4OTex: 4D Object Texturing from Diffusion Priors

Qianru Li, Yunfei Deng
Technical University of Munich
qianru.li@tum.de, philips.deng@tum.de

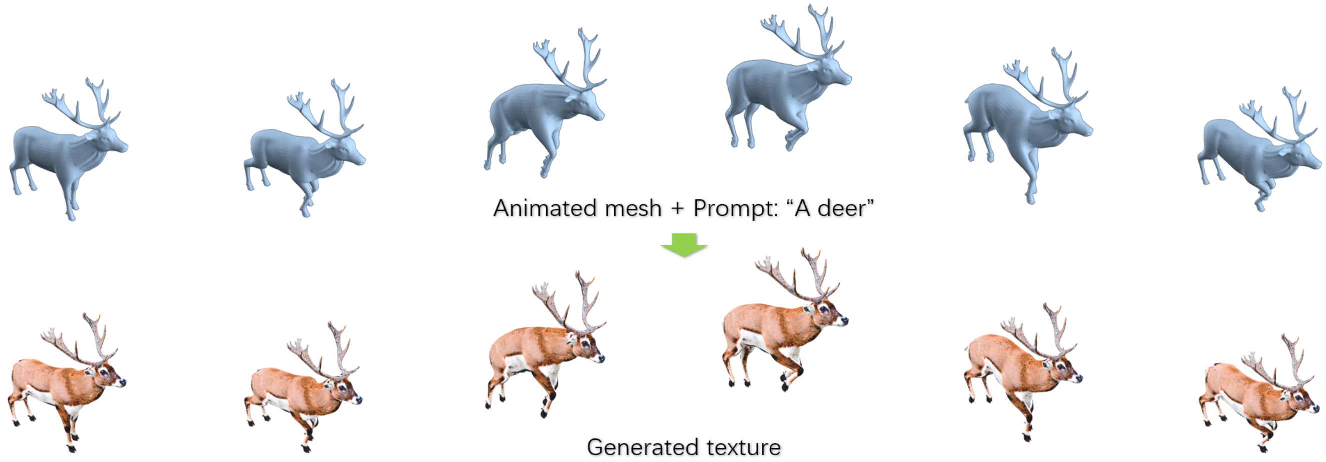

Figure 1. We introduce 4OTex, a text-driven texture synthesis architecture for 4D objects. Given an animated mesh and the text prompt as input, 4OTex generates high-quality texture via depth-to-image diffusion priors.

Abstract

We present 4OTex, a generative method for realistic texture generation on animated meshes from diffusion priors. Existing approaches suffer from shading effects and focus on static object texturing, while this work deals with these problems with physically based rendering (PBR) representation and training with multi-frame inputs. Moreover, a timestep selection mechanism is used for coarse-to-fine texture producing which stabilizes the training process. Our method outputs high-quality textures that are compatible with the rendering pipeline of mainstream game engines. Experimental results on DeformingThings4D animations showcase the effectiveness of our method.

1. Introduction

Generating textures for animated meshes requires extensive manual labor and complex workflow using traditional techniques for animation creation. With the priors from a large pretrained model, this process can be simplified and automated. Existing works [2, 3] suffer from shading effects due to the usage of a single texture map, which does not take the lighting influence into consideration. Moreover, previous works [4, 11, 15] focus on static object texturing, which inspires us to explore the creation of textures on animated objects. In this work, we propose a 4D object texturing pipeline that can update the appearance of the animated surface iteratively. Our proposed 4OTex outperforms existing methods for high-quality texture creation. We summarize our technical contributions as follows.

- We leverage a Physically Based Rendering Representation as texture field for shading effects removal.
- We design a efficient pipeline that distills high-quality surface information from depth-conditioned diffusion priors.
- With timestep selection and multi-frame training, we incorporate a coarse-to-fine mechanism to accelerate the training process, and a simple strategy for input with multiple frames.

2. Related Work

2.1. Texture Generation on Static Objects

Recently, texture generation is becoming a notable trend in text-to-3D generation. TEXTure [11] and Text2Tex [2] leverage a iteratively projection and refinement method for texturing a given mesh using depth-conditioned diffusion models. However, their results are inherited with noticeable artifacts at viewpoint switching areas. SceneTex [3] proposes a texturing pipeline for rooms which generates RGB texture maps containing baked-in highlights or shadows, disabling realistic rendering in downstream tasks. To tackle with this limitation, Fantasia3D [4] and DreamMat [15] incorporate BRDF to generate more realistic appearances. Our method uses a similar PBR representation while taking directly the 2D texture coordinates as input for inference rather than 3D point positions.

2.2. Texture Generation on Animated Objects

Tex4D [1] tackles similar task with ours leveraging a latent aggregation in UV space and a video diffusion model to enhance spatial and temporal consistency. However, the generated 4D texture using this method lacks details. We optimize the visual quality of the texture through a multi-resolution strategy. Make-It-Vivid [13] first generates the texture for the static object with the neutral pose, then animates the object with text captioning. In comparison, our method takes multi-frame meshes as input for training, which reduces the artifacts on the generated texture through motion.

3. Method

The proposed pipeline is shown in Figure. 2.

3.1. Physically Based Rendering Representation

To reduce shading effects on generated texture, 4OTex applies a Physically Based Rendering representation for texture field. As shown in Figure. 3, PBR model contains three materials [4, 9]: base color image k_d, roughness and metallic image $k_{rm} \in \mathbb{R}^2$, and a normal variation image $k_n \in \mathbb{R}^3$. Given a 2D position $p \in \mathbb{R}^2$ in texture uv coordinate generated by xatlas, we leverage a multi-resolution hashgrid embedding β combined with MLP Γ parameterized by γ to perform the mapping $p \rightarrow (k_d, k_{orm}, k_n)$, which can be formulated as:

$$(k_d, k_{rm}, k_n) = \Gamma(\beta(p), \gamma). \tag{1}$$

Once the PBR materials are predicted, a Bidirectional Reflectance Distribution Function (BRDF) [5, 7] f is utilized for rendering the pixel color L along the direction of a camera pose w given uv position p:

$$\begin{aligned} L(p,w) &= L_d(p,w) + L_s(p,w) \\ &= k_d(1-m)\int_\Omega L_i(p,w_i)(w_i \cdot n_p)\,dw_i \\ &\quad + \int_\Omega \frac{DFG}{4(w \cdot n_p)(w_i \cdot n_p)} L_i(p,w_i)(w_i \cdot n_p)\,dw_i, \end{aligned} \tag{2}$$

where n_p denotes the surface normal at p, $\Omega = \{w_i : w_i \cdot n_p \geq 0\}$ represents a hemisphere with the incident direction w_i. The Fresnel term, the shadowing-mask term, and the GGX distribution of normal are given by F, G, and D respectively. L_i is the incident light predefined by an environment map.

By aggregating the computed pixel values from sampled uv positions in one view, the rendered image for this camera pose $x = \{L(p,w)\}$ is generated for further updating. Moreover, the rendered images from multi-view camera poses can also be back-projected to the texture space to generate a single texture map combining k_d, k_{rm}, k_n, and L_i.

3.2. Variational Score Distillation

4OTex adopts a pre-trained Diffusion model with depth condition as priors to update the PBR texture. The Variational Score Distillation (VSD) [14] process is composed with two stages, the objective of the first stage can be formulated as:

$$\mathcal{L}_{\text{VSD}}(\gamma) := \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_{\phi_{\text{pre}}}(x;y,d,t) - \epsilon_\phi(x;y,d,t))\frac{\partial x}{\partial \gamma}], \tag{3}$$

where t represents the chosen timestep, whose selection mechanism is explained in detail in Sec. 3.3. y and d are the given prompt and the depth condition, $w(t) = \sqrt{1-\bar{\alpha}_t}$ denotes the loss weight at chosen timestep t, in which $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. ϕ_{pre} and ϕ represent the pre-trained diffusion model with depth control and the trainable LoRA model separately, which in this stage are both frozen. $\epsilon_{\phi_{\text{pre}}}$ and ϵ_ϕ are the noise predicted by the two models respectively. In the second stage, the LoRA model is unfrozen for training. The updating loss for this stage is formulated as:

$$\mathcal{L}_{\text{LoRA}}(\phi) = \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_\phi(x;y,d,t) - \epsilon)], \tag{4}$$

where $\epsilon \sim \mathcal{N}(0,1)$ is the randomly sampled noise.

3.3. Timestep Selection

Following DreamTime [6], the first part of timestep selection mechanism involves computing the weighting function, which is formulated as:

$$\begin{aligned} W(t) = \frac{1}{Z} W_d(t) \cdot W_p(t), Z = \sum_{t=1}^{T} W_d(t) \cdot W_p(t), \\ W_d(t) = \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}, W_p(t) = e^{-\frac{(t-T/2)^2}{2s^2}}, \end{aligned} \tag{5}$$

where T is the number of diffusion steps, which in this case is 1000, and s is empirically set as 25. After acquiring the weighting function $W(t)$, the second part of timestep selection is getting the sampled timestep for each iteration i (number of iterations: N) via:

$$t(i) = \arg\min_{t'} |\sum_{t=t'}^{T} W(t) - \frac{i}{N}|. \tag{6}$$

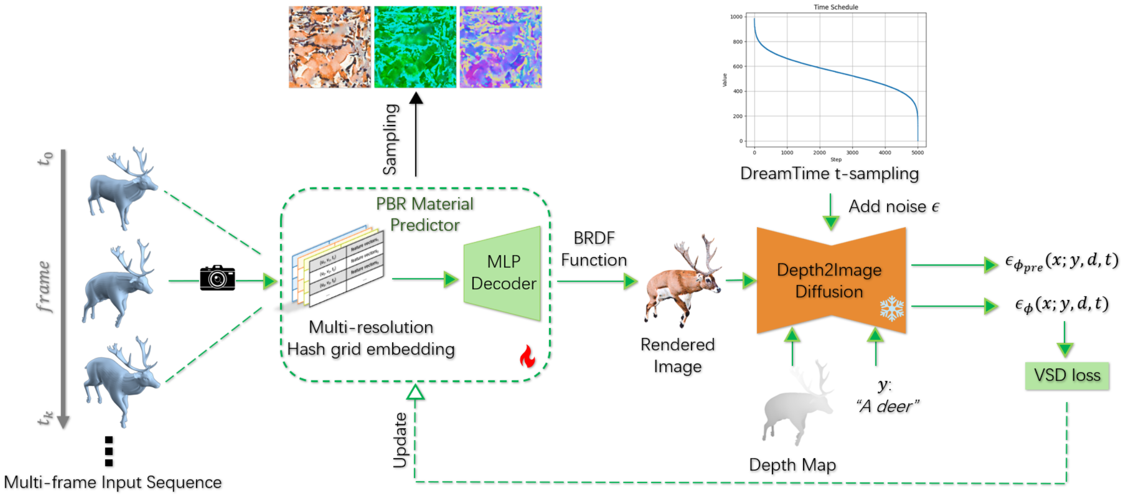

Figure 2. **Overview of 4OTex pipeline.** 4OTex takes multi-frame animated mesh sequence as input. For each iteration, one specific view of the chosen frame is rasterized and passed to the PBR material predictior, which contains a multi-resolution hash grid embedding and a MLP, mapping the uv coordinate to PBR materials. The predicted materials are then leveraged for rendering the RGB image via BRDF function. Through a depth-conditioned diffusion model and a effective t-sampling method, the predicted noises both from frozen Stable Diffusion and the LoRA module are used for VSD loss computing, which updates the PBR material predictor.

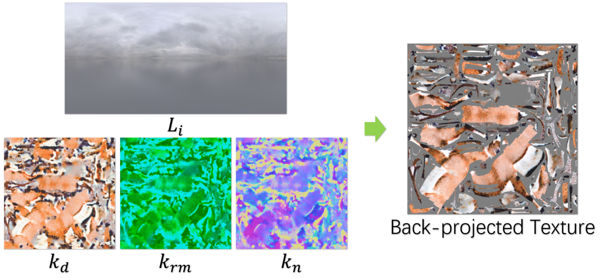

Figure 3. We represent the object surface as a set of spatially varying materials and an environment map following a standard PBR model, which can then be combined to be back-projected to one single RGB texture without shading effects.

This weighted non-increasing t-sampling strategy (illustrated in Figure. 2) effectively narrows down the solution space in a coarse-to-fine manner, significantly accelerating the convergence process. Additionally, by aligning the diffusion sampling steps closely with the hierarchical nature of 3D optimization, it leads to higher-quality reconstructions with improved detail fidelity.

3.4. Training with Multi-frame Inputs

To guide the texture adjust to the motion changes on animated meshes, for each iteration, 4OTex takes randomly-chosen mesh frame from all frames as input for training. Since the model sees more information on the whole sequence, this simple multi-frame training strategy can reduce artifacts (asymmetric patterns) on the generated texture, in particular for asymmetric motions.

4. Experiments

4.1. Dataset and Experimental Setup

We evaluate our method using the DeformingThings4D dataset [8], comprising 1,972 animated sequences across 31 categories. Specifically, we chose three representative objects: a ground dragon, a flying dragon, and a jumping deer.

In training, viewpoints are randomly sampled at each step, with elevation angles drawn from $U(0°, 60°)$ and azimuth angles from $U(-170°, 180°)$. The field of view gradually narrows from $50°$ to $30°$ following a coarse-to-fine strategy, initially capturing global features and later refining details. The training runs for 5,000 iterations, using learning rates of 1×10^{-3} (general optimization) and 1×10^{-4} (LoRA fine-tuning). All computations utilize FP16 precision to optimize GPU memory and speed without loss of performance.

Experiments are conducted on a single NVIDIA L40 GPU (48 GB VRAM), converging within roughly 1 hour.

4.2. Comparisons with Other Methods

We benchmark the proposed method against several texture synthesis approaches, namely DreamMat [15] and Text2Tex [2]. To thoroughly evaluate these methods, we selected 3 distinct objects from the DeformingThings4D dataset [8], each accompanied by one textual prompt.

4.2.1 Quantitative Analysis

In Tab. 1, we show that our method outperforms the compared methods with higher values of metrics.

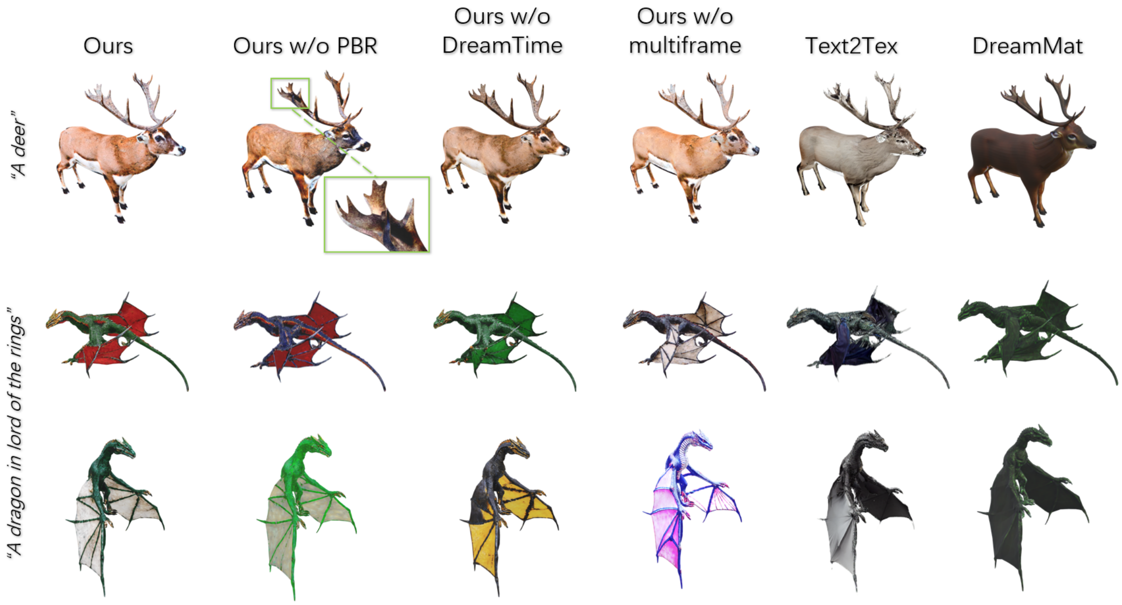

Figure 4. **Ablation study and comparison with other methods.**

Method	CLIP ↑	IS ↑
DreamMat [15]	27.94	2.64
Text2Tex [2]	27.83	2.41
4OTex (Ours) w/o PBR	27.44	2.71
4OTex (Ours) w/o DreamTime	27.42	3.00
4OTex (Ours) w/o multiframe	26.28	2.81
4OTex (Ours)	**28.24**	**3.14**

Table 1. **Quantitative results.** We introduce the CLIP score (CLIP) [10] and the Inception Score (IS) [12] for quantitative comparisons to assess how well the generated textures align with the input prompts and to quantify the visual quality of the synthesized textures. Our findings indicate that our method produces textures of the highest quality.

4.2.2 Qualitative Results

In Fig. 4, visual inspections confirm that our method produces textures with more detailed and realistic features, closely adhering to the textual descriptions. Notably, textures generated by our approach exhibit fewer artifacts and improved coherence, especially when applied to dynamically deforming objects.

4.3. Ablation Study

We conduct ablation studies to validate the effectiveness of the key components of our method, namely, multi-frame input training, DreamTime timestep sampling, and PBR representation. Quantitative comparisons for these modules are summarized in Tab. 1, which demonstrates the effectiveness of all the key components. The qualitative results of the ablation experiments in Figure. 4 show that with PBR representation, the shading effects are removed, especially compared with the rectangular area.

5. Conclusion

We introduce 4OTex, a text-driven texture synthesis architecture for 4D objects. Given an animated mesh and the text prompt as input, 4OTex generates high-quality texture via depth-to-image diffusion priors. At its core, 4OTex applies physically based rendering (PBR) representation and training with multi-frame inputs to improve the texture quality. Moreover, a timestep selection mechanism is used for coarse-to-fine texture producing which stabilizes the training process. Our method outputs high-quality textures on DeformingThings4D animations.

6. Discussion

Despite the promising results, our method still has several limitations. First, the proposed method is computationally intensive, requiring a significant amount of GPU memory and computational resources, which further increase the training time for generating the texture. Second, although shading effects can be removed using the PBR representation, another issue, namely asymmetric textures on object surfaces, remains challenging. We believe with multi-view aware diffusion priors, this limitation can be mitigated.

References

[1] Jingzhi Bao, Xueting Li, and Ming-Hsuan Yang. Tex4d: Zero-shot 4d scene texturing with video diffusion models. *arXiv preprint arXiv:2410.10821*, 2024. 2

[2] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18512–18522, 2023. 1, 3, 4

[3] Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21081–21091, 2024. 1, 2

[4] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 1, 2

[5] Robert L Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1):7–24, 1982. 2

[6] Yukun Huang, Jianan Wang, Yukai Shi, Boshi Tang, Xianbiao Qi, and Lei Zhang. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In *ICLR*, 2024. 2

[7] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 2

[8] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12706–12716, 2021. 3

[9] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 2

[10] 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 *ICLR*, 2021. 4

[11] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 1

[12] Edward J Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*, pages 87–96. PMLR, 2017. 4

[13] Junshu Tang, Yanhong Zeng, Ke Fan, Xuheng Wang, Bo Dai, Kai Chen, and Lizhuang Ma. Make-it-vivid: dressing your animatable biped cartoon characters from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6243–6253, 2024. 2

[14] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36:8406–8441, 2023. 2

[15] Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, et al. Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 1, 2, 3, 4