Retrieve-and-generate: Image-to-3D framework with mesh retrieval and texture generation

SangEun Lee, Wonseok Chae, Hyeon-Jin Kim
Sudogwon Research Division
Electronics and Telecommunications Research Institute
Seongnam, Korea
sange1104@etri.re.kr, wschae@etri.re.kr, jini@etri.re.kr

***Abstract*—In recent years, significant advancements have been made in the field of image-to-3D reconstruction. However, one of the persistent challenges that remains unresolved is messy surfaces of generated outputs. To tackle this issue, we introduce a novel retrieve-and-generate scheme specifically designed for image-to-3D reconstruction tasks. Our method involves a two-stage process: first, the model retrieves the most similar 3D mesh corresponding to the input image; second, it generates a texture map that can be accurately mapped onto the retrieved mesh. In the initial retrieval phase, our approach leverages a pre-trained multi-modal joint representation to identify the 3D mesh that closely resembles the input image within the embedding space. Subsequently, texture generation module generates a realistic texture reflecting the input image, which leads to the complete 3D object reconstruction when this texture is mapped to the retrieved mesh. We have observed that our retrieve-and-generate approach significantly enhances the quality of the reconstructed 3D objects from a single input image. This improvement in reconstruction performance demonstrates the efficacy of our proposed method and its potential to advance the state-of-the-art in image-to-3D reconstruction technology.**

***Index Terms*—Image-to-3D reconstruction, 3D object generation, 3D shape retrieval, Texture synthesis**

I. Introduction

The task of image-to-3D reconstruction aims at generating a 3D object from a given 2D image. It holds significant potential in reducing the manual labor required to create 3D assets across various fields such as animation, graphics, and gaming. This process is considered to be challenging due to the need to predict and generate the parts of the object that are not visible in the input image, using only the limited information available.

With the advent and success of generative AI, recent studies struggling this task have gained a remarkable progress by leveraging the 3D ability inherited in the image generation models. Recent studies [1], [2] utilize different views of the object produced by image generation models and reconstruct 3D objects in an end-to-end manner based on Neural Radiance Fields (NeRF) systems.

Despite these advances, the visual quality of the generated 3D outputs still falls short of those created by skilled human designers. A primary issue lies in the uneven and often messy surfaces of the reconstructed models, which detracts from their overall realism and usability in real-world applications.

In this work, we introduce a novel retrieve-and-generate scheme for the image-to-3D reconstruction task. Prior works in 3D mesh retrieval have focused on effectively querying similar shapes from large 3D databases, by leveraging deep learning techniques to enhance the retrieval performance [3], [4]. On the other hand, texture generation methods have advanced in generating high-fidelity texture maps that accurately align with 3D geometry while simultaneously reflecting the given conditions [5], [6]. Inspired by these research backgrounds, our approach contrasts the current end-to-end paradigms by incorporating a mesh retrieval stage that queries the most similar 3D mesh corresponding to the input image, and a texture generation stage that generates a texture map that is designed to be mapped onto the retrieved mesh.

Our empirical results demonstrate that this retrieve-and-generate approach significantly enhances the quality of the reconstructed 3D objects compared to existing models. By effectively combining the strengths of mesh retrieval and texture generation, our method produces 3D objects with superior visual fidelity and surface smoothness, marking a substantial improvement over previous methods. This novel framework not only advances the technical capabilities in the field of image-to-3D reconstruction but also provides a practical solution for creating high-quality 3D assets with reduced human labor.

In short, our main contributions can be described as follows.

- To our best knowledge, it is the first time to introduce a new approach for image-to-3D reconstruction which integrates mesh retrieval and texture generation, contrasting the traditional end-to-end paradigms.
- The proposed method significantly improves the visual fidelity and surface smoothness of 3D objects compared to existing models, addressing the common issues of uneven and messy surfaces in the reconstructed models.
- While our method is challenging to generate a 3D object that is perfectly identical to the input image, it has the powerful advantage of producing high-quality objects at high speed, which can be applied directly to industries that require a large amount of 3D objects.

The remainder of the paper is organized as follows: Section II discusses recent studies, including 3D reconstruction, 3D shape recognition and retrieval, and texture generation. Section III explains our retrieve-and-generate scheme. Section IV introduces the dataset used in the experiment and compares

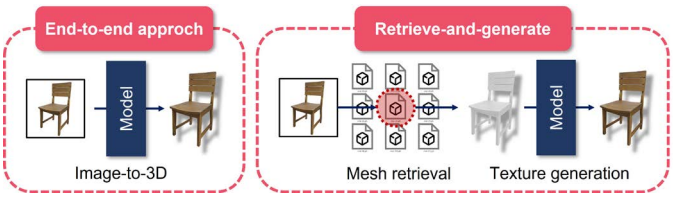

Fig. 1. Comparison between existing end-to-end approaches and our retrieve-and-generate paradigm.

the results from our method with previous studies. Finally, Section V concludes the article with an analysis of this study.

II. Related Work

A. 3D Reconstruction

To effectively learn how to reconstruct 3D object from a single image, a substantial amount of 3D data is required for model training, which is often costly to collect. To address this, recent approaches have utilized image generation models with high generative capabilities to guide 3D reconstruction. For instance, DreamFusion [7] employs the image generation model Imagen [8] to generate images of the desired object from various viewpoints based on text input, which a NeRF is used to reconstruct the 3D object. Similarly, Zero-1-to-3 [2] enhances the Stable Diffusion model by training it on different viewpoints, enabling more accurate viewpoint synthesis for 3D reconstruction from a single image. Magic123 [1] integrates Stable Diffusion as 2D prior information and Zero-1-to-3 [2] as 3D prior information to create a generative model that maintains both complex object generation capability and 3D consistency.

NeuralLift-360 [9] generates a 3D object from a single image with 360° views by utilizing a depth-aware neural radiance representation guided by denoising diffusion models and a CLIP-guided sampling strategy, enhanced with a ranking loss for rough depth estimation. RealFusion [10] reconstructs a 3D object by fitting a NeRF and utilizing a diffusion-based conditional image generator, combined through the DreamFusion method to integrate the input view, conditional priors, and regularizers.

Make-It-3D [11] optimizes a NeRF using constraints from the reference image and a 2D diffusion model for novel views, and converts the coarse model into textured point clouds and enhances realism with diffusion priors, leveraging high-quality textures from the reference image. Fantasia3D [12] produces a 3D object by disentangling geometry and appearance, using a hybrid scene representation for geometry learning with surface normals as input for an image diffusion model, and incorporating a spatially varying bidirectional reflectance distribution function (BRDF) for photorealistic rendering. Additionally, SV3D [13] utilizes a video generation model for multi-view synthesis, significantly improving the quality and 3D consistency of the generated objects.

On the other hand, numerous studies have explored Gaussian splatting, a method that represent the scene with 3D Gaussians, achieving fast, interpretable, and high-fidelity rendering [14]–[18].

B. 3D Shape Recognition and Retrieval

This task aims at accurately identifying corresponding 3D shapes from given 2D images or text descriptions. Lin et al. [3] proposes a method to retrieve 3D shapes through instance and category-level contrastive learning between single images and 3D shapes. Liu et al. [4] proposes a method that combines various large-scale 3D datasets and automatically filters and enriches text descriptions to perform multi-modal contrastive learning between 3D point clouds, images, and text.

C. Texture Generation

Recent studies have introduced texture synthesis task that maintains the structure of a given mesh while generating a texture map that incorporates specific conditions using generative models. Texture Fields [5] presents a generative model that predicts the colors of point clouds in 3D space to restore the texture, when given a 3D object model and a reference image. TEXTure [19] introduces a pre-trained depth-to-image diffusion model to generate textures on 3D shapes conditioned by text prompt, with a novel iterative scheme with trimap partitioning to create seamless textures from multiple viewpoints.

In addition, Point-UV Diffusion [20] combines Point Diffusion, which predicts the colors of some point clouds, with UV Diffusion, which improves the quality of texture images projected onto a 2D UV map. This approach enables the generation of high-quality textures for meshes of various geometric shapes. Furthermore, InTeX [6] is an interactive text-to-texture synthesis method with a unified depth-aware inpainting model to integrate depth information and inpainting cues, addressing 3D inconsistencies and enhancing generation speed.

III. Proposed Model

Our framework consists of a sequential process of two modules, where mesh retrieval stage searches the most similar object with the image, and then texture generation stage produces a texture image which can be mapped to the retrieved mesh, as shown in figure 2.

A. Mesh Retrieval

OpenShape [4] is multi-modal joint representations of text, images, and 3D point clouds, which enable both images and 3D objects to be mapped into a unified embedding space. The key idea is to train a point cloud encoder that aligns 3D shape embeddings with CLIP's text and image embedding spaces through multi-modal contrastive learning.

Specifically, we utilize a large-scale 3D shape database, ShapeNet [21], where each 3D mesh is converted into point clouds. These point clouds are then encoded by a point cloud encoder. To retrieve the most similar 3D object to an input image, we calculate the cosine similarities between the encoded image and all the encoded point clouds. The 3D mesh

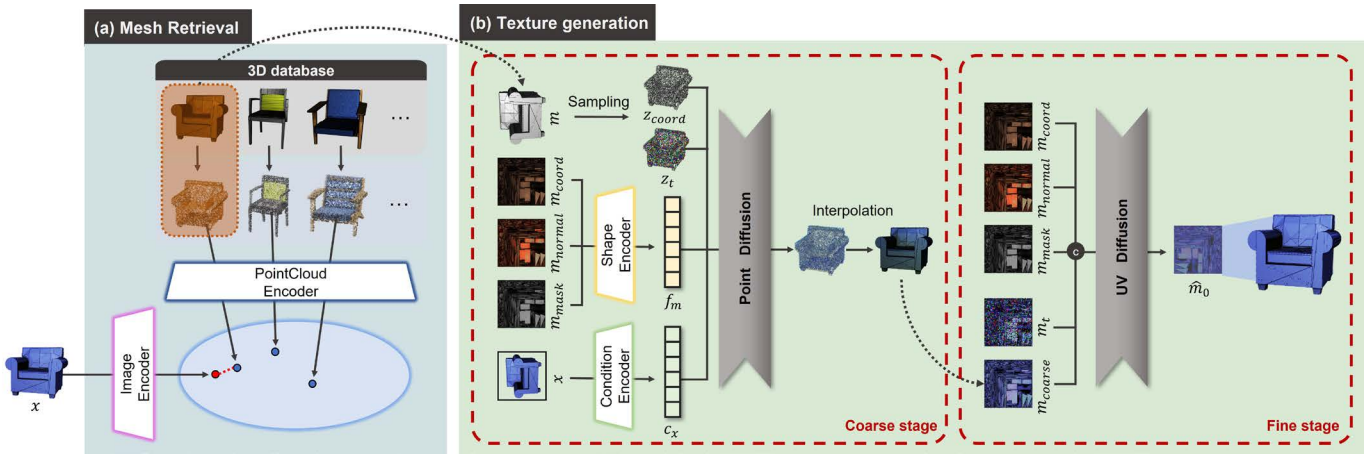

Fig. 2. Architecture of our proposed image-to-3D diffusion model with mesh retrieval and texture generation.

m with the highest similarity score is then selected for further processing. The detailed equation is presented as follows,

$$p* = \underset{i \in 0,1,\ldots,N}{\operatorname{argmax}} \frac{e_i(x) \cdot e_p(p_i)}{\|e_i(x)\| \cdot \|e_p(p_i)\|} \tag{1}$$

where x and p_i represent an input image and the i-th point cloud in the 3D database which consists of N point clouds, respectively. e_i and e_p refer to the image encoder and the point cloud encoder. The point cloud with the highest cosine similarity score is selected as the point cloud $p*$, which is then converted back to the mesh m, serving as one of the input variables for the next stage.

B. Texture Generation

In texture generation module, the texture image is generated which is able to be mapped to the retrieved mesh m as well as reflecting the style of input image. We adopted Point-UV diffusion [20], which is a coarse-to-fine texture generation model that produces a high-quality texture map conditioned by a text or an image. Point-UV diffusion operates in a manner where coarse stage generates a rough texture map, while fine stage enhances the fidelity of generated map.

In coarse stage, points are sampled from mesh m using farthest point sampling strategy. A set of coordinates and colors for sampled point clouds are defined as z_{coord} and z_0. After adding noise for t steps in diffusion forward process, noisy state is defined as z_t. Also, shape encoder converts m_{mask}, m_{normal}, m_{coord} into an embedding vector f_m, where m_{mask}, m_{normal}, m_{coord} represent the mask map, the normal map, and the UV coordinate map, respectively. In addition, an input image x is passed to the condition encoder, resulting an embedding vector c_x. Consequently, Point Diffusion model predicts the color of point clouds, given z_{coord}, z_t, f_m, and c_x, as follows:

$$\hat{z_0} = D^1_{\theta_1}([z_{coord}, z_t, f_m, c_x, t]) \tag{2}$$

where $D^1_{\theta_1}$ refers to the forward process of Point Diffusion model and $\hat{z_0}$ represents the color of the point cloud.

The remaining unsampled points are colored using the KNN interpolation method. The completed point cloud is then converted back into a mesh to extract the texture, thereby creating the coarse map m_{coarse}. This map m_{coarse} is subsequently passed to the fine stage for further processing.

Fine stage also improves the fidelity of the texture map by predicting the denoised state from a noisy input over several steps. With texture map m_{coarse}, noisy state m_t, shape information m_{mask}, m_{normal}, and m_{coord}, UV diffusion model generates the refined texture map $\hat{m_0}$, as follows:

$$\hat{m_0} = D^2_{\theta_2}([m_{coarse}, m_t, m_{mask}, m_{normal}, m_{coord}, t]) \tag{3}$$

where $D^2_{\theta_2}$ indicates UV diffusion model.

Finally, our framework reconstructs a complete 3D object with this texture map $\hat{m_0}$ and retrieved mesh m.

IV. Experiment

A. Dataset

ShapeNet [21] is one of the rich 3D object repository comprising approximately 52.5k 3D models that encompass a diverse range of categories and concepts. This dataset is regarded as a critical benchmark for training and evaluating 3D tasks. In this study, we experimented our method using categories of chair and table from the ShapeNet dataset, which include 6,777 and 8,511 objects, respectively. Figure 3 presents examples from the dataset used in our study. The first row showcases examples from the chair dataset, while the second row displays examples from the table dataset.

B. Result

Figure 4 qualitatively shows the reconstruction result of our method and baseline models including Zero-1-to-3 [2] and Magic123 [1]. Upon examination, it is evident that our model excels in producing detailed and 3D-consistent objects that accurately reflect the input image. In contrast, the other models tend to generate surfaces that are incomplete and uneven, highlighting the superiority of our approach in achieving a

Fig. 3. Examples of the dataset used in our experiment.

higher level of detail and consistency in the reconstructed 3D objects. This demonstrates the effectiveness of our method in overcoming common issues faced by existing models, thereby providing more reliable and visually appealing 3D reconstructions.

Fig. 4. Comparison of 3D object reconstruction. Ours (a) refers to the result of mesh retrieval module, where (b) indicates the final output of our model.

V. Conclusion

In this work, we proposed a novel retrieve-and-generate framework for image-to-3D task, which is a new scheme to overcome the low quality issue of mesh. Experimental results showed that our method can reconstruct 3D objects of high quality that reflects the input image as well as handle the current issue of low quality mesh. The key idea of our framework is that mesh retrieval module first queries the most similar mesh with the input image, and then texture generation module produces the visually matching texture to be mapped to the mesh output. We believe our research opens up new venues in which 3D assets are needed, such as animation, online games, and movies.

However, since our method relies on the most similar mesh with the input image, it cannot generate the exact same mesh with the image. Nevertheless, we expect that the applicability of our work to practical usage in 3D industry would be wider than previous works due to the higher reconstruction performance of our model. Several advances including expanding the available meshes or generating high-quality meshes would be addressed in the future work.

Acknowledgements

This work was supported by internal fund/grant of Electronics and Telecommunications Research Institute(ETRI). [24ZT1100, ICT convergence technology support and development based on local industry in the metropolitan area]

References

[1] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov *et al.*, "Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors," *arXiv preprint arXiv:2306.17843*, 2023.

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9298–9309.

[3] M.-X. Lin, J. Yang, H. Wang, Y.-K. Lai, R. Jia, B. Zhao, and L. Gao, "Single image 3d shape retrieval via cross-modal instance and category contrastive learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 405–11 415.

[4] M. Liu, R. Shi, K. Kuang, Y. Zhu, X. Li, S. Han, H. Cai, F. Porikli, and H. Su, "Openshape: scaling up 3d shape representation towards open-world understanding," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] M. Oechsle, L. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger, "Texture fields: Learning texture representations in function space," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4531–4540.

[6] J. Tang, R. Lu, X. Chen, X. Wen, G. Zeng, and Z. Liu, "Intex: Interactive text-to-texture synthesis via unified depth-aware inpainting," *arXiv preprint arXiv:2403.11878*, 2024.

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

[8] 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*, vol. 35, pp. 36 479–36 494, 2022.

[9] D. Xu, Y. Jiang, P. Wang, Z. Fan, Y. Wang, and Z. Wang, "Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4479–4489.

[10] L. Melas-Kyriazi, I. Laina, C. Rupprecht, and A. Vedaldi, "Realfusion: 360deg reconstruction of any object from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8446–8455.

[11] J. Tang, T. Wang, B. Zhang, T. Zhang, R. Yi, L. Ma, and D. Chen, "Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 22 819–22 829.

[12] R. Chen, Y. Chen, N. Jiao, and K. 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*, 2023, pp. 22 246–22 256.

[13] V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani, "Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion," *arXiv preprint arXiv:2403.12008*, 2024.

[14] S. S. Mallick, R. Goel, B. Kerbl, F. V. Carrasco, M. Steinberger, and F. De La Torre, "Taming 3dgs: High-quality radiance fields with limited resources," *arXiv preprint arXiv:2406.15643*, 2024.

[15] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[16] S. Niedermayr, J. Stumpfegger, and R. Westermann, "Compressed 3d gaussian splatting for accelerated novel view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 349–10 358.

[17] H. Yu, J. Julin, Z. Á. Milacski, K. Niinuma, and L. A. Jeni, "Cogs: Controllable gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 624–21 633.

[18] H. Ouyang, K. Heal, S. Lombardi, and T. Sun, "Text2immersion: Generative immersive scene with 3d gaussians," *arXiv preprint arXiv:2312.09242*, 2023.

[19] E. Richardson, G. Metzer, Y. Alaluf, R. Giryes, and D. Cohen-Or, "Texture: Text-guided texturing of 3d shapes," in *ACM SIGGRAPH 2023 conference proceedings*, 2023, pp. 1–11.

[20] X. Yu, P. Dai, W. Li, L. Ma, Z. Liu, and X. Qi, "Texture generation on 3d meshes with point-uv diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4206–4216.

[21] G. Li, B. Choi, J. Xu, S. S. Bhowmick, K.-P. Chun, and G. L.-H. Wong, "Shapenet: A shapelet-neural network approach for multivariate time series classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 9, 2021, pp. 8375–8383.