Tactile DreamFusion: Exploiting Tactile Sensing for 3D Generation

Ruihan Gao[1] Kangle Deng[1] Gengshan Yang[1] Wenzhen Yuan[2] Jun-Yan Zhu[1]
[1]Carnegie Mellon University [2]University of Illinois Urbana-Champaign
https://ruihangao.github.io/TactileDreamFusion/

***Abstract*—3D generation methods powered by 2D diffusion priors often struggle to synthesize realistic geometric details, resulting in overly smooth surfaces or inaccurate geometry baked into albedo textures. We present TactileDreamFusion, a framework that incorporates tactile sensing as an additional modality to enhance geometric details in 3D generation. By learning a unified 3D texture field and jointly refining visual textures guided by high-resolution tactile normals, our approach produces high-fidelity textures with accurate alignment between appearance and geometry. We demonstrate results on both text-to-3D and image-to-3D tasks.**

***Index Terms*—visual-tactile synthesis, 3D generation**

I. Introduction

Generating high-fidelity 3D assets is essential for applications in gaming, VR/AR, and robotics simulation. Recent advances in generative models [1], [2], neural rendering [3], [4], and large-scale datasets [5], [6] have enabled 3D asset creation from a single image [7] or a text prompt [8], [9]. However, these methods often produce overly smooth geometry or bake fine details into the albedo map without true surface variation.

Two key challenges remain: the lack of high-resolution geometry in current datasets [5], [6] and the difficulty of specifying geometric textures in language. To address these gaps, we propose leveraging tactile sensing to capture fine-grained surface detail for 3D generation.

Given a text prompt or input image, we generate a base mesh with an albedo map and capture tactile normals using GelSight sensors [10], [11]. These tactile signals are converted into normal maps, and we train a TextureDreambooth to synthesize diverse texture patches. A lightweight 3D texture field is learned to jointly optimize visual and tactile textures using diffusion-based guidance.

Our method further enables multi-part texture synthesis by aggregating 2D diffusion-based part segmentation into a 3D label field. Experiments demonstrate that our approach produces coherent, high-resolution geometry and textures with accurate cross-modal alignment.

II. Method

Tactile Data Acquisition. To capture high-resolution geometric details, we use the GelSight sensor [10], [11], which applies photometric stereo to measure fine surface normals at micrometer resolution. We collect tactile data by pressing the sensor against object surfaces and process the output height maps with high-pass filtering to isolate high-frequency textures. Non-contact regions are masked out, and the resulting height maps are converted back to normal maps.

The project is partially supported by the Amazon Faculty Research Award, Cisco Research, and the Packard Fellowship. Kangle Deng is supported by the Microsoft Research PhD Fellowship. Ruihan Gao is supported by the A*STAR National Science Scholarship (Ph.D.).

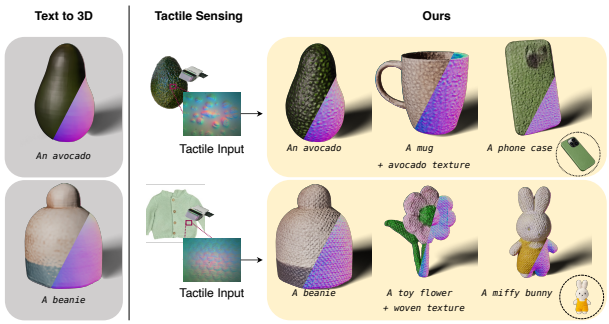

Fig. 1. Our method refines 3D generation using tactile input, improving geometric details and visual realism.

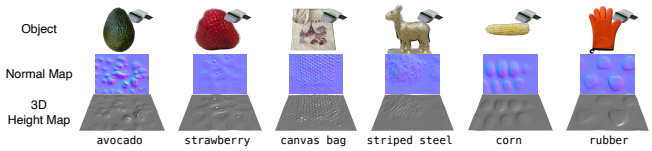

Fig. 2. **`TouchTexture` dataset.** We collect tactile normal data from 18 daily objects featuring diverse tactile textures. To demonstrate the local geometric intricacies, we show the tactile normal map and a 3D height map for each object. Please refer to the supplement for the full set of our data.

Figure 2 shows examples of our `TouchTexture` dataset, which contains 18 daily objects with various tactile textures. These tactile exemplars are used for both initialization and guidance of our texture field learning.

Base Mesh and Texture Field. Given a text or image prompt, we generate a base mesh M with an albedo UV map using Wonder3D [12]. To incorporate tactile signals, we learn a 3D texture field $\beta(p) = (\mathbf{c}, \mathbf{n}_{\text{T}})$ using multi-resolution hash encoding [13], where $\mathbf{c}$ is albedo and $\mathbf{n}_{\text{T}}$ is the tactile normal.

Rendering is performed via a differentiable rasterizer $\mathcal{R}$ [14], compositing mesh normals $\mathbf{n}_{\text{B}}$ with tactile normals $\mathbf{n}_{\text{T}}$ via the TBN matrix for shading:

$$\mathbf{n} = \mathbf{Q}_{\text{TBN}} \cdot \mathbf{n}_{\text{T}}. \tag{1}$$

Texture Refinement with Tactile Guidance. We optimize the texture field with reconstruction and diffusion-based refine-

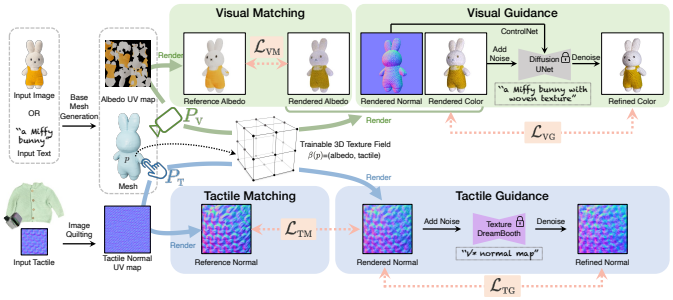

Fig. 3. **Method overview.** Given an input image or text prompt, our method generates a mesh with high-quality visual and normal textures using a 3D texture field. We jointly optimize albedo and tactile normal textures with diffusion-based visual and tactile guidance, leveraging distinct camera sampling for visual and tactile supervision. A customized Texture Dreambooth further refines tactile alignment to match the input exemplars.

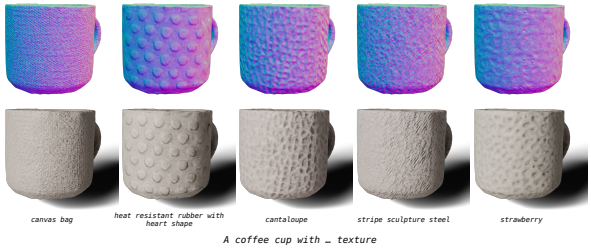

Fig. 4. **Diverse textures with the same object.** With additional texture cues from tactile data, we can synthesize diverse textures with the same coarse shape for customized designs.

ment losses:

$$\mathcal{L} = \lambda_{VM}\mathcal{L}_{VM} + \lambda_{TM}\mathcal{L}_{TM} + \lambda_{VG}\mathcal{L}_{VG} + \lambda_{TG}\mathcal{L}_{TG}. \quad (2)$$

The visual matching loss $\mathcal{L}_{VM}$ supervises the albedo field against the UV-projected texture. The tactile matching loss $\mathcal{L}_{TM}$ aligns the tactile normal field with synthesized normal maps from tactile exemplars using image quilting [15].

For refinement, we employ a normal-conditioned ControlNet [16] for visual guidance ($\mathcal{L}_{VG}$) and a LoRA-finetuned Texture DreamBooth [17], [18] for tactile guidance ($\mathcal{L}_{TG}$), both implemented as multi-step denoising diffusion processes:

$$\mathcal{L}_{VG} = \left|\left|\hat{\mathbf{I}}_{\mathrm{c}} - \mathbf{I}_{\phi}\right|\right|_1 + \mathcal{L}_{LPIPS}, \quad \mathcal{L}_{TG} = 1 - \cos(\hat{\mathbf{I}}_{\mathrm{T}}, \mathbf{I}_{\psi}). \quad (3)$$

Multi-Part Texture Assignment. To support multi-material generation, we segment object parts via diffusion attention maps [19], [20], merging multi-view results into a 3D part label field supervised by a cross-entropy loss. This allows part-specific tactile supervision by adapting $\mathcal{L}_{TM}$ and $\mathcal{L}_{TG}$ with per-part masks.

Our approach unifies tactile sensing and diffusion priors to enable 3D texture generation with aligned visual and geometric details.

III. Experiment

We validate our method on text-to-3D and image-to-3D generation tasks using the `TouchTexture` tactile dataset, consisting of 18 real-world materials such as strawberry skin, striped steel, and canvas bag. Each material includes five high-resolution GelSight [10], [11] tactile patches with paired descriptions. One patch is used to initialize the texture field via image quilting [15], and the remaining are used to train Texture DreamBooth.

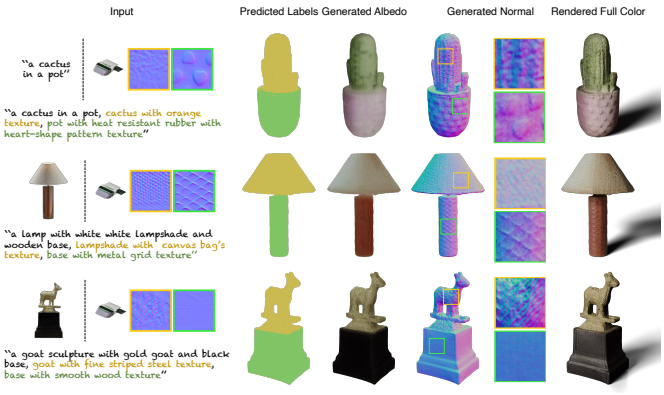

Fig. 5. **Multi-part texture generation.** Our method assigns different tactile textures to object parts specified by text or image prompts. We show predicted labels, albedo, normals, and full-color renderings, with zoom-in patches highlighting the generated normal textures.

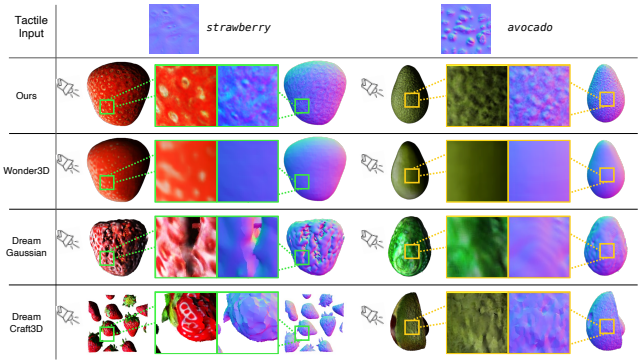

Fig. 6. **Baseline comparison.** Compared to the SOTA image-to-3D (Wonder3D and DreamGaussian) and text-to-3D (DreamCraft3D) baselines, our method produces significantly more plausible low-level geometry. For a fair comparison, we use the same input image for the first three rows.

Single-Texture Generation and Transfer. Our method generates coherent visual and geometric textures from both text and image prompts, demonstrating accurate albedo-normal alignment as shown in Figure 1. We further showcase in Figure 4 the flexibility by transferring different tactile textures (e.g., metal grid, rubber, canvas) onto the same object, enabling user-driven customization with real material priors.

Multi-Part Textures. In Figure 5, we assign distinct tactile textures to different semantic parts of an object (e.g., cactus and pot) using text prompts and diffusion-based segmentation. Our method produces consistent, segmented textures guided by the learned 3D label field.

Comparison with Baselines. We compare our method with DreamCraft3D [21], DreamGaussian [22], and Wonder3D [12]. As shown in Figure 6, our approach produces finer normal details and stronger color-geometry alignment, while DreamCraft3D exhibits overfitting artifacts and the "Janus" effect. In a user study with 1,000 responses on AMTurk, participants preferred our results in over 85% of cases for both texture appearance and geometric details.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[2] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathain, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning (ICML)*, 2015.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision (ECCV)*, 2020.

[4] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, 2023.

[5] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[6] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[7] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[8] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *International Conference on Learning Representations (ICLR)*, 2023.

[9] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[10] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[11] S. Wang, Y. She, B. Romero, and E. Adelson, "Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[12] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt *et al.*, "Wonder3d: Single image to 3d using cross-domain diffusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[13] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[14] S. Laine, J. Hellsten, T. Karras, Y. Seol, J. Lehtinen, and T. Aila, "Modular primitives for high-performance differentiable rendering," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, 2020.

[15] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *ACM SIGGRAPH*. Association for Computing Machinery, 2001.

[16] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[17] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[19] J. Tian, L. Aggarwal, A. Colaco, Z. Kira, and M. Gonzalez-Franco, "Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] S. Ge, T. Park, J.-Y. Zhu, and J.-B. Huang, "Expressive text-to-image generation with rich text," in *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[21] J. Sun, B. Zhang, R. Shao, L. Wang, W. Liu, Z. Xie, and Y. Liu, "Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior," in *International Conference on Learning Representations (ICLR)*, 2024.

[22] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," in *International Conference on Learning Representations (ICLR)*, 2024.