COSE474-2024F: Final Project
"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 overcome this, we propose a text-guided approach that transforms a single image into a 3D model using both image and textual inputs to specify the object for segmentation.

Our method leverages the Contrastive Language-Image Pre-training model (CLIP) and the Segment Anything Model (SAM) to achieve precise text-driven segmentation of 2D images. Once segmented, these images are then processed by DreamGaussian to generate the 3D models. This integration allows for more accurate and efficient 3D model generation from single images with complex backgrounds.

2. Related Works

2.1. Semantic Segmentation

The Segment Anything Model(SAM) (Kirillov et al., 2023) is an innovative approach in the field of image segmentation, particularly notable for its ability to effectively handle diverse segmentation tasks across a broad range of categories.

However, SAM exhibits limitations when tasked with segmenting images based on text prompts. This challenge stems from the model's architecture, which does not integrate the nuances and specificities of text-based instructions into its segmentation process, while highly adept at interpreting visual content. Instead, SAM relies on box or point prompts to guide its segmentation. This limitation highlights a critical area for further development in integrating linguistic and visual understanding in segmentation models.

To overcome this, Grounded SAM (Ren et al., 2024) employs a annotated box derived from the grounding DINO object detection (Liu et al., 2024b) and performs segmentation by initially specifying the area characterized by text prompts. This approach significantly reduces computational costs and improves segmentation performance.

However, Grounded SAM is constrained by its reactivity to specific vocabularies only, which can make comprehensive semantic segmentation challenging. In this paper, to address this issue, the CLIP model (Radford et al., 2021) has been used to improve text-based segmentation performance. CLIP uses a contrastive learning framework to effectively align images and text within a shared latent space. This enables robust zero-shot transfer capabilities to downstream tasks by learning generalized visual and textual representations, thus facilitating a more adaptive and flexible approach to semantic segmentation across diverse textual prompts.

2.2. 3D Generation Guided by Single Image

One Image to 3D generation is a technique that reconstructs three-dimensional models from a single image, often using 2D diffusion models to optimize the 3D model creation process. These methods leverage various innovative approaches to enhance the fidelity and efficiency of the generated models.

RealFusion (Melas-Kyriazi et al., 2023) uses diffusion-based image generation to create novel views of an object, which are then utilized to construct its 3D geometry. Zero123 (Liu et al., 2023b) is a neural network, specifically designed for zero-shot 3D shape generation from a single-view image. Isotropic3D (Liu et al., 2024a) employs a single CLIP embedding and Explicit Multi-view Attention (EMA) to produce more geometrically accurate and consistent 3D models across different views. These techniques predominantly use multi-view images to produce the 3D geometry through a neural radiance field (NeRF) (Mildenhall et al., 2020).

Additionally, One-2-3-45 (Liu et al., 2023a) utilizes a network based on Zero123 to generate multi-view images and 3D convolution for creating SDF volumes, noted for its speed and consistency. DreamGaussian (Tang et al., 2024), also based on Zero123, utilizes 3D Gaussian Splatting (Kerbl et al., 2023) for fast rendering and optimization, surpassing traditional neural network field methods in efficiency. In this paper, DreamGaussian is specifically employed to generate 3D models from segmented 2D images, demonstrating significant advancements in the field.

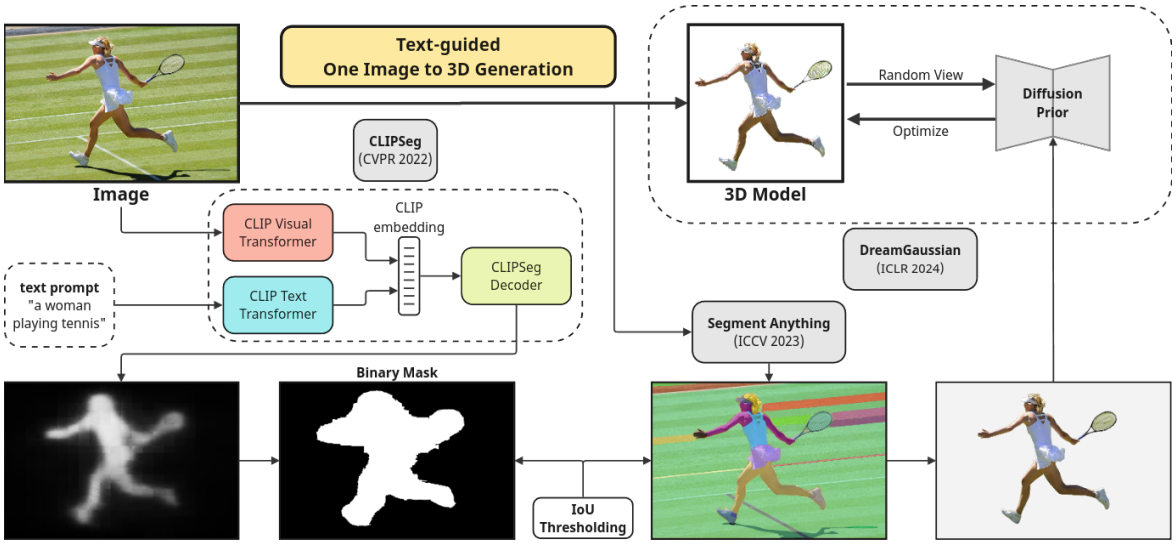


Figure 1. Overview of Text-guided One Image to 3D Generation

3. Methods

3.1. Overview

We suggest a technique for converting a single image into 3D models by semantically understanding and manipulating specific regions of images. This method utilizes the capabilities of the CLIP and SAM models. Our goal is to selectively process parts of the image that are semantically significant based on a specific textual description.

3.2. Text-guided Segmentation

Firstly, we leverage the CLIP model's capabilities in a fine-tuned version specifically designed for image segmentation, known as CLIPSeg. This model has been pretrained to understand complex semantic relationships between text and images. The result of CLIPSeg is processed to produce a binary segmentation mask that highlights regions of the image corresponding to the text query.

In parallel, we use the SAM model, which generates multiple segmentation masks based on the content of the image without textual context. The SAM model is designed to automatically identify and segment various objects within an image, making it highly effective for detailed and diverse image analyses.

The synergy between CLIPSeg and SAM is crucial. After SAM identifies potential objects within the image, the binary mask produced by CLIPSeg refines these predictions by comparing them against the segmentation driven by the textual description. This combination allows for high precision in isolating and enhancing areas of the image that are most relevant to the provided text.

To integrate the segmentation results from the SAM and CLIPSeg models, we compute the Intersection over Union (IoU) between the masks generated by each technique. Given that the binary mask from CLIPSeg covers a larger area than the SAM-segmented mask, the union in our IoU calculation is constrained by the SAM mask's boundaries. If the IoU exceeds a predefined threshold of 0.5, it signifies substantial overlap and agreement between the models on the identified region. When this condition is met, we combine the masks to create a more precise and detailed segmentation of the image.

3.3. Image to 3D

In the final step, we utilize Dreamgaussian, an advanced technique for converting segmented 2D images into detailed 3D models. It employs generative Gaussian splatting along with Zero-1-to-3 XL, a 2D diffusion prior, which facilitates the transformation process. This method involves the progressive densification of 3D Gaussians to simplify the optimization, enabling efficient initialization of both geometry and appearance attributes.

4. Experiments

4.1. Dataset and Setting

For our evaluation, we use the PhraseCut (Wu et al., 2020) for quantitative analysis and the MS COCO 2017 (Lin et al., 2014) for qualitative analysis. All experiments are performed and measured on an Ubuntu 22.04 system, equipped with an AMD Ryzen 7 5800X 8-core processor and an NVIDIA RTX 4090 GPU.

	Ours	Grounded SAM
IoU ↑	0.303	**0.394**
IoU (based on Ground Truth) ↑	**0.694**	0.610
time (s) ↓	3.560	**0.383**

Table 1. **Quantitative Comparison.** Semantic segmentation performance of our method and Grounded SAM on PhraseCut dataset.

4.2. Quantitative Results

4.2.1. Text-guided Segmentation

The segmentation task was performed on 100 images from the PhraseCut dataset, responding to a total of 821 text prompts (phrases). The Intersection over Union (IoU) metric was the primary focus of our evaluation.

Although our IoU scores were lower compared to the state-of-the-art method Grounded SAM, it's important to note that our method is not designed for instance segmentation. Our approach outperforms Grounded SAM when assessing IoU against ground truth, despite requiring more processing time due to the use of CLIPSeg, which demands significant computational resources.

A significant drawback of Grounded SAM is its dependency on outputs from Grounding DINO; without these outputs, semantic segmentation followed by SAM is infeasible. This limitation underscores the robustness and self-sufficiency of our approach in semantic segmentation scenarios.

Additionally, the reliability of the IoU metric is affected by inaccuracies in ground truth annotations, such as with the "busy sidewalk" image, where the annotations do not fully represent the scene's semantic details, casting doubt on the IoU's definitive value as a performance metric.

4.3. Qualitative Results

4.3.1. Text-guided Segmentation

Our method demonstrated commendable performance in semantic segmentation tasks, showing significant alignment with ground truth annotations across diverse scenarios. These results substantiate our model's ability to effectively interpret and segment complex scenes based on textual prompts, reflecting a robust understanding of various semantic elements within the images.

4.3.2. Image to 3D

While our method effectively produced high-quality 3D models from known viewpoints, it has yet to resolve the Janus problem in unseen views. Despite this, the 3D models viewed from familiar perspectives displayed detailed accuracy and closely matched the original images.

(a) original image (b) GT mask (c) our mask

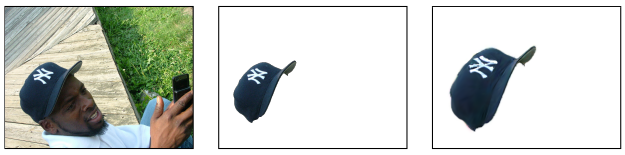
Figure 2. prompt **"blue sky"**

(a) original image (b) GT mask (c) our mask

Figure 3. prompt **"busy sidewalk"**

(a) original image (b) our mask (c) 3D model

Figure 4. 3D model from image guided by prompt **"a teddybear"**



(a) original image (b) our mask (c) 3D model

Figure 5. 3D model from image guided by prompt **"a black cap"**

5. Future Direction

Our research will focus on enhancing model performance through hyperparameter optimization, specifically targeting variables related to Intersection over Union (IoU) and the masks generated by CLIPSeg. This fine-tuning aims to improve segmentation accuracy considerably.

Additionally, we plan to extend the training of pretrained models, like SAM and CLIPSeg, using an expanded dataset. This approach will adapt their functionalities to meet our specific requirements more effectively. Increasing the dataset size will also improve the models' ability to generalize across various scenarios.

Finally, we aim to train the Dreamgaussian model on real-world data to effectively tackle the Janus problem. Achieving this will allow us to produce more realistic 3D models that handle varying viewpoints, enhancing the utility of our research in practical applications.

References

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. 3d gaussian splatting for real-time radiance field rendering, 2023. URL https://arxiv.org/abs/2308.04079.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., and Zitnick, C. L. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

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, 2024a. 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.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024b. URL https://arxiv.org/abs/2303.05499.

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.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *CoRR*, abs/2003.08934, 2020. URL https://arxiv.org/abs/2003.08934.

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.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., and Zhang, L. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. URL https://arxiv.org/abs/2401.14159.

Tang, J., Ren, J., Zhou, H., Liu, Z., and Zeng, G. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation, 2024. URL https://arxiv.org/abs/2309.16653.

Wu, C., Lin, Z., Cohen, S., Bui, T., and Maji, S. Phrasecut: Language-based image segmentation in the wild, 2020. URL https://arxiv.org/abs/2008.01187.