Journal of Multimedia Information System VOL. 11, NO. 4, December 2024 (pp. 241-248): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2024.11.4.241

Accurate Single Image to 3D Using View-Specific Neural Renderer

U-Chae Jun[1†], Jaeeun Ko[1†], Kibeom Hong[1*]

Abstract

Synthesizing a 3D model from a single 2D image is a significant challenge in computer vision and 3D modeling. Previous single image-to-3D methods generate multi-view images from a single image first and then feed these images to Neural Radiance Fields (NeRF) for 3D reconstruction. Therefore, visual consistency across viewpoints of these generated multi-view images directly affects the accuracy of 3D reconstruction. However, the previous methods tend to generate view-inconsistent images due to the projective ambiguity of a single image. To address the view inconsistency, we propose a viewpoint-specific learning method for single image-to-3D reconstruction using variants of NeRF. By introducing viewpoint-specific self-attention to NeRF, our method specializes the learning for viewpoints, enabling accurate 3D reconstruction even with visually discontinuous multi-view images. Experimental results demonstrate that the proposed method outperforms state-of-the-art single image-to-3D techniques by generating more accurate and coherent 3D models.

Key Words: Multi-View Generation Model, 3D Reconstruction, View Consistency.

I. INTRODUCTION

Humans can imagine the 3D shape of an object from a single camera view. To achieve a high level of generalization, humans rely on strong geometric prior knowledge, such as symmetry, which is built up through various visual explorations [1]. By emulating the powerful human ability to reason in 3D, recent single image-to-3D generation methods [2-6] are evolving to predict more creative and complex shapes by utilizing geometric information and prior knowledge gained from large and diverse datasets.

3D reconstruction from single image have been accomplished in two consecutive stages: 1) multi-view image generation, and 2) 3D reconstruction. Firstly, multi-view images are generated from a single image with a generative model. After that, variants of Neural Radiance Field (NeRF) [7-10] are used for 3D reconstruction using those generated multi-view images. Therefore, the quality of 3D reconstruction could be highly dependent on the visual continuity of the multi-view generated images. However, a single image inherently contains large projective ambiguity. As a result, multi-view images generated from a single image tend to have visual discontinuities. Fig. 1 illustrates representative examples of view-inconsistency in multi-view image generation from a single image.

Due to this property of the framework, previous works have explored multi-view consistency to improve 3D reconstruction quality in two main approaches. The first approaches [5,11] employs indirect loss functions, such as Score Distillation Sampling (SDS) and Contrastive Language-Image Pre-training (CLIP), to train a NeRF-based model. The indirect loss functions help to yield various and creative 3D results from visually discontinuous images, but this variety decreases the accuracy of 3D reconstruction. Additionally, a large-scale pretrained 2D image generative model is used to utilize indirect prior knowledge, which significantly increases processing times. The second approaches [2,6] focuses on improving the view consistency during multi-view image generation to use direct image loss functions, such as Mean Squared Error (MSE), in 3D reconstruction. Using direct image loss functions provides more accurate results and faster processing speeds under the assumption of high visual consistency between multi-view images. However, their performance heavily depends on the consistency of the generated multi-view images, leading to the lack of generalizability.

Recently, some efforts have been made to address these issues, including the use of 3D self-attention blocks for training [12] and the parameterization of camera information [3]. Nevertheless, view consistency problems still arise during the generation of multi-view images using a single image due to the projective ambiguity of a single-

Manuscript received November 21, 2024; Revised December 05, 2024; Accepted December 09, 2024. (ID No. JMIS-24M-11-034)

Corresponding Author (*): Kibeom Hong, +82-2-2077-7959, kb.hong@sookmyung.ac.kr

[†] Authors are equally contributed.

[1] Department of Computer Science, Sookmyung Women's University, Seoul, Korea, wjsdbco@sookmyung.ac.kr, rhwodms1223@sookmyung.ac.kr, kb.hong@sookmyung.ac.kr

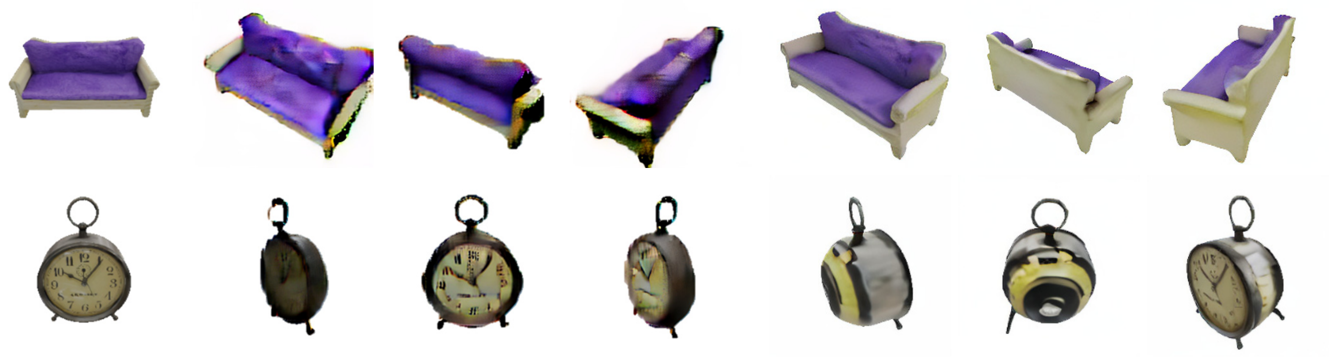

Fig. 1. An example of view inconsistent generation of previous methods. (a) results of a state-of-the-art single image-to-3D generation method [6], and (b) the proposed method.

view image.

To alleviate the multi-view inconsistent problems, we propose a novel 2D-to-3D method robust to view inconsistency by proposing a view-specific neural renderer for NeRF-based models under the assumption that the given multi-view images can be visually inconsistent. Especially, we introduce the novel self-attention layers within Multilayer Perceptron (MLP) in NeRF, which collaboratively use multi-view features to obtain a single accurate result. By incorporating the *view-specific* self-attention, the model independently specializes the learning for each viewpoint, effectively capturing the geometric and structural properties of objects. Our approach ensures accurate 3D reconstruction even in the presence of view discontinuities. We demonstrate the superiority of our proposed single image-to-3D method through extensive comparative experiments against previous state-of-the-art approaches. Notably, our method achieves significant improvements in accuracy even for visually inconsistent images.

II. RELATED WORK

2.1. Single Image-to-Multi-View Image Generation

Single image-to-multi-view image generation is a long-standing challenge in computer vision. Recently, image generation has been significantly advanced by the development of image diffusion models. Image diffusion models, such as stable diffusion [13], have demonstrated strong generalization abilities thanks to their training on large-scale datasets [14-16]. By iteratively denoising noise samples, diffusion models have been able to generate high-quality images and have become foundational for various generation tasks, including novel view synthesis. Single image to multi-view image generation methods have achieved substantial progress by leveraging diffusion models and can be categorized into 3D supervised learning methods and 2D lifting methods.

3D supervised learning methods [8-9] train a generator using 3D databases [14-16]. These methods can efficiently generate high-quality multi-view images. However, they are limited in generalization due to the limited scale of available 3D data. 2D lifting methods [11,17] utilize prior knowledge of pretrained diffusion models. Magic3D [18] utilizes a 2D diffusion model to optimize the 3D representation through SDS loss. Recently, SyncDreamer [6] introduces a synchronized multi-view diffusion model that ensures multi-view image consistency by processing each view simultaneously using a shared noise estimation and an attention mechanism.

Despite various efforts, however, the generation of multi-view images based on a single image is still prone to view discontinuities due to the projective ambiguity of a single-view image. Therefore, in this paper, we propose a novel method to address visual discontinuities in multi-view images by introducing a view-specific self-attention method for accurate 3D reconstruction.

2.2. 3D Reconstruction Using Multi-View Images

In recent years, neural network-based 3D reconstruction techniques, such as NeRF, have advanced significantly [7-10]. NeRF [19] is a coordinate-based neural scene representation method that combines neural networks with graphics principles to effectively represent 3D content. It takes the position and viewpoint direction of points in 3D space as input and predicts the color (RGB) and density of those points using an MLP network. Mip-NeRF [7] utilizes volumes of varying resolutions to estimate RGB color values and volume densities across multiple scales. This method minimizes information loss and enhances visual quality in complex scenes. PixelNeRF [8] synthesizes images from a new perspective using only 13 input images, compared to the 50-100 images required by previous NeRF-based models. Neural Surface Reconstruction (NeuS) [10] reconstructs 3D model based on multi-view images. While tradetional NeRF models use volume density fields to synthesize novel views, NeuS utilizes Signed Distance Field (SDF) to more accurately represent 3D surfaces.

These existing NeRF-based methods use MLP networks to predict volume density or SDF and color based on location inputs. In our work, we utilize the original NeRF model to validate the effectiveness of our proposed method for a fair comparison. The proposed method is applicable and can be implemented easily to variants of the NeRF model.

III. METHOD

Fig. 2 describes the overall architecture of our proposed single image-to-3D reconstruction based on view-specific self-attention layers. Inspired by variants of NeRF [7-10], our framework takes 3D locations and viewing directions as inputs along with the NeRF and produces RGB colors and density. Different from other related methods, our framework contains n specialized MLPs for the given n images generated from multiple viewpoints. This approach enables our model to be aware of information from other views during inference, ensuring view-consistent results.

3.1. View-Specific MLP Layer

We have followed the MLP structure of the conventional NeRF [19], which consists of 9 fully connected layers. It takes 3D location $x = (x, y, z)$ and viewing direction $\theta = (\theta, \phi)$ in spherical coordinate as input and outputs volume density σ and RGB color $c = (r, g, b)$. The MLP first uses the input 3D coordinates x with 8 fully connected layers to produce density σ and a feature vector, and the feature vector is then combined with the viewing direction θ through a 9th-fully connected layer to output the view-dependent RGB color c. Thus, similar to variants of NeRF [7-10], our MLP is defined as follows:

$$(c, \sigma) = MLP(x; \theta). \tag{1}$$

Our MLP network is composed of shared layers (blue block in Fig. 2) and view-specific layers (orange block in Fig. 2) for view-specific learning. For the given n views of images, the shared layers learn the common features across all n views. In contrast, view-specific layers are specialized for each view, where each layer learns view-specific features exclusively. Finally, these n view-specific features are adaptively incorporated via a self-attention layer.

In our experiments, we have trained the MLP architecture by 3 stages. First, a single MLP for all viewpoints. Then, n MLP networks with shared layers and n independent layers are trained for view-specific learning. Finally, we train n MLP layers by adaptively incorporating the view-specific features with the self-attention layer across the viewpoints.

3.2. Volume Rendering

After obtaining the RGB color and density values from the MLP, the process continues with volume rendering. For volume rendering, we have followed the previous related works [4-6,19]. Volume rendering determines pixel colors of a 2D image by accumulating radiance colors of 3D points along rays. These rays are cast from the camera through each pixel into the scene, and numerous points are sampled along these rays. Sampled 3D points x along a ray r pass through an MLP, which outputs a density σ and a RGB color c. These outputs are then alpha-composited from the back of ray to the camera, resulting in the final rendered RGB color for a pixel as follows:

$$\hat{C}(r) = \sum_i w_i c_i\,, \qquad w_i = \alpha \prod_{j<i} (1 - \alpha_j), \\ \alpha_i = 1 - \exp(-\sigma_i \| x_i - x_{i+1} \|). \tag{2}$$

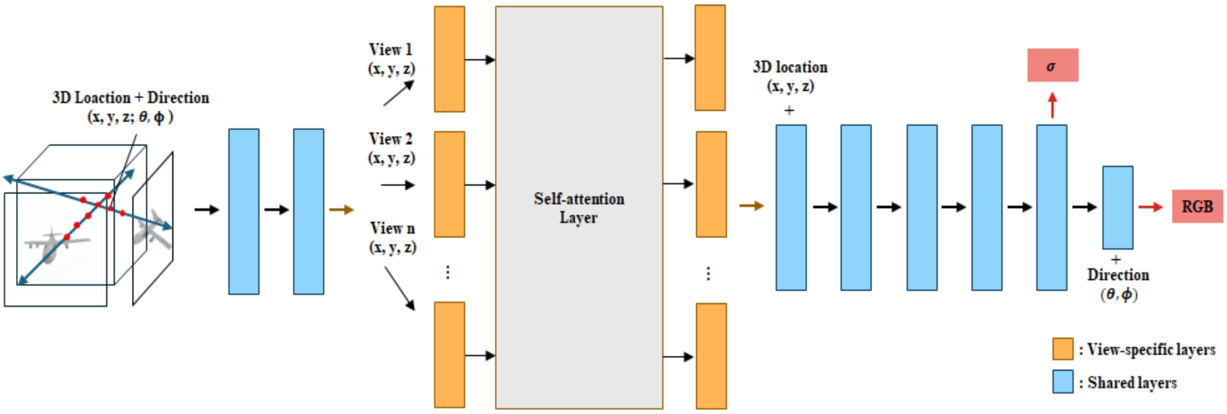

Fig. 2. Pipeline of proposed method. We introduce a self-attention layer in MLP of NeRF-based models to collaboratively incorporate multi-view features to obtain a single accurate result feature.

To ensure our model outputs desired ground truth (GT) colors and results, we train our model using the following loss function:

$$\mathcal{L} = \sum_{r \in \mathcal{R}} \left\| \hat{C}(r) - C(r) \right\|_2^2 , \tag{3}$$

where R is the set of camera rays r, $\hat{C}(r)$ is the color obtained by our model from camera ray r, and $C(r)$ is the GT color from r.

IV. EXPERIMENTS

To validate the performance of the proposed method, we compare the performance with recent single image-to-3D generation models, *i.e.*, Zero-1-to-3 [4], RealFusion [5] and SyncDreamer [6]. Zero-1-to-3 [4] utilizes a diffusion model to learn how to control the camera extrinsic. Zero-1-to-3 takes a single image and the relative pose of the camera as inputs to generate a novel view image corresponding to the given pose. The generated multi-view images are then reconstructed in 3D using the Score Jacobian Chaining (SJC) loss function. RealFusion [5] is a single-image-based 3D reconstruction that leverages existing 2D diffusion models to learn 3D geometry and appearance by sampling images from different viewpoints and minimizing rendering loss.

In the experiment, we have calculated three metrics to evaluate the model's performance in image generation from a novel viewpoint following the recent related works [6,20]: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). PSNR measures the maximum possible pixel value and the MSE between the two images, with higher values indicating less distortion and better quality [21]. The Structural Similarity Index Measure (SSIM) assesses the similarity between images based on luminance, contrast, and structural consistency. A higher SSIM value indicates better structural consistency between images [21]. Learned Perceptual Image Patch Similarity (LPIPS) is a metric that leverages deep learning to measure perceptual differences between images, making it useful for evaluating the quality of complex images. A lower LPIPS value signifies greater similarity between the generated image and the real image, indicating better quality [22].

For evaluation, we have utilized 1,030 3D data points from the GSO dataset [14] that were not used in the training of the comparison models, including our model. The GSO dataset primarily consists of high-quality 3D-scanned household items, which are diverse in terms of shapes, textures, and sizes, making it suitable for evaluating object-level 3D reconstruction. Therefore, the GSO dataset allows us to measure the generalization performance on new inputs that the model has not encountered during training.

Table 1 presents the quantitative comparisons with previous methods. The results demonstrate that our model significantly outperforms the previous methods across the evaluation metrics. Specifically, the proposed method demonstrates higher consistency of generated 3D models and the ability to generate high-quality images from novel viewpoints. However, our model increases the overall complexity due to the computation of self-attention for each view in the view-specific layer, leading to a longer generation time compared to other model [6]. In particular, our model takes far less time than Zero 1-to-3 [4] and RealFusion [4] and needs a little longer time but comparable to SyncDreamer [6]. Even with the little higher computational cost, our model outperformed the other methods, showing consistently better multi-view coherence and 3D reconstruction quality than the other methods on all evaluation metrics.

In addition, Fig. 3 illustrates the results of a single-image-to-3D reconstruction. These results were generated from the multi-view images shown in Fig. 4, which consist of four views rendered at 0°, 90°, 180°, and 270° viewpoints. These multi-view images were generated using SyncDreamer [6], and we conducted experiments on both simple and complex objects to demonstrate the model's generalization ability. For each object in generating multi-view images, we render an input view image with a size of 256×256. RealFusion [5] shows an acceptable multi-view consistency for simple objects but fails to produce visually plausible images for complex objects. Zero-1-to-3 [4] produces visually plausible geometry but produces inaccurate 3D reconstructions that are overly smoothed. The results show that the indirect loss, such as SDS and SJC, is limited to efficiently derive visual continuity from inconsistent images. In contrast, our proposed method produces more accurate 3D reconstructions, by efficiently addressing visual discontinuities across viewpoints using view-specific learning. In addition, our method can produce images that are semantically consistent with the input image and multi-view consistent in color and geometry.

Table 1. Quantitative comparisons with previous methods.

Method/ metric	PSNR (↑)	SSIM (↑)	LPIPS (↓)	Time (min)
RealFusion [5]	15.26	0.722	0.283	90
Zero 1-to-3 [4]	18.93	0.779	0.166	40
SyncDreamer [6]	20.02	0.783	0.159	6
Ours	**20.32**	**0.798**	**0.153**	**7**

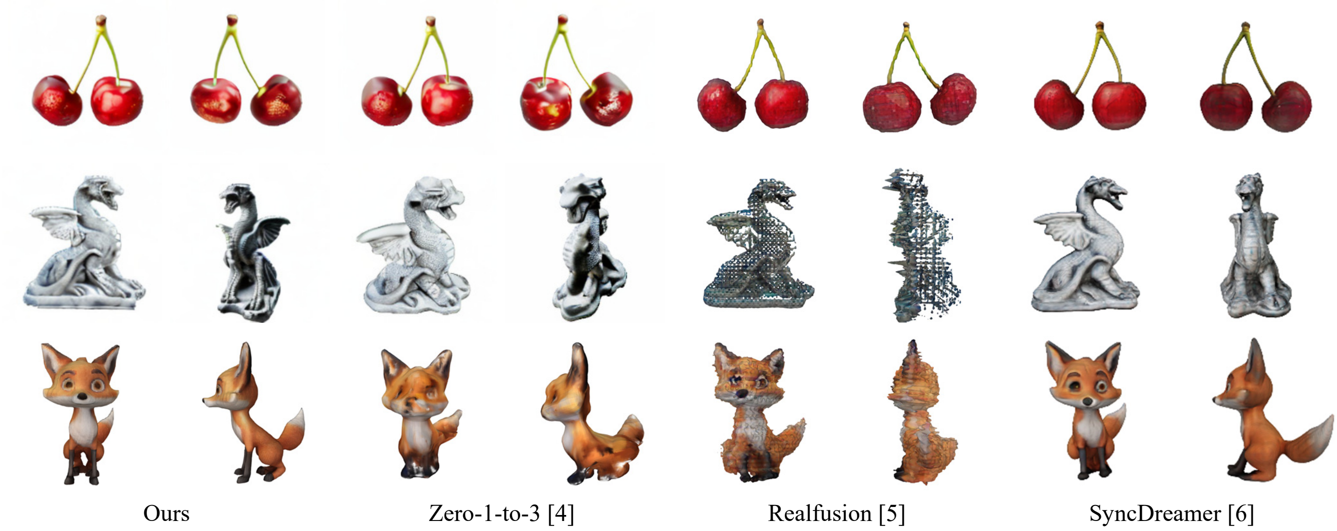

Fig. 3. Qualitative comparisons with state-of-art single image-to-3D techniques: Zero1-to-3 [4], RealFusion [5], and SyncDreamer [6].

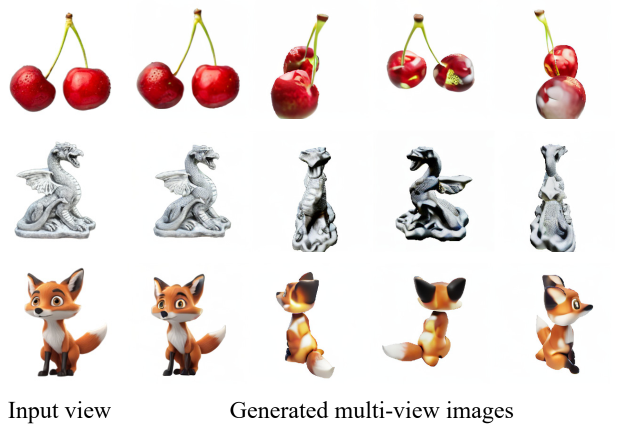

Fig. 4. Multi-view images from single input view using state-of-the-art single image-to-multi-view generation method [6].

V. CONCLUSION

In this paper, we propose a novel method to effectively address visual discontinuities in multi-view images for single-image-based 3D reconstruction. Previous single image-to-3D methods generate multi-view images from a single image first and then feed these images to NeRF for 3D reconstruction. Therefore, visual continuity across viewpoints of these generated multi-view images directly affects the accuracy of 3D reconstruction. However, current multi-view generation models from a single image face the issue of view incoherence, due to the lack of geometric information. To this end, we introduce a novel 2D-to-3D framework robust to view discontinuity by proposing a view-specific neural renderer for NeRF-based models. We propose a *view-specific* self-attention network within the MLP in NeRF, which collaboratively incorporates multi-view features to obtain a single accurate result. Furthermore, by training view-specific NeRF using direct image loss functions, we can achieve more accurate 3D reconstruction.

Through experiments, the proposed method outperforms previous state-of-the-art single image-to-3D methods, demonstrating the improved 3D reconstruction accuracy of our method. Additionally, the proposed method can be easily implemented in various NeRF model variants.

However, despite its strong performance, the proposed method has limitations in handling diverse real-world datasets that may contain significant noise or highly variable input conditions. To address this challenge, preprocessing steps such as noise reduction techniques or outlier detection could improve the robustness of our framework. We hope that the proposed method provides broader applicability in fields such as computer vision and image processing.

ACKNOWLEDGMENT

This research was supported by the funding provided by the Seoul Metropolitan Government in 2024 under the "2024 AI Technology Commercialization Support Program" (Grant No. SY240213). The authors also sincerely express their gratitude to Professor Jiwoo Kang of Sookmyung Women's University for her valuable guidance and support during the preparation of this paper.

REFERENCES

[1] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, no. 2, pp. 115-147, 1987.

[2] X. Long, Y. C. Guo, C. Lin, Y. Liu, Z. Dou, and L. Liu, et al., "Wonder3D: Single image to 3D using cross-domain diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,

2024, pp. 9970-9980.

[3] V. Voleti, C. H. Yao, M. Boss, A. Letts, D. Pankratz, and D. Tochilkin, et al., "Sv3d: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion," in *Proceedings of the European Conference on Computer Vision*, Cham: Springer, 2025, pp. 439-457.

[4] 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.

[5] L. Melas-Kyriazi, I. Laina, C. Rupprecht, and A. Vedaldi, "RealFusion: 360° 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.

[6] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, and T. Komura, et al., "SyncDreamer: Generating multiview-consistent images from a single-view image," in *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[7] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 2021, pp. 5855-5864.

[8] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelNeRF: Neural radiance fields from one or few images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4578-4587.

[9] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, and P. P. Srinivasan, "Ref-NeRF: Structured view-dependent appearance for neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5481-5490.

[10] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021, pp. 27171-27183.

[11] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D Diffusion," in *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[12] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang, "MVDream: Multi-view diffusion for 3D generation," in *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684-10695.

[14] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, and E. VanderBilt, et al., "Objaverse: A universe of annotated 3D objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13142-13153.

[15] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs, "Large-scale multi-view stereopsis evaluation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 406-413.

[16] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, "Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10901-10911.

[17] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, and H. Su, et al., "ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 8406-8441.

[18] C. H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, and X. Huang, et al., "Magic3D: High-resolution text-to-3D content creation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 300-309.

[19] 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," *Communications of the ACM*, vol. 65, no. 1, pp. 99-106, 2021.

[20] 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. 1-25, 2023.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586-595.

AUTHORS

U-Chae Jun is currently pursuing her B.S. degree in the Department of Software at Sookmyung Women's University, Korea, since 2020. Her research interests include Computer Vision, Computer Graphics, and Artificial Intelligence.

Jae-Eun Ko is currently pursuing her B.S. degree in the Department of Software at Sookmyung Women's University, Korea, since 2021. Her research interests include Computer Vision, Computer Graphics, and Artificial Intelligence.

Kibeom Hong is currently an assistant professor in the Department of Software at Sookmyung Women’s University, Korea, since 2024. He received B.S. and the Ph.D. degrees in computer science from Yonsei University, Seoul, Korea in 2023. His research interests include generative models, neural style transfer and domain generalization.