An Object is Worth 64x64 Pixels: Generating 3D Object via Image Diffusion

Xingguang Yan[1] **Han-Hung Lee**[1] **Ziyu Wan**[2] **Angel X. Chang**[1,3]
[1]Simon Fraser University [2]City University of Hong Kong [3]Canada-CIFAR AI Chair, Amii

omages.github.io

Abstract

This paper presents "Object Images," a novel method that represents 3D shapes with UV-patch structures as 64x64 pixel images, encapsulating geometry and appearance details. By converting complex 3D shapes into a 2D format, we leverage image generation models like Diffusion Transformers to produce realistic 3D models with UV-patches. This patch-based format naturally describes local regions of the generated object, making editing easier and enhancing the creative design process.

1 Introduction

Generative models for 3D assets face significant challenges due to geometric and semantic irregularities. Unlike 2D images, 3D shapes often involve complex topologies and intricate sub-structures, making them hard to model using standard approaches. Existing methods [6, 7, 5] often sacrifice geometric or semantic structures to the data pre-processing, which limits their effectiveness in applications requiring such structures.

To address these challenges simultaneously, we propose generating 3D shapes as Multi-Chart Geometry Images [4], leveraging UV charts in human-modeled 3D assets. This patch-based representation breaks shapes into smaller, semantically meaningful segments, allowing for easier, targeted editing and interaction, which is crucial for designing interactive, human-in-the-loop generative AI systems.

Our method transforms mesh geometry into a 12-channel image representation, termed "Object Images" (omages), which retains both material properties and semantic structures. By using image-based generative models, this technique supports creating textured 3D meshes with UV maps and materials, as demonstrated on the ABO dataset [1]. It is also relatively easy to edit afterwards.

2 Method

Our method introduces the Object Image (omage) representation, which encodes a 3D shape as a set of UV-mapped surface patches that are packed into a regular 2D image. As illustrated in Fig. 1, this representation preserves both geometric and material information, including albedo, normal, metalness, and roughness maps, enabling efficient reconstruction of photo-realistic 3D objects. By rasterizing the geometry and material data into a 12-channel image, we maintain the integrity of the original 3D structure and support interactive editing.

We utilize the Diffusion Transformer architecture [2] to model the distribution of these omages. This approach leverages transformers' ability to handle long-range dependencies and unordered sets, addressing the challenge of patch-based representation where the spatial arrangement in 2D does not affect the 3D shape. We first train a model to generate the four geometric channels, followed by a second stage to generate the remaining material channels, conditioning on the geometry.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

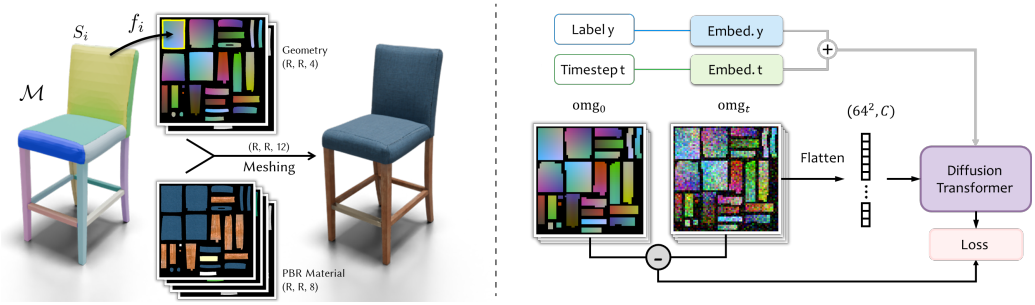

Figure 1: Method overview. **Left**: We assume the mesh $\mathcal{M}$ has patch decomposition $\{S_i\}$, and has single-valued uv-map f_i that flattens patch S_i into the 2D uv-domain. Together with the material maps, Object Images can represent high-quality photo-realistic object. **Right**: We train the image diffusion generative model with Diffusion Transformer. The input noised Object Image, omg, is first flattened into a sequence before passing into the transformer to predict the clean omg_0.

Figure 2: Examples of label-conditioned Omage-64 generation results. The left side displays results for 'ottoman', 'bed', 'exercise equipment', 'painting', 'lamp' 'vanity', 'plant pot', 'chair', 'pillow' and 'lamp'. Even at this resolution, thin structures are successfully generated. On the right, a scene with three objects generated by our method is shown, highlighting our capability in material generation.

To create omages from 3D assets with UV maps, we employ a UV-atlas repacking method that handles overlapping regions, out-of-boundary UVs, and excessive patches by merging vertices and retaining a maximum number of patches. High-resolution omages are then downsampled using boundary snapping and sparse pooling techniques, ensuring boundary information is preserved while reducing gaps between patches for efficient model training.

In Fig. 2 we show the label-conditioned generation results of our method trained on the ABO [1] dataset. After the DiT module generate geometry, we utilize Imagen [3] to generate material maps.

3 Conclusion

In this work, we introduce "Object Images," a multi-channel image representation that encodes potentially topology-complicated 3D objects with both semantic structures and appearance information, enabling a new approach to 3D generation that supports patch- and part-level edits. However, our current method has limitations, such as visible seams due to the inability to guarantee seamless patch fusion and constraints on resolution. In the future, we aim to generate higher-resolution omages that capture more detailed geometric and appearance details.

References

[1] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik. ABO: Dataset and benchmarks for real-world 3D object understanding. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, pages 21126–21136, 2022.

[2] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proc. Int. Conf. on Computer Vision*, 2023.

[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[4] P. V. Sander, Z. J. Wood, S. Gortler, J. Snyder, and H. Hoppe. Multi-chart geometry images. In *Eurographics Symposium on Geometry Processing*. Eurographics Association/Association for Computing Machinery, 2003.

[5] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan. InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

[6] X. Yan, L. Lin, N. J. Mitra, D. Lischinski, D. Cohen-Or, and H. Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022.

[7] B. Zhang, J. Tang, M. Niessner, and P. Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023.