Querying 3D point clouds exploiting open-vocabulary semantic segmentation of images

Ashkan Alami[1,2], Fabio Remondino [1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: (aalami, remondino)@fbk.eu

[2] Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy

KEYWORDS: deep learning, point cloud, query, open-vocabulary

ABSTRACT
While deep models have advanced the 3D data analysis and demonstrated impressive results, they often struggle to generalize to new classes that are absent from the training dataset. Recently, open-vocabulary and zero-shot models have addressed this problem. However, these models are still relying on some data for training and fine-tuning for specific tasks. This requirement limits them to real-world applications. In this research, we propose an open-vocabulary method for point cloud segmentation, which does not require additional training data beyond the images and point cloud from the survey scene. By using the capabilities of the power of 2D open-vocabulary models and geometric features from the 3D data, combined with an XGBoost-guided region growing algorithm, our approach segments the queried objects directly in 3D scenes. We evaluate our method on 3D benchmark datasets, such as Replica and ScanNet, showing its practicality and scalability to real-world scenarios with limited data.

1. INTRODUCTION

Scene understanding is one of the most important aspects in photogrammetry, remote sensing and computer vision (Fooladgar, et al., 2015; Heipke et al., 2020). The goal is normally to extract semantic information, such as object location or identification, from images or 3D point cloud. This ability has many possible uses in robotics, autonomous driving, territorial monitoring, smart city applications, augmented reality, etc. Traditionally, methods for understanding 3D scenes relied heavily on geometric and sensor-specific features (Weinmann, et al., 2017; Grilli and Remondino, 2020). While these methods are still effective in case of small annotation sets or projects with uncommon classes, we have witnessed the rise of deep learning methods, in particular for 3D point cloud segmentation, with models such as PointNet (Qi, et al., 2016), KPConv (Thomas et al., 2019), and Point transformer (Zhao et al., 2021). However, despite their success, these models face challenges in generalizing across diverse object classes due to the limited range of training (2D and 3D) data. In response to these limitations, zero-shot learning techniques have been developed. In image processing, models like MDETR (Kamath, et al., 2021), CLIP (Radford, et al., 2021) or Grounding DINO (Liu, et al., 2023) are beginning to combine text and vision. Recently, zero-shot learning approaches have been applied to 3D data such as point clouds. mainly by using large-scale pre-trained Vision-Language Models (VLMs) to increase performance in previously unexplored classes. These models (Chen et al., 2023; Ding et al., 2023; Huang et al., 2024; Zhang et al., 2023) successfully transfer knowledge from 2D to 3D understanding, allowing for the segmentation and recognition of unusual items in 3D space. For instance, OpenScene (Peng, et al., 2022) proposes a zero-shot approach for 3D scene understanding that co-embeds dense 3D point features with image pixels and text in the CLIP feature space. However, these approaches often require some level of training and lack generalization to diverse data types. For example, OpenScene requires training the 3D model on projected CLIP space features. Similarly, OpenMask3D (Takmaz, et al., 2023), a model for open-vocabulary 3D instance segmentation, uses Mask3D (Schult, et al., 2022) for class-agnostic masks identification and retrieves semantic information via CLIP. Beside some current limitations and challenging, the tasks of querying and accessing 3D point clouds can lead to multiple interesting applications.

1.1. Paper aims

In this paper, we present a novel, training-free method for open-vocabulary 3D point cloud segmentation. Our approach integrates Vision-Language Models with traditional geometric features and Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) techniques, allowing effective segmentation without dataset-specific training. This makes it an adaptable and cost-effective solution for querying and inspecting a wide range of 3D datasets.

2. RELATED WORKS

2.1. 3D point cloud analysis

3D point cloud analysis includes detecting and segmenting objects within a scene and understanding their relationships. Traditionally, these methods rely on geometric features that are derived from the spatial arrangement and relationships of the points cloud. These features are used with the classical clustering methods like DBSCAN, RANSAC, and random forest, that cluster the point into meaningful groups (Ni et al., 2017; Czerniawski et al., 2018; Grilli and Remondino, 2020; Zhou et al., 2022; You et al., 2024). With the rise of deep learning, researchers have begun to use deep models on point clouds, and some models, such as PointNet (Qi et al., 2017), have shown promising results in point cloud classification. Since then, a variety of models have been developed, highlighting improved performance (Phan et al., 2018; Hu et al., 2020; Shinohara et al., 2020; Zhao et al., 2021). Recent models, such as Mask3D (Schult et al., 2023), take advantage of the transformer's architecture to become the state-of-the-art in object segmentation. However, one significant limitation of these models is their reliance on fixed labels from training data, which limits their adaptability across different object classes and reduces robustness when encountering novel or untrained categories.

2.2. Open vocabulary in images

Open vocabulary models aim to overcome the limitations of the classic deep models by detecting or segmenting the novel classes during the test. In computer vision, these models are bridging between the language and vision (Lu et al., 2019; Li et al., 2022; Jia et al., 2021; Singh et al., 2022). A notable example is CLIP (Radford et al., 2021), which uses two different encoders to create a shared embedding space between

images and text, connecting the vision and language. Some other models use a different strategy, combining encoded features from images and text. For example, Grounding DINO (Liu et al., 2023) employs two encoders for image and text, combining their features via a fusion module that selects relevant queries from an image. Another example is YOLO-World (Cheng et al., 2024), which is an improvement on open-vocabulary object detection from the YOLO series (Redmon, 2016; Redmon and Farhadi, 2018; Ao Wang, 2024).

2.3. Open vocabulary in 3D scenes

By the success of open vocabulary models in images, researchers started to explore how to adapt these methods for point cloud segmentation. However, due to the scarcity of data and the large size of point clouds, directly applying image-based approaches is inadequate. As a result, open vocabulary methods for point clouds are leveraging power from image-based open vocabulary models and applying it to the 3D world. One approach is distilling the information from vision-language models (VLMs) into the point cloud (Delitzas et al., 2023; Zhang et al., 2023; Huang et al., 2024). Notable example OpenScene (Peng et al., 2023), which uses OpenSeg (Ghiasi et al., 2022), an open-vocabulary image segmenter, to extract text-based representations from images. These representations are then projected onto 3D points, allowing a 3D model to be trained on these points. Other methods are creating 3D masks from point clouds and then aligning these masks with language information extracted from the corresponding images, establishing a link between 3D structures and text descriptions. For instance, OpenMask3D (Takmaz et al., 2023) employs Mask3D (Schult et al., 2023) to generate agnostic object masks and then uses CLIP to project text features onto these masks. Similarly, Open3DIS (Nguyen et al., 2024) expands on this idea, refining both mask proposals and mask classifications with image data, leading to a better performance. Although open-vocabulary methods perform outstandingly in scene understanding, several limitations need to be addressed. Even if the methods do not necessarily require annotated data, they still need relevant training samples for the training. In addition, when applied to the unique survey scenes that differ significantly from common 3D datasets, these methods often produce poor results, struggling to generalize beyond the training data. As a solution for this limitation, we propose our method, which requires no additional training data and suitable for any 3D scene.

3. METHODOLOGY

The proposed method assumes to have a point cloud of a surveyed scene, an image dataset representing the scene and the corresponding camera poses (dashed box in Figure 1). Given these input data, to achieve our aim, the following steps are executed:

- **Search the object(s) of interest through a query:** The approach utilizes the YOLO-World (Cheng et al., 2024) and Grounding DINO (Liu et al., 2023) open-vocabulary object detection model to perform image queries. Generates precise bounding box for the target object (e.g., a sofa).
- **Extract the object's mask:** The bounding box is fed into the Segment Anything Model (SAM) (Kirillov, et al., 2023; Ren, et al., 2024) to extract the object mask around the target object. The mask is passed through a simple kernel to shrink it and prevent any misprojection of pixels around the object.
- **Projection in 3D:** Given the scene's point cloud and the camera parameters, the pixels within object's mask are projected onto the 3D points. The process involves:
  - Voxelization: Such discretization process simplifies and enhances the efficiency of subsequent processes.
  - Ray casting: Rays from the camera's viewpoint are projected into the 3D space using the Open3D library (Zhou, et al., 2018); ray intersections with the voxel grid are determined identifying which voxels are impacted by the passing-through rays.
  - Voxel labeling: Intersected voxels are labelled assigning tags from the 2D image masks.
  - Point labeling: Each point contained in the labelled voxel receives the same label.
  - Refining the projection: To ensure projection accuracy, DBSCAN is applied to the projected points, with only the largest cluster retained.
- **Label assigning and cluster merging:** For each point, the semantic label with the highest detection confidence across projections is assigned. Nearby clusters with identical semantic labels are then merged to refine the segmentation. Following this, DBSCAN is applied to the labeled points to further cluster them, and an instance ID is assigned to each cluster.
- **Geometric feature calculation** for each point.
- **3D object refinement:** The projected points normally do not represent the entire queried object. Therefore, leveraging on geometric features and a clustering method, similar points are added to create a more complete representation. The following steps are performed for each semantic Cluster for example i^{th} is S_i:
- **Creating data:** The color and features of each point is considered as the input data. For training, we only used the points that a mask is projected on. The ground truth is a binary array whether the label is S_i or not.

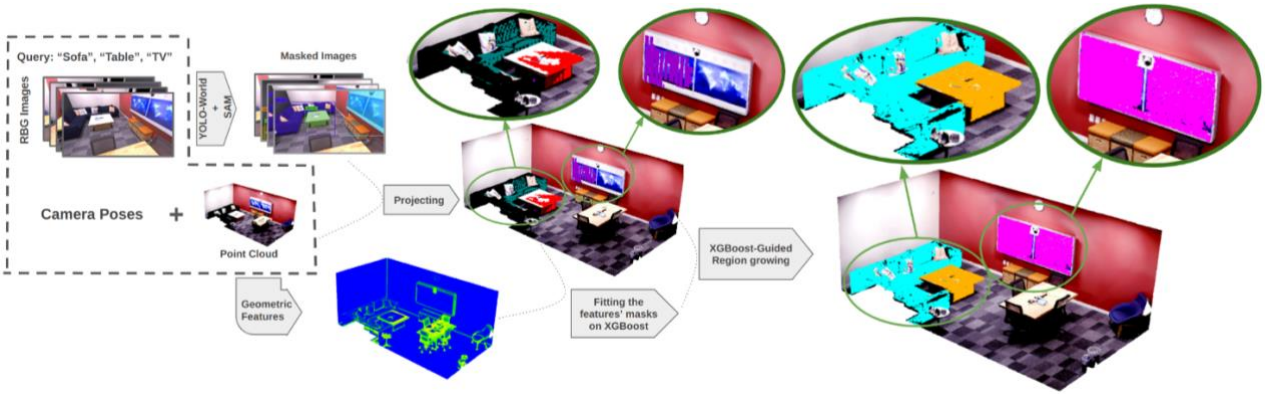

Figure 1. The proposed pipeline with its input data (dashed box) and different steps: extracted masks of desired objects using an open-vocabulary tool; projection onto the point cloud; geometric features; fitting the features on XGBoost; region growing with the XGBoost.

- **Create and train a model:** We used XGboost, which is a scalable machine learning system for tree boosting. The model trained and fitted to the training data.
- **Region growing** for each cluster with semantic label S_i we do the following (see Algorithm 1):
  - Choose a Point: A point is the cluster is chosen.
  - Picking Nearest Neighbors: Pick the K nearest neighbors of the chosen point that are not yet part of the cluster.
  - Predict Semantic Label: The model is used to predict the label for each of the neighbors.
  - Expand the Cluster: If more than half of these neighbors are predicted to belong to S_i, these points are added to the cluster.
  - Repeating: Repeating this process for each point in the expanded cluster until no additional points can be added.
- **Merging extended clusters:** As the final step, extended clusters with the intersection are merged and the bigger ones with higher confidence score from the detection are considered as the final semantic label.

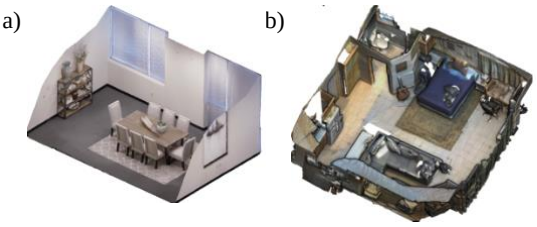

Figure 2: The Replica[1] (a) and ScanNet[2] indoor scenes used for the evaluation of the proposed methodology.

4. EXPERIMENTS SETUP

Datasets. The method was evaluated on two indoor datasets - the Replica[1] (Straub, et al., 2019) and ScanNet[2] (Dai, et al., 2017). Replica is a high-resolution synthetic dataset designed for 3D scene understanding. It contains detailed indoor environments with various object classes and ground truth annotations. Our method was evaluated on the scenes: *room0, room1, room2, office0, office1, office2, office3, office4*. For each scene the RBG images (8-10 images per scene), corresponding camera parameters, and the point cloud are used. We aimed to segment 48 objects that can be found within these scenes. In addition, an experiment was conducted on the ScanNetv2 validation set, which consists of 312 scenes. In this experiment, in contrast to Replica, about 150 images per scene are used to segmenting 20 objects that are defined in the dataset. Due to the low quality of data, floor is detected by RANSAC and excluded from the ScanNet's scenes.

2D Models. For object detection, we utilize YOLO-World and Grounding DINO, though it can be substituted with any open vocabulary or even close-set object detectors. In the experiment on ScanNet, only YOLO-World is used, since in initial testing this model has shown a better performance on the ScanNet images. For detection confidence, 0.45 is used for YOLO-World and 0.4 for Grounding DINO. For segmentation, the Segment Anything Model (SAM) with the default ViT-H SAM model is used.

Region Growing method. As input to the XGBoost model, we used the color of each point along with features such as linearity, planarity, omnivariance, anisotropy, eigenentropy, verticality, sphericity, flatness, compactness, curvature change, and shape index. To calculate these features, we first determine the maximum-minimum distance between one pair of points, then compute the features for three radii: 1.5, 5, and 10 times this distance. These values were selected based on dataset quality and initial testing. In the region-growing algorithm, 5 neighboring points are checked for each point (k=5), and for the XGBoost threshold, 0.96 is applied.

Metrics. Our method is evaluated for both semantic segmentation and instance segmentation, following the evaluation procedures of ScanNet. For semantic segmentation, the Intersection over Union (IoU) is calculated for each semantic mask. For instance segmentation, the average precision (AP) is calculated. According to ScanNet evaluation method, AP scores are averaged over the overlap range of [0.5 : 0.95 : 0.05] at mask overlap levels of 50% and 25%.

Algorithm 1 - Region Growing Algorithm

Require: Point cloud $P \in R^{N\times 3}$
Require: Feature vector $X \in R^{N\times r}$
Require: Initial points indices $I = \{i_1, \dots, i_m\}$
Require: Trained model M
Require: Number of neighbors K
Require: Confidence threshold θ
Ensure: Cluster of points C
Function RegionGrowing*(P, X, I, M, k, θ)*
  tree ← KDTree(*P*)
  $C \leftarrow I$ {Initialize cluster with initial points}
  $Q \leftarrow I$ {Queue of points to process}
  While $Q \neq \emptyset$ *do*
    $P \leftarrow Q$.pop(0) {Dequeue point}
    $(d, N) \leftarrow$ tree.query*(P[p], k + 1)*
    $N \leftarrow N[1:]$ {removing the chosen point from neighbors}
    $Conf \leftarrow$ M.predict_proba($X[N]$) [:, 1] {Get the confidence}
    $valid\_N \leftarrow \{ n_i \in N \mid conf[i] > \theta \text{ and } n_i \notin C \}$
    if $|valid\_N| \geq \frac{k}{2}$ *then*
      $C \leftarrow C \cup valid\_N$
      $Q \leftarrow Q \cup valid\_N$
    end if
  end while
return C
end function=0

5. RESULTS AND DISCUSSION

The proposed method is evaluated for both semantic and instance segmentation, focusing mainly on comparison with the projected mask as baseline. As reported in Table 1 and 2, our method shows notable improvements in segmentation across various objects categories, particularly for larger or more distinct objects like sofas and benches.

As mentioned, few images per scene are used for the experiment. Remarkably, testing with only 2-3 images still achieved effective segmentation. This highlights that the method is effective in scenarios with limited 2D images. Additionally, as the approach leverages foundation models, it is adaptable to various pre-trained 2D models, making it suitable for specific tasks where a pre-trained 2D model is available. As an additional evaluation, the model was tested on semantic segmentation using the ScanNet validation dataset, indicating great performance without requiring training, as shown in Table 3. It highlights the approach's adaptability to diverse data. Moreover, since the method does not require training, extensive or similar datasets are no longer required. It performs well with only the desired scene images, making it efficient and flexible to a variety of scenarios. As reported in Figures 3 and 4, the proposed method performs well across various scenarios and objects and the refinement process allow to detect object quite completely.

[1] https://github.com/facebookresearch/Replica-Dataset

[2] http://www.scan-net.org/

Evaluation on Replica				
model	**metric**	**Projection**	**Refinement**	**Δ**
Ground. DINO	AP	0.0	0.06	**+0.06**
	AP_{50}	0.01	0.13	**+0.12**
	AP_{25}	0.01	0.25	**+0.24**
	IoU	0.14	0.22	**+0.08**
YOLO-World	AP	0.01	0.02	**+0.01**
	AP_{50}	0.02	0.06	**+0.04**
	AP_{25}	0.04	0.21	**+0.17**
	IoU	0.10	0.21	**+0.11**

Table 1. Results of semantic and instance segmentation on the Replica dataset, with some selected objects. The proposed method is the column Refinement. Evaluations are conducted using YOLO-World and Grounding DINO. The last column (Δ) highlights the improvement achieved by the refinement step. In the averaged results, objects that are not detected by the models are excluded from the table.

Segmentation results for selected objects in Replica				
object	**model**	**Projection**	**Refinement**	**Δ**
bench	Ground. DINO	0.12	0.23	**+0.11**
	YOLO-World	0.28	0.64	**+0.36**
chair	Ground. DINO	0.09	0.20	**+0.11**
	YOLO-World	0.09	0.25	**+0.16**
clock	Ground. DINO	0.26	0.43	**+0.17**
	YOLO-World	0.20	0.39	**+0.19**
picture	Ground. DINO	0.21	0.25	**+0.04**
	YOLO-World	0.22	0.36	**+0.14**
pillow	Ground. DINO	0.11	0.16	**+0.05**
	YOLO-World	0.22	0.32	**+0.10**
sofa	Ground. DINO	0.08	0.24	**+0.16**
	YOLO-World	0.22	0.54	**+0.32**

Table 2. Objects' segmentation performances for Grounding DINO and YOLO-World models, comparing results between the baseline projection method and our proposed refinement step. Performance is evaluated using the IoU metric. The last column (Δ) illustrates the improvement achieved by the proposed refinement.

Some dependencies and challenges of the proposed approach are:

Dependence on 2D models: The method is dependent on 2D models. If the 2D model fails to detect or incorrectly recognizes an object, false semantic information will be mapped to the point cloud and create an error in segmentation. As shown in Figure 5, an example of false detection and resulting errors in the final segmentation. Therefore, a strong and reliable 2D model is required. Furthermore, the target object must be visible in the images for segmentation, otherwise the object will not be detected by model.

Projection accuracy: Compared to common point cloud segmentation methods, we exclude the depth information in our projections, resulting in some missed projections. Inaccurately projected points may cause problems. For example, if a projected mask for a "chair" overlaps with another object, such as a "table", a portion of the table may be mislabelled as a chair during refinement.

Quality of point cloud data: High-quality point clouds are recommended, as detailed and well-defined geometric features improve refining accuracy. Refinement is less effective with low-quality data, for instance in our test small objects like wall plugs were problematic. In these cases, even a minor misprojection can cause huge segmentation errors.

Variety of object detection: The method performs best when a variety of objects are queried and detected, as the refinement relies on distinguishing between different objects' features. With a limited scope of objects, the method may over-segment and include unintended areas.

6. CONCLUSIONS

In this work, we introduced a training-free method for open-vocabulary object segmentation in 3D point clouds. By leveraging the abilities of open-vocabulary foundation models for images and geometric features from point clouds, we developed an XGBoost-based region-growing method that can detect and segment any desired object in a 3D scene. Unlike other common methods, this approach does not require depth information or additional training data, making it highly adaptable for different scenarios, especially when there is a single survey scene. Furthermore, the method can use any 2D detector, whether open-set or closed-set, enhancing its flexibility and adaptability. However, the performance is highly dependent on the 2D detector and projection, which can pose challenges in cases of missed detection or incorrect projections due to camera calibrations. As a future direction, this approach could be explored with outdoor and LiDAR data.

REFERENCES

Ao Wang, Hui Chen, L. L. e. a., 2024. YOLOv10: Real-Time End-to-End Object Detection. *arXiv preprint arXiv*:2405.14458.

Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., Wang, W., 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip. *Proc. CVPR*, 7020-7030.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. Proc. 22nd *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, pp. 785-794.

Evaluation on ScanNet								
Average IoU	cabinet	bed	chair	sofa	table	door	window	bookshelf
0.244625	0.453	0.391	0.47	0.003	0.044	0.22	0.001	0.375

Table 3. Semantic segmentation performance on the ScanNet validation set, with selected objects as examples. Average IoU excludes wall, floor and undetected objects by 2D detector.

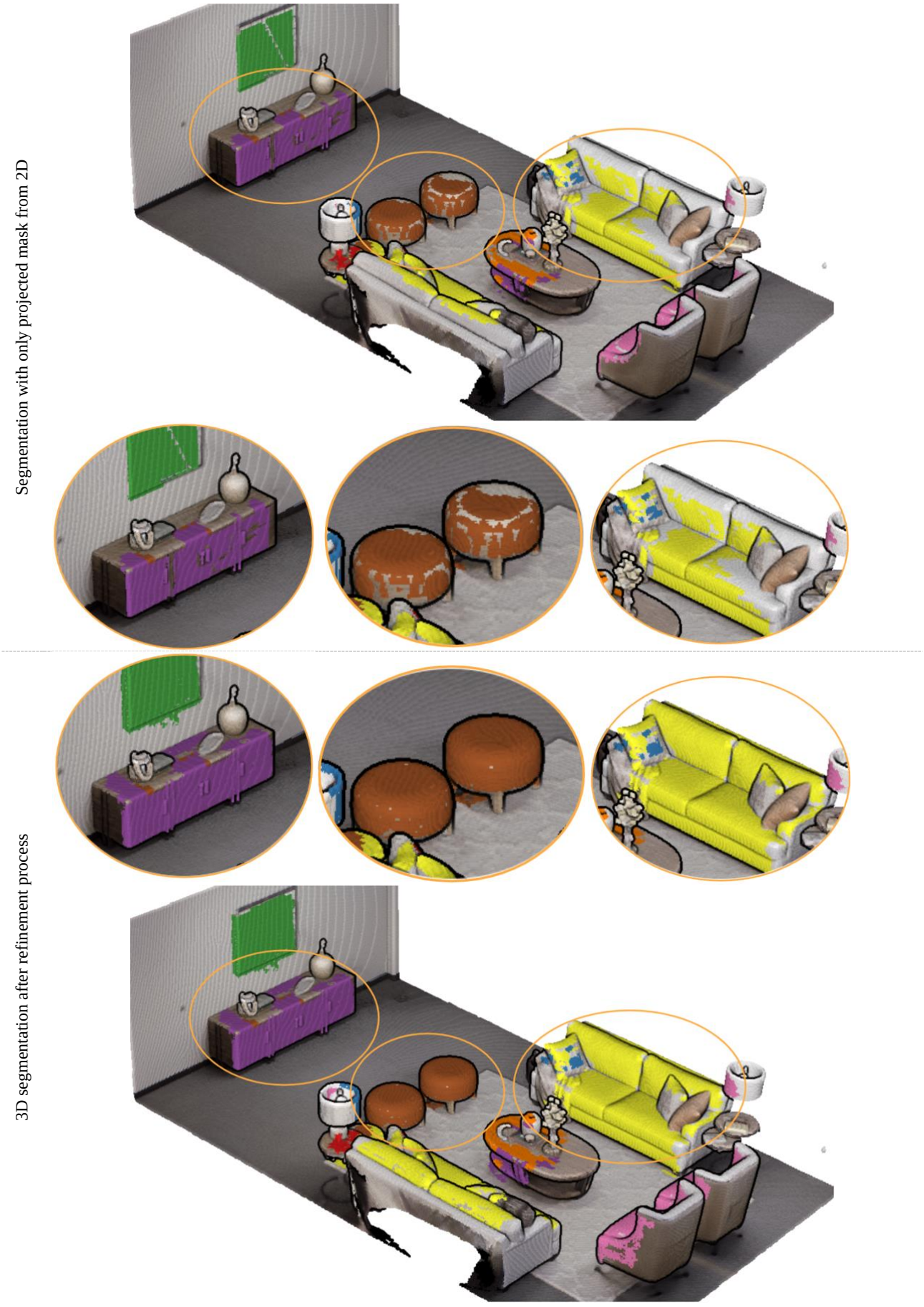

Figure 3. Results on a scene from the Replica dataset showing the 3D scene with projected labels from the images (top) the refined queried 3D scene after proposed method (bottom).

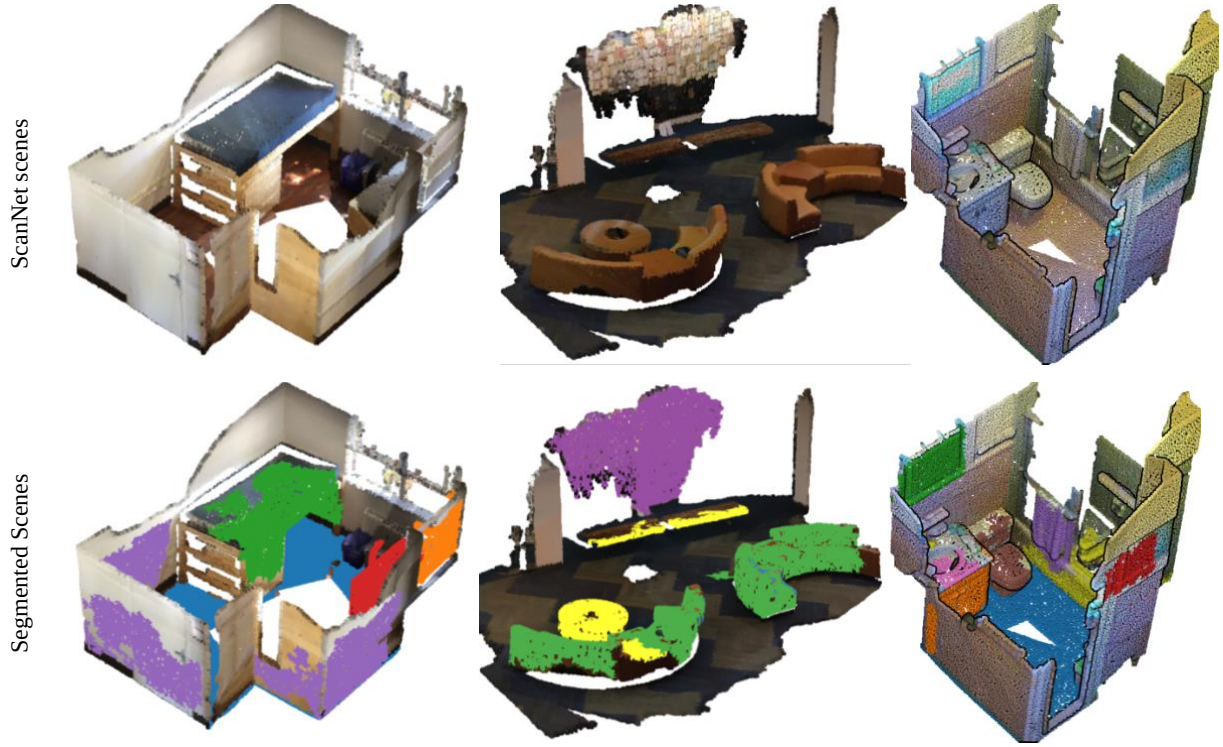

Figure 4. Results using the ScanNet dataset showing a robust performance across diverse environment, including an indoor room, an outdoor bench and a shower.

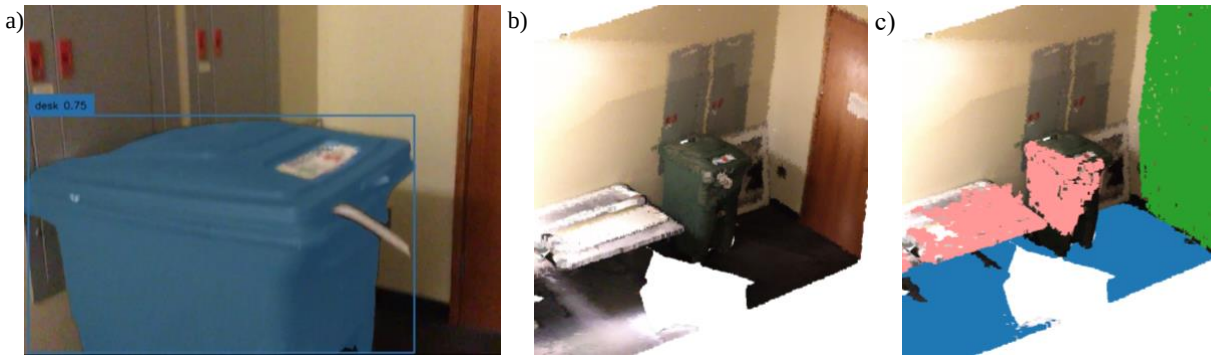

Figure 5. The impact of errors in the 2D model on 3D segmentation. The 2D model (left) misclassifies a bin as a desk, leading to incorrect labeling in the final 3D segmentation (right). Both the bin and the adjacent desk are assigned the same label (pink), highlighting the propagation of errors from 2D to 3D models.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y., 2024. Yolo-world: Real-time open-vocabulary object detection. *Proc. CVPR*, pp. 16901-16911.

Czerniawski, T., Sankaran, B., Nahangi, M., Haas, C., Leite, F., 2018. 6D DBSCAN-based segmentation of building point clouds for planar object classification. *Automation in Construction*, 88, 44-58.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Proc. CVPR*.

Delitzas, A., Parelli, M., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., Hofmann, T., 2023. Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. *arXiv preprint arXiv*:2306.02329.

Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X., 2023. Pla: Language-driven open-vocabulary 3d scene understanding. *Proc. CVPR*, 7010–7019.

Fooladgar, F., Kasaei, S., 2015. Semantic segmentation of rgb-d images using 3d and local neighbouring features. *Proc. Int. Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp/ 1-7.

Ghiasi, G., Gu, X., Cui, Y., Lin, T.-Y., 2022. Scaling open-vocabulary image segmentation with image-level labels. Proc. ECCV, pp. 540–557.

Grilli, E., Remondino, F., 2020. Machine learning generalisation across different 3D architectural heritage. *ISPRS International Journal of Geo-Information*, 9(6), 379.

Heipke, C., Rottensteiner, F., 2020. Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. *Geo-spatial Information Science*, 23(1), 10–19.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. *Proc. CVP*R, pp. 11108–11117.

Huang, Z., Wu, X., Chen, X., Zhao, H., Zhu, L., Lasenby, J., 2024. OpenIns3D: Snap and Lookup for 3D Open-vocabulary Instance Segmentation. European Conference on Computer VisionJia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *International conference on machine learning*, PMLR, 4904–4916.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N., 2021. Mdetr - modulated detection for end-to-end multi-modal understanding. *Proc. ICCV*, pp. 1760–1770.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al., 2023. Segment anything. *Proc. CVP*R, 4015-4026.

Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International conference on machine learning, PMLR, 12888–12900.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J. et al., 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv*:2303.05499.

Lu, J., Batra, D., Parikh, D., Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Nguyen, P., Ngo, T. D., Kalogerakis, E., Gan, C., Tran, A., Pham, C., Nguyen, K., 2024. Open3dis: Open-vocabulary 3D instance segmentation with 2d mask guidance. *Proc. CVPR*, pp. 4018–4028.

Ni, H., Lin, X., Zhang, J., 2017. Classification of ALS Point Cloud with Improved Point Cloud Segmentation and Random Forests. *Remote Sensing*, 9(3).

Peng, S., Genova, K., Jiang, C. M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., 2023. Openscene: 3d scene understanding with open vocabularies. Proceedings of the *Proc. CVP*R.

Phan, A. V., Le Nguyen, M., Nguyen, Y. L. H., Bui, L. T., 2018. Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Networks*, 108, 533–543.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. CVP*R, 652–660.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning.*

Redmon, J., 2016. You only look once: Unified, real-time object detection. *Proc. CVP*R.

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv*:1804.02767.

Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B., 2023. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. *Proc. ICRA*.

Shinohara, T., Xiu, H., Matsuoka, M., 2020. FWNet: semantic segmentation for full-waveform LiDAR data using deep learning. *Sensors*, 20(12), 3568.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D., 2022. Flava: A foundational language and vision alignment model. *Proc. CVP*R, pp. 15638-15650.

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., Newcombe, R., 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv*:1906.05797.

Takmaz, A., Fedele, E., Sumner, R. W., Pollefeys, M., Tombari, F., Engelmann, F., 2023. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. *Proc. NeurIPS*.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. *Proc. ICCV*, 6411–6420.

Weinmann, M., Hinz, S., Weinmann, M., 2017. A hybrid semantic point cloud classification-segmentation framework based on geometric features and semantic rules. *PFG–Journal*, Vol. 85, pp. 183-194.