2D guided 3D gaussian-based binary segmentation

Yang Yang
Applied Computing.
Lincoln University
Christchurch, NZ
Yang.Yang6@lincolnuni.ac.nz

Green Richard
Computer Science.
Canterbury Univeristy
Christchurch, NZ
richard.green@canterbury.ac.nz

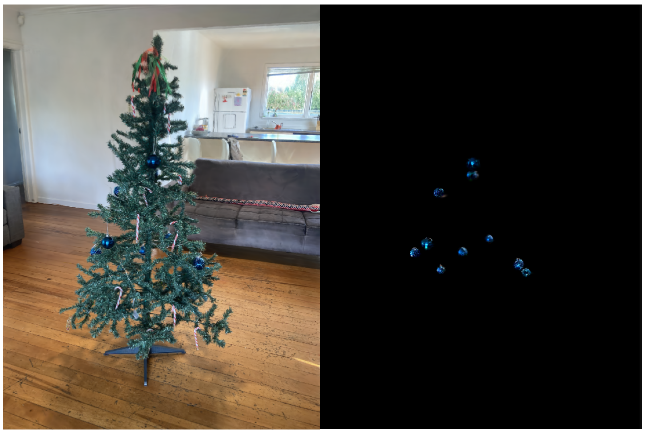


Fig. 1. Our segmentation method effectively utilizes Gaussian modeling to discern and delineate objects within a three-dimensional space. The image showcases the effective isolation of ornaments from the Christmas tree.

***Abstract*—This paper proposes a method to prompted binary segmentation of 3D point clouds by leveraging a combination of advanced techniques, including Grounding DINO for 2D object detection, SAM for segmentation, and the state-of-the-art 3D scene representation method, 3D Gaussian splatting. Our method decouples the training of the 3D Gaussian model from the object segmentation process, enhancing the flexibility and efficiency. To facilitate the discrimination between target and non-target points, we augment each trained 3D Gaussian with an additional attribute representing the likelihood of a point belonging to the target. By rendering the 3D Gaussian model solely based on the new likelihood attribute, rather than the traditional color attribute, we generate a specialized possibility map. This map is then compared against the multi-view binary masks of target objects predicted by Grounding DINO and SAM. The proposed learning process yields comprehensive insights into the likelihood of each point belonging to the target objects. The results show that the proposed approach can achieve accurate segmentation of target object points with the trained attribute $p > 0.9$.**

Keywords—3D gaussian, grounding DINO, SAM, 3D segmentation

I. Introduction

In the intricate domain of 3D object detection and segmentation, accurate and efficient methodologies are pivotal, particularly in sectors like agriculture where they directly influence decision-making and resource optimization. Traditional methods, primarily centered around 2D imagery, have provided substantial groundwork but often fail to navigate the complexities of 3D environments. This is especially evident in applications such as fruit counting in agriculture, where accurate yield estimation is crucial for efficient yield forecasting and resource allocation. Recent technological advancements have attempted to bridge this gap, transitioning from 2D to 3D object detection to overcome the inherent limitations of 2D-based methodologies. However, while notable strides have been made, challenges persist, particularly when dealing with vast spatial expanses and the intricate details of numerous objects, such as fruits on a tree. This paper aims to address these challenges by introducing an approach that integrates state-of-the-art techniques from both 2D object detection and 3D scene reconstruction. Our method harnesses the power of Grounding DINO for refined 2D object detection, SAM for meticulous segmentation, and 3D Gaussian splatting for advanced scene representation. Our proposed framework separates the training of the 3D Gaussian model from the segmentation process, which introduces a modular aspect to the workflow. Additionally, it assigns a probability value to each data point, allowing for more streamlined and adaptable processing pipelines. Moreover, the specialized focus on binary segmentation allows for a tailored approach in distinguishing between the object of interest and the background, thereby improving the system's efficiency in scenarios where a clear demarcation is essential.

II. Background

A. 2D-based object segmentation

The realm of fruit counting has seen a proliferation of 2D image detection techniques, bolstered by deep learning advancements [3, 6, 9, 10, 18]. Methods like Faster R-CNN have been pivotal for such tasks. The state-of-the-art Segment Anything Model (SAM) [16] excels particularly in image segmentation for the ability to segment images effectively without task-specific training samples. While these approaches have showcased significant advancements in automating fruit counting processes, they inherently operate within the confines of two-dimensional visual data and encounter limitations, especially when scaling to the 3D complexities of extensive agricultural spaces. Recent advancements, such as the introduction of the YOLOv7 framework, with its attention mechanism, has made strides in addressing some of these challenges by improving multi-object detection in video sequences [7]. Nonetheless, for comprehensive and accurate counting across vast orchards, these methods need to evolve into the 3D domain, where they

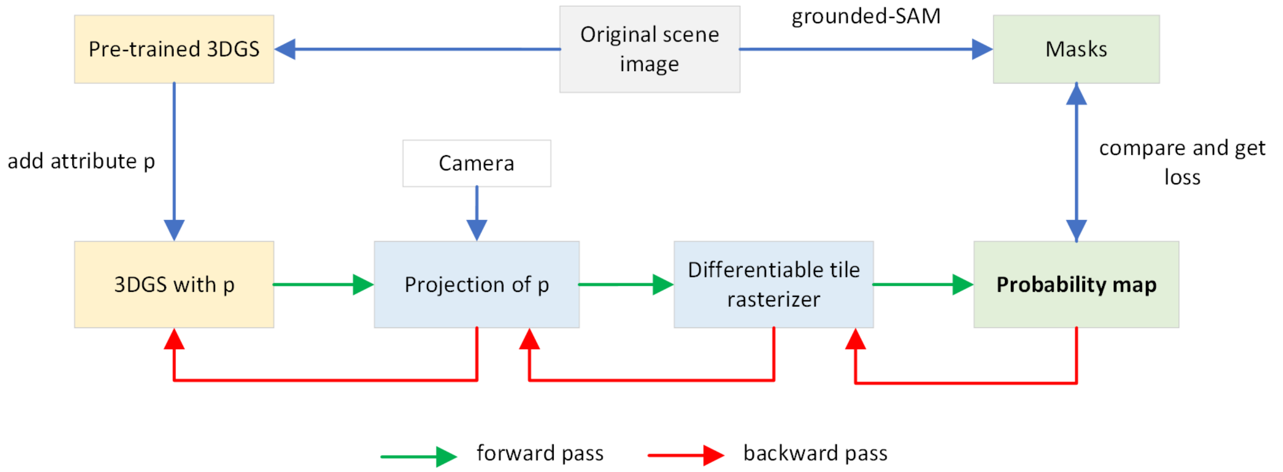


Fig. 2. The flowchart of the segmentation process using a 3D Gaussian Splatting (3DGS) model enhanced with an added attribute $\boldsymbol{p}$ for probability. The process begins with a pre-trained 3DGS to which the attribute $\boldsymbol{p}$ is added. This modified 3DGS is then used to project the attribute $\boldsymbol{p}$ throught the camera's perspective, creating a probability map via a differentiable tile rasterizer. The probability map is compared with the mask to calculate the loss.

can better manage the spatial dynamics of fruit positioning and density, a challenge not fully met by their 2D counterparts.

B. *3D point cloud segmentation*

The techniques of Structure from Motion (SfM) [5] have advanced the field of 3D reconstruction by translating sequential 2D imagery into expansive three-dimensional point clouds, laying the foundation for the subsequent phases of object detection and segmentation. The emergence of point cloud as a byproduct of these methods offers a depth of spatial information hitherto unavailable with traditional 2D approaches. Based on the generated point cloud, various methodologies such as [14] emerged for object detection within this point cloud. VoxNet [4] integrates a volumetric Occupancy Grid representation with 3D Convolutional Neural Network (3D CNN) for real-time object detection. However, the computational demands associated with 3D CNN are considerable. In response to the computational hurdles posed by VoxNet, subsequent advancements in object detection within point clouds have been introduced to refine the balance between computational efficiency and detection accuracy. Notably, [8] VoteNet have emerged as novel methodologies. Based on [11] PointNet and [12] PointNet++, VoteNet adapted Hough voting scheme for deep learning and 3D point clouds where each point in the input cloud votes for the potential center of an object to which it may belong. Independent from 3D voxel grids or multi-view images, it is particularly effective in scenarios where the objects are well-separated and not densely packed. While effective, the quality of the predictions heavily relies on the quality and density of the input point clouds. VoteNet struggle with objects with minimal presence in the point cloud, which is often the case with smaller fruits. This underscores the need for high quality point cloud data in applications leveraging VoteNet for accurate object detection, especially in the nuanced task of fruit counting where precision is important.

C. *Gaussian splatting-based segmentation*

The paper by [15] introduces Neural Radiance Fields (NeRF) as a novel representation technique for scenes, enabling continuous evaluation at any point in space from sparse point cloud inputs. While training and rendering with NeRF can be computationally intensive, it paved the way for the development of 3D Gaussian splatting [13]. This innovative method stands out for its ability to achieve rapid and high-quality 3D scene reconstruction.

Expanding upon this foundation, [2] Gaussian grouping and [1] Segment any 3D Gaussians (SAGA) has leveraged [16] Segment Anything (SAM) generated masks for object segmentation, integrating advanced detection and segmentation methods with Gaussian splatting. In the case of Gaussian grouping, a key feature is it's requirement for consistent identity across multiple views mask inputs. To address this challenge, a zero-shot tracker is employed within video sequence of multi-view images, which enhances the capacity of the system to track multiple objects effectively during camera movement. However, in specific applications like fruit counting on trees, the Gaussian grouping approach encounters practical limitations. The sheer number of fruits in orchards makes tracking each individual fruit not only computationally demanding but also costly in terms of resources. Furthermore, Gaussian grouping intertwines the segmentation process with 3D gaussian model training, potentially hampers its flexibility. This integration can lead to constrains when adapting the methodology to varied and complex scenarios, such as the diverse and unpredictable environments typically found in agricultural settings. While SAGA distinct in its delineation of segmentation and 3D gaussian model training, it falls short in its ability to identify all target objects through prompts. Addressing these limitations to fit for fruits detection and segmentation, this paper proposes a method that synthesizing the strengths of both SAGA and Gaussian grouping.

This approach capitalizes on the segmentation capabilities in three-dimensional space, integrating advanced techniques from 2D object detection and segmentation. Specifically, it

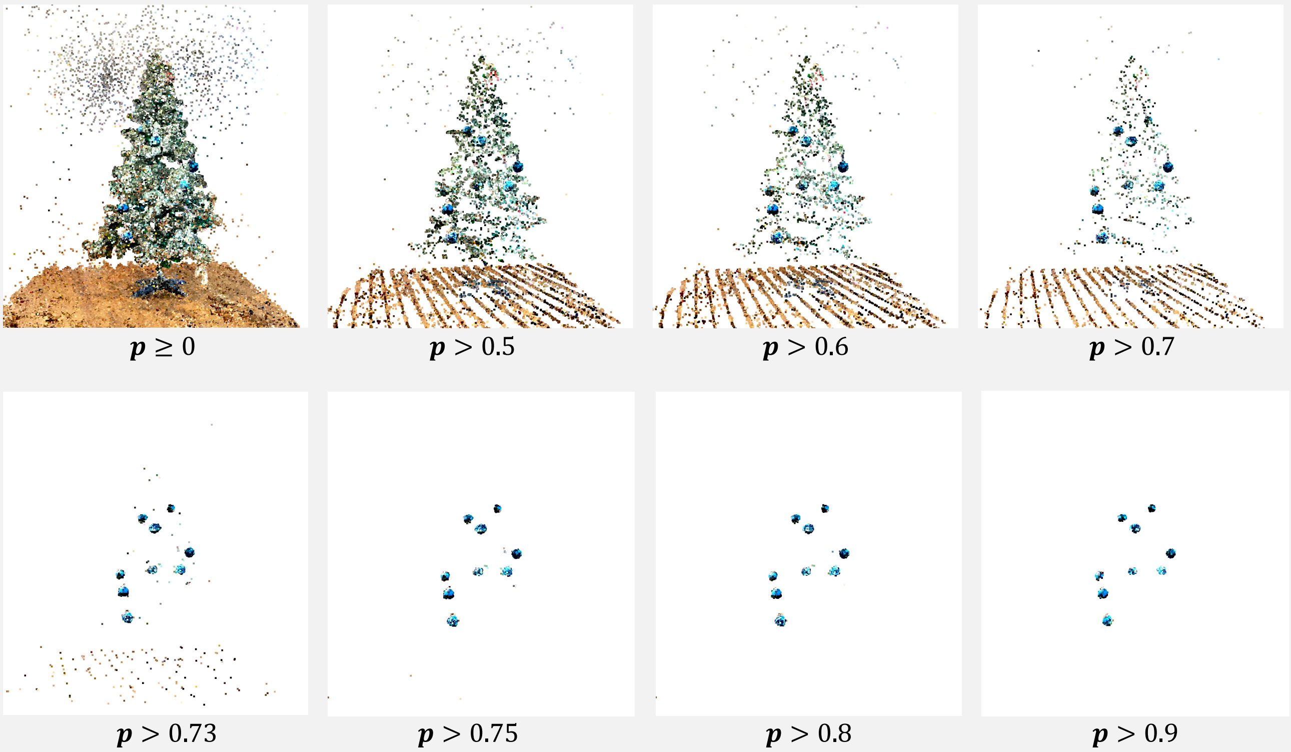


Fig. 3. Visualization and segmentation of a trained point cloud using open3d with varying probability thresholds. In this visualization, the trained point cloud is rendered in open3d, with points selectively filtered based on the attribute $\boldsymbol{p}$. As depicted in the sequence, it becomes apparent that thresholds exceeding 0.75 yield a distinctly clear and precise separation of the target object from the surrounding environment.

employs [17] Grounding DINO and SAM, coupled with 3D scene reconstruction through Gaussian splatting. This integration allows for a more nuanced and detailed analysis of the scenes, facilitating the identification and segmentation of target objects, such as fruits, in complex environments.

The proposed method marks an improvement in the field of fruit detection and segmentation, particularly in challenging environments like orchards where the density and distribution of fruits can vary greatly. Despite the challenges inherent in navigating the intricacies of detection and segmentation within the volumetric space of point clouds, this paper seeks to overcome these obstacles. The primary objective is to establish a robust framework capable of not only identifying the presence of objects but also precisely delineating and segmenting them within the rich, three-dimensional context of point cloud information. By integrating and building upon existing techniques, this approach aims to offer a more flexible, efficient, and accurate solution to the challenges of fruit counting and yield estimation in agriculture. Through this advancement, the research seeks to contribute to the broader goal of precision agriculture, where such technological innovations can lead to more informed decision-making and resource optimization.

III. Method

The method is implemented in Python using the Pytorch framework combined with C++ and custom CUDA kernels for rasterization part.

A. 3D Gaussian splatting pretrained model

The methodology outlined for 3D object segmentation in this paper adopts a multi-faceted approach (refer to Figure 2) that begins with the generation of a 3D point cloud from multi-view images. This process is facilitated using COLMAP. Once the point cloud is generated, it serves as the foundational dataset for training the Gaussian splatting model.

A key innovation in this approach is the modification of the Gaussian model file to include an additional attribute, termed ‘p’, for each point in the point cloud. This ‘p’ attribute is designed to represent the probability of each point belonging to the target object, reformulating the 3D object segmentation challenge into a binary classification problem.

The methodology entails two main components:

1. **Prediction Phase**: The predicted value is derived from the image rasterized by the 3D gaussian model from the perspective of the camera. This process involves mapping the 3D points back to a 2D plane from the viewpoint of the camera, thus providing a predicted probability map.
2. **Grounded Truth Alignment**: The ground truth value is obtained from the corresponding mask image, which precisely delineates the target object within the scene.

The crux of the method lies in minimizing the discrepancy between these predicted values and the ground truth. By reducing this difference, each point’s probability of belonging to the target object is effectively ascertained. To

TABLE I. CHRISTMAS TREE DATASET TRAINING DETAILS

Model	Training details			mBIoU (%)			
	iteration	***Total points***	***Time lapsed***	***0.7***	***0.75***	***0.8***	***0.9***
Christmas tree	3000	1,418,600	22m28s	52.7	17.4	1.1	0
Christmas tree	7000	1,418,600	57m03s	65.0	53.2	37.7	2.28
Christmas tree	10000	1,418,600	79m24s	66.3	57.9	46.6	15.2
Christmas tree	12000	2,112,900	127m28s	67.3	60.2	50.6	20.0

achieve this, the technique employs Binary Cross-Entropy, a standard loss function in binary classification tasks. This

function quantifies the difference between the predicted probabilities and the actual binary labels (whether the pixel is projected by the target object or not).

B. *Grounded Segment Anything mask input*

The proposed methodology in this research utilizes Grounded-SAM, an approach that combines stable diffusion with Grounding DINO and SAM for enhanced object detection and segmentation. Once the target objects are accurately detected and their boundaries established, the Segment Anything (SAM) is used to meticulously draw segmentation masks over the identified objects within their respective bounding boxes. The next step involves the use of the crafted mask images to train the Gaussian model. Each point is assigned an additional attribute – the likelihood of its belonging to the target object. This additional attribute is the key of the model's ability to distinguish between the target objects and the surrounding environment.

C. *Binary segmentation*

In this methodology, the objective is to distinguish between points that belong to the target objects and those that are part of the surrounding environment in a 3D point cloud. To achieve this, we introduced an extra attribute 'p' for each point, representing the probability that the point is part of the target object. This attribute complements existing point attributes such as position, scaling, rotation, color, and opacity.

- Gaussian Splatting Rendering:

In the Gaussian splatting rendering process, the color of each pixel is computed as a weighted sum of the colors of all points N overlapping that pixel ordered by depth. This calculation is based on the transparency (α) of each point:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (1)$$

Where C is the color rendered at a pixel, c_i is the color of point i, and α_i is the transparency of point i.

- Rasterization of Probability Map:

By integrating the probability attribute p (representing the likelihood of a point belonging to the target object), a similar approach is used to create a probability map:

$$G = \sum_{i \in N} p_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (2)$$

Here, G represents the rasterized probability map, and p_i is the probability that point i belongs to the target object. The resulting image is a grayscale map where the brightness of each pixel indicates the probability of the presence of the target object (with black background).

- Loss calculation.

The loss L is computed using Binary Cross-Entropy between the rasterized probability map and the ground truth masks. Binary Cross-Entropy is a suitable choice here as it measures the distance between two probability distributions – the predicted probabilities and the actual binary labels (0 or 1) in the ground truth masks.

To ensure the stability of the learning process and maintain the probabilities within a valid range, the values of p are clamped to remain between 0 and 1 after each iteration. This clamping is crucial to prevent probabilities from reaching non-physical values (less than 0 or greater than 1).

By employing this approach, our methodology effectively turns the problem of 3D object segmentation into a manageable binary classification task. The use of a probability map as an output of the Gaussian splatting process, coupled with a well-suited loss function, enables a nuanced and precise segmentation of target objects from the surrounding environment in the 3D point cloud.

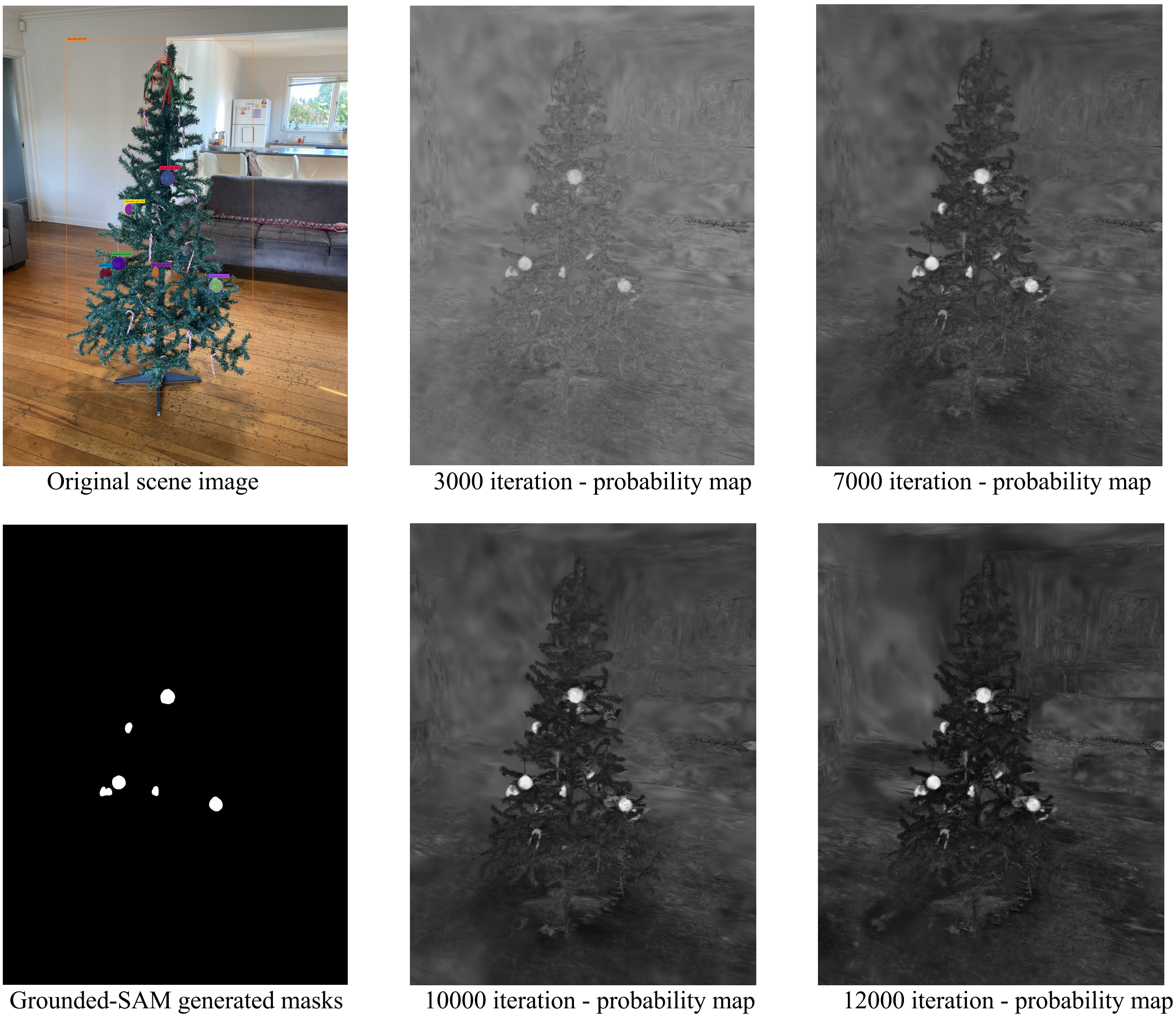
IV. RESULTS

A. *Experiment on an indoor static object – a Christmas tree*

In this study, we utilized a pre-constructed Gaussian model of a Christmas tree, encompassing 79 camera viewpoints and adorned with 10 blue spherical ornaments, to evaluate our segmentation model. A new attribute, denoted as p, was integrated into the pre-existing 3D Gaussian model, which represents the probability of each point belonging to the target object (in this case, the ornaments). The rasterization methodology employed is akin to standard Gaussian splatting [13], with the key distinction being the projection of attribute p's value instead of the conventional RGB values, resulting in the generation of a grayscale probability map as shown in figure 4.

Optimization of the model was carried out using the Adam optimizer, set at a learning rate of 0.0005. Through a series of experimental iterations, it was observed that a total of 7000 iterations were sufficient for the loss value to plateau, indicating optimal modal training. This observation was consistent across models trained for different iteration counts – specifically, models pre-trained for both 7000 iterations (yielding 1,418,600 points) and 30,000 iterations (resulting in 2,112,900 points). The dataset, featuring the Christmas tree, was trained using an RTX2080 GPU. The duration of training

Identify applicable funding agency here. If none, delete this text box.



Original scene image

3000 iteration - probability map

7000 iteration - probability map

Grounded-SAM generated masks

10000 iteration - probability map

12000 iteration - probability map

Fig. 4. Probability map rendered during model evaluation. As the iteration count increases, the trained model renders a probability map where the distinction between the object and the background becomes progressively more pronounced.

corresponding to different iteration counts is detailed in Table I.

In analyzing the results (refer to Figure 3), a threshold was applied to the probability values (p). It was observed that when $p > 0.75$, the model proficiently delineated the target objects (ornaments) from the background, yielding a clear and accurate segmentation. For the purpose of evaluation, the mean Binary Intersection over Union (mBIoU) metric was employed. Given the grayscale nature of the predicted images (as illustrated in Fiture 4), a specific threshold was set to binarize these images for the mBIoU calculation, as this metirc necessitates both the predicted and ground truth masks to be in binary format. The mBIoU scores, corresponding to varying threshold levels, are presented in Table I.

The experimentation underscores the efficacy of our proposed method in accurately segmenting target objects form complex backgrounds in 3D environments, demonstrating promising potential for broader applications in various fields.

B. *Experimental Evaluation on the LERF-Mask dataset*

To compare out model with other models, we tested our model with LERF-Mask dataset [2] and the result is shown in Table II.

TABLE II. COMPARING BETWEEN DIFFERENT MODELS

Model	**Figurines mBIoU (%)**			
LERF	30.6			
Gaussian Grouping	67.9			
Ours (only with promts: green apple)	0.7	0.75	0.8	0.9
	56.8	41.6	16.2	0

V. Limitation

The effectiveness of our segmentation approach is intrinsically linked to the precision of 2D object detection masks. As such, the quality of segmentation is heavily dependent on the accuracy of these 2D segmentation outcomes. However, this dependency is somewhat mitigated when multiple viewpoints are utilized, each providing different detection results for a single object. In scenarios where varying perspectives are available, the impact of any discrepancies in 2D detection accuracy is reduced, leading to a more robust and reliable segmentation performance.

VI. Conclusion

In conclusion, the method presented in this paper provides a robust and versatile framework for prompted binary segmentation of 3D point clouds, effectively bridging the divide between 2D object detection and 3D scene representation. The results showcased in this study highlight the system's ability to enhance segmentation accuracy while maintaining a balance between computational efficiency and detailed scene reconstruction. This approach significantly contributes to the ongoing advancement in the field of 3D object detection and segmentation.

Looking ahead, future research endeavors will focus on integrating VoteNet for the identification of object centroids within 3D point clouds. This addition aims to facilitate accurate object counting, further expanding the applicability of our methodology in various practical scenarios. By leveraging VoteNet's capabilities, we anticipate improvements in both the precision of object localization and the efficiency of object enumeration, thereby addressing some of the current limitations and opening new avenues in 3D spatial analysis.

References

[1] J. Cen *et al.*, "Segment Any 3D Gaussians," *arXiv.org*, Dec. 01, 2023. https://arxiv.org/abs/2312.00860 (accessed Jan. 24, 2024).

[2] M. Ye, "Gaussian Grouping: Segment and edit anything in 3D scenes," arXiv.org, Dec. 01, 2023. https://arxiv.org/abs/2312.00732

[3] S. Tu *et al.*, "Passion fruit detection and counting based on multiple scale faster R-CNN using RGB-D images," *Precision Agriculture*, vol. 21, no. 5, pp. 1072–1091, Jan. 2020, doi: 10.1007/s11119-020-09709-3.

[4] "VoxNet: A 3D Convolutional Neural Network for real-time object recognition," *IEEE Conference Publication | IEEE Xplore*, Sep. 01, 2015. https://ieeexplore.ieee.org/document/7353481

[5] M. Crocco and C. Rubino, "Structure from Motion with Objects," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, doi: 10.1109/cvpr.2016.449.

[6] M. Rahnemoonfar and C. Sheppard, "Deep count: fruit counting based on deep simulated learning," *Sensors*, vol. 17, no. 4, p. 905, Apr. 2017, doi: 10.3390/s17040905.

[7] J. Hu, C. Fan, Z. Wang, J. J. Ruan, and S. M. Wu, "Fruit detection and counting in apple orchards based on improved Yolov7 and Multi-Object tracking methods," *Sensors*, vol. 23, no. 13, p. 5903, Jun. 2023, doi: 10.3390/s23135903.

[8] C. R. Qi, "Deep hough voting for 3D object detection in point clouds," *arXiv.org*, Apr. 21, 2019. https://arxiv.org/abs/1904.09664

[9] N. Häni, P. Roy, and V. Isler, "Apple Counting using Convolutional Neural Networks," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2018, doi: 10.1109/iros.2018.8594304.

[10] M. V. Afonso *et al.*, "Tomato fruit detection and counting in greenhouses using deep learning," *Frontiers in Plant Science*, vol. 11, Nov. 2020, doi: 10.3389/fpls.2020.571299.

[11] C. R. Qi, "PointNet: Deep learning on point sets for 3D classification and segmentation," *arXiv.org*, Dec. 02, 2016. https://arxiv.org/abs/1612.00593

[12] C. R. Qi, "PointNet++: deep hierarchical feature learning on point sets in a metric space," *arXiv.org*, Jun. 07, 2017. https://arxiv.org/abs/1706.02413

[13] B. Kerbl, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *arXiv.org*, Aug. 08, 2023. https://arxiv.org/abs/2308.04079

[14] C. Rubino, M. Crocco and A. Del Bue, "3D Object Localisation from Multi-View Image Detections," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1281-1294, 1 June 2018, doi: 10.1109/TPAMI.2017.2701373.

[15] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, Dec. 2021, doi: 10.1145/3503250.

[16] A. Kirillov, "Segment anything," *arXiv.org*, Apr. 05, 2023. https://arxiv.org/abs/2304.02643

[17] S. Liu, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," *arXiv.org*, Mar. 09, 2023. https://arxiv.org/abs/2303.05499

[18] J. P. Vásconez, J. Delpiano, S. Vougioukas, and F. A. Cheein, "Comparison of convolutional neural networks in fruit detection and counting: A comprehensive evaluation," *Computers and Electronics in Agriculture*, vol. 173, p. 105348, Jun. 2020, doi: 10.1016/j.compag.2020.105348.