COSE474-2024F: Final Project Proposal
"Text-guided One Image to 3D Generation"

Seohyeon Park [1]

1. Introduction

In recent years, the field of computer vision has seen significant advancements in image-to-3D model generation. These models enable a wide range of applications in virtual reality (VR), gaming, and industrial design. However, a common limitation of existing models is their requirement to be devoid of complex backgrounds for input images. To address this challenge, we propose to develop a text-guided single image to 3D model.
Our framework integrates the CLIP model for textual understanding to guide the segmentation of a single 2D image, followed by constructing multi-view representations using Zero123. These representations are rapidly transformed into 3D models through Instant NGP.

2. Problem definition & challenges

The primary problem of existing image-to-3D generation models is the difficulty of processing images containing complex backgrounds. Traditional image-to-3D models struggle with accurately segmenting the object of interest from these backgrounds, leading to erroneous or incomplete 3D representations.
One of the challenges is the accurate identification of objects, guided by textual descriptions. The system must interpret textual descriptions and correlate them with visual data accurately.
Another challenge is the speed of the image-to-3D conversion process. For applications such as virtual reality, the conversion process needs to be not only accurate but also exceedingly fast. This demands optimization to ensure operation in real-time without sacrificing the quality.

3. Related Works

CLIP (Radford et al., 2021) leverages a contrastive learning framework to align images and text in a shared latent space, enabling robust zero-shot transfer to downstream tasks by learning generalized visual and textual representations.
Furthermore, Zero123 (Liu et al., 2023b) is a neural network designed for zero-shot 3D shape generation from single-view images. It leverages 2D image priors and novel view synthesis techniques to produce coherent 3D shapes.
Moreover, Instant NGP (Müller et al., 2022) is a rapid neural scene representation using a multi-resolution hash grid encoding, resulting in real-time reconstruction of 3D scenes from 2D images with high fidelity.

4. Datasets

Replica dataset (Straub et al., 2019) will be used for the training. It is a high-quality 3D dataset designed for photorealistic indoor scene reconstruction, featuring precise geometric and semantic annotations.
For outdoor scenes, Mip-NeRF 360 dataset (Barron et al., 2022), which provides a collection of 360-degree scenes used for training neural radiance fields (NeRF) models, will be used.

5. State-of-the-art methods and baselines

RealFusion (Melas-Kyriazi et al., 2023) uses a neural radiance field (NeRF) for the 3D geometry and diffusion-based image generation models to generate novel views of the object. They combine the DreamFusion method with regularizers.
In addition, One-2-3-45 (Liu et al., 2023a) leverages a 2D diffusion model called Zero123 to generate multi-view images, which are then reconstructed into a 3D model using a generalizable SDF-based neural surface method. Unlike traditional optimization-heavy approaches, this method is faster and more consistent, showing superior performance in both synthetic and real-world scenarios.
Lately, Isotropic3D (Liu et al., 2024) introduces an approach by utilizing a single CLIP embedding without L2 supervision loss, employing Explicit Multi-view Attention (EMA) to generate more geometrically accurate and consistent 3D models across different views.

6. Schedule

Week 1: Read SOTA methods and set up the environment
Week 2: Preprocess datasets
Week 3: Train network architecture
Week 4: Experiment with datasets
Week 5: Write the final project paper

References

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., and Hedman, P. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022.

Liu, M., Xu, C., Jin, H., Chen, L., T, M. V., Xu, Z., and Su, H. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023a. URL https://arxiv.org/abs/2306.16928.

Liu, P., Wang, Y., Sun, F., Li, J., Xiao, H., Xue, H., and Wang, X. Isotropic3d: Image-to-3d generation based on a single clip embedding, 2024. URL https://arxiv.org/abs/2403.10395.

Liu, R., Wu, R., Hoorick, B. V., Tokmakov, P., Zakharov, S., and Vondrick, C. Zero-1-to-3: Zero-shot one image to 3d object, 2023b. URL https://arxiv.org/abs/2303.11328.

Melas-Kyriazi, L., Rupprecht, C., Laina, I., and Vedaldi, A. Realfusion: 360° reconstruction of any object from a single image, 2023. URL https://arxiv.org/abs/2302.10663.

Müller, T., Evans, A., Schied, C., and Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H. M., Nardi, R. D., Goesele, M., Lovegrove, S., and Newcombe, R. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.