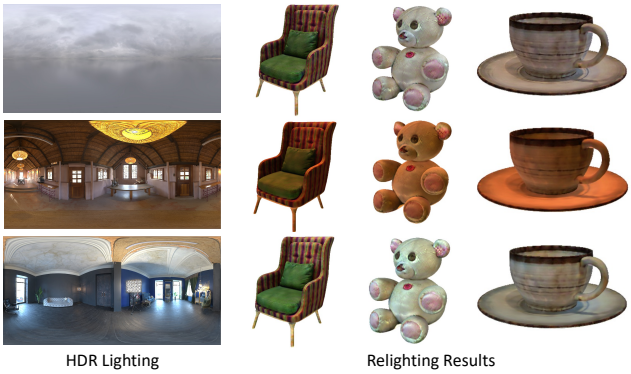

Figure 11. **Example of relighting results.** The textures are relit by the original HDR environment maps (first row) and the novel maps (second and third rows).

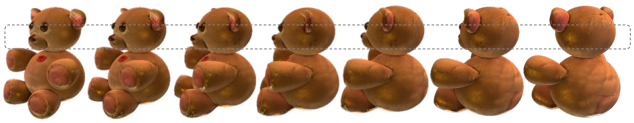

Figure 12. **View-dependent specular highlights.**

6. Additional Experiments

6.1. Texture Diversity

By using different random seeds, our framework can generate diverse textures, as shown in Figure 10.

6.2. Relightable and View-dependent Texture

Since our synthesized texture contains albedo, metallic, and roughness maps, the target objects with the synthesized appearance can be relit, as shown in Figure 11.

As mentioned in discussion, our method may suffer from lighting/specularity baked-in issues and tend to obtain more diffuse textures. However, we still observe view-dependent specular hightlights for an example with low roughness and high metallic under novel HDR lighting in Figure 12.

Figure 13. **Captured shapes.** Input images are from Figure 6.

Figure 14. **Additional Ablation study.** If replacing generic diffusion model ϕ with personalized model or applying classifier guidance scale 7.5, some random patterns might appear in the synthesized texture. If we freeze the camera encoder ρ, the result might be worse or more noisy than our full method.

6.3. Captured Shapes

In Figure 13, we show transfer results on *captured* meshes of solid art objects from a commercial 3D scanner [34]. Captured shapes contain imperfect surfaces (highlighted in a circle), which might lead to unexpected patterns or defects.

6.4. Additional Ablation Studies

We analyze the effectiveness of ControlNet and the design of score distillation in Section 4.3. We perform an additional ablation study in Figure 14. If we replace generic diffusion model ϵ_ϕ with the personalized diffusion model ϵ_ψ or apply classifier free guidance weight 7.5, the result tends to introduce random patterns which does not exist in the input images. If we choose to freeze the camera encoder weights ρ, the result becomes worse or more noisy than our full method.

Table 3. **Ablation study** on image-based texturing w.r.t. CLIP image-based feature similarity. Although *w/o ControlNet* and *w/ ControlNet (Depth)* achieve higher similarity score, the transfer results tend to ignore target shape and directly paint the texture without reasoning the geometry. Among the remaining ablative methods, our full method achieves the highest CLIP similarity w.r.t. reference images.

	CLIP similarity ↑
w/o ControlNet	*0.8394*
w/ ControlNet (Depth)	*0.8320*
SDS, w/o CFG	0.8101
SDS, CFG 100	0.7983
w/o LoRA removed	0.8110
Personalized model as ϕ	0.8218
CFG weight as 7.5	0.8218
w/o camera encoder ρ updated	0.8267
Ours	**0.8296**

We also quantitatively evaluate the importance of each component in our system, as shown in Table 3. We use image-based CLIP feature to measure the similarity between reference images and the rendered images. To ensure fair evaluation, the background of both reference and rendered images are masked with white color. Our full method achieves the highest similarity score among the ablative baselines except *w/o ControlNet* and *w/ ControlNet (Depth)*. As shown in Figure 7, these two methods tend to ignore the target shape and directly paint the texture without adapting to geometry. Thus, they could reach higher score by painting the original texture regardless of the shape. We also observe that SDS results tend to be saturated or blurry and cannot recover the texture from the inputs. Keeping LoRA in the generic diffusion model ϵ_ϕ will introduce random patterns to the synthesized texture.