RE0: Recognize Everything with 3D Zero-shot Open-Vocabulary Instance Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we introduce a novel zero-shot 3D instance segmentation framework called **RE0**. We leverage the 3D geometry information in 3D point cloud, the projection relationship between 3D point cloud and multi-view 2D posed RGB-D frames and the semantic features extracted by CLIP from multi-view 2D posed RGB-D frames to address the challenge of 3D instance segmentation. Specifically, we utilize Cropformer to extract mask information from multi-view posed images, combined with projection relationships to assign point-level labels to each point in the point cloud, and achieve instance-level consistency through inter-frame information interaction. Then, we employ projection relationships again to assign CLIP semantic features to the point cloud and achieve aggregation of small-scale point clouds. Due to the particularity of zero-shot 3D instance segmentation, we introduce the 3D open-vocabulary task to evaluate our method. Notably, **RE0** does not require any additional training and can be implemented by supporting only one inference of Cropformer and one inference of CLIP. Experiments on ScanNet200 benchmark show that our method achieves higher quality segmentation than the previous zero-shot methods. Besides, our method even surpasses the human-level annotations in many cases. Our project page is available at https://recognizeeverything.github.io/

1 Introduction

With the development of technologies such as autonomous driving, robotics, and virtual reality[1, 5, 41], 3D instance segmentation, a fundamental task in 3D computer vision, is increasingly demonstrating its importance. Its target is to predict 3D object instance masks from input 3D scenes like meshes, point clouds, and posed RGB-D frames. Traditional 3D instance segmentation methods[2, 7, 9, 26, 31, 33, 35, 39] are data-driven, and are trained on close-set dataset. Although these methods have made some progress, they still cannot solve the increasing requirements of data and resources.

In 2D segmentation area, Segment Anything Model[11] brings a breakthrough. After training on SA-1B dataset, SAM can segment any unknown image without further training. Previous methods like [6, 36, 37] utilize projection, graph neural network, and other information to build the connection between 2D and 3D to realize 3D segmentation. These methods sometimes do not generate results that meet our expectations due to the granularity control relationship of the SAM Prompt encoder. Sometimes the granularity is too fine, and sometimes it is not fine enough, as shown in Fig 1. We believe that, on the one hand, this is because it is difficult to manually control the granularity of the masks produced by SAM. On the other hand, these methods still have certain flaws in keeping the consistency of 3D instances.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

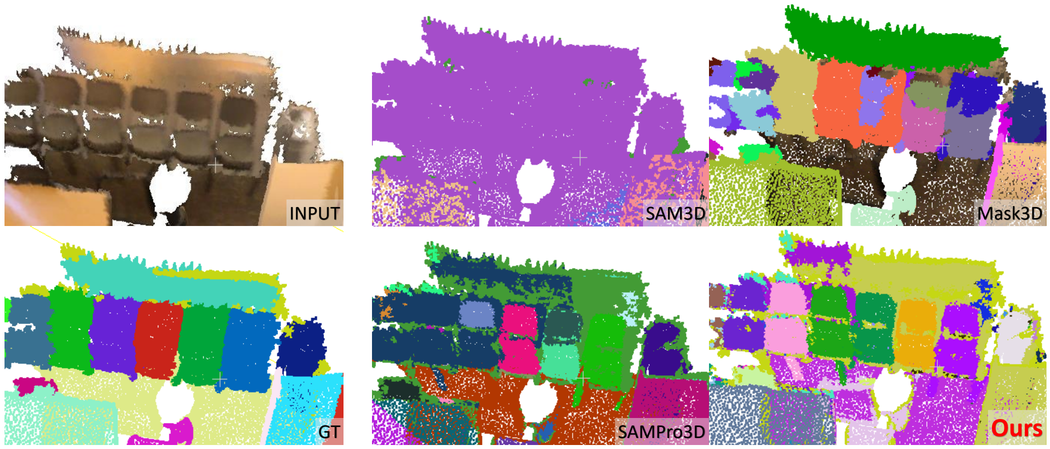

Figure 1: **Comparison of related works**. The visualization results of different methods are shown above. Input of this figure contains six chairs and one rubbish bin. Recognizing the six similar neighboring chairs is hard. For zero-shot methods like SAM3D and SAMPro3D, they either completely collapse or recognize adjacent objects as the same category; for training-based method, Mask3D feels ambiguity on this scene; however, our framework **RE0** has the ability to segment all the six chairs completely and accurately.

To solve these issues, we propose a novel framework called **RE0** for indoor scenes. Followed by some previous works, RE0 uses a pre-trained 2D segmentation model to generate masks for RGB-D frames. Then, we use a Mask-Based Segmentation approach which leverages the projection relationship between 2D and 3D to achieve consistency across mask frames and produce a preliminary 3D segmentation result. Subsequently, a Mask-Based Merge Module is employed to exploit the projection relationship and CLIP semantic features to integrate fine-grained segmentation results into a complete segmentation granularity which aligns with CLIP semantic features.

However, zero-shot 3D instance segmentation presents a common challenge: the evaluation of segmented point cloud instances within standard close-set datasets is hindered by the difficulty in determining the correspondence between point clouds. To address this challenge, we have drawn on the 3D Open-vocabulary task proposed by OpenMask3D[29]. After performing 3D zero-shot instance segmentation, we incorporate a CLIP Semantic Addition module for RE0. It assigns the semantics of corresponding representative objects to the point cloud instances and facilitates the evaluation of our segmentation results. Furthermore, we have designed an evaluation metric which is specifically designed to directly evaluate zero-shot 3D instance segmentation.

In summary, our contributions are as follows:

- This paper proposes a novel framework called **RE0** to achieve zero-shot 3D instance segmentation. This method achieves unified consistency between 2D and 3D, as well as between 3D and 3D. The segmented results also conform to the semantic granularity.
- In order to facilitate the evaluation, this paper has also done the corresponding processing for the 3D open- vocabulary segmentation task, i.e., the RE0 framework can accomplish the 3D zero-shot open-vocabulary instance segmentation task. Besides, we design a new metric to demonstrate the performance advantages of our framework.
- Experiments conducted on ScanNet200 benchmark have shown that our method has achieved state-of-the-art (SOTA) standards among methods that perform zero-shot 3D instance segmentation. Furthermore, it has exhibited considerable performance in the 3D open-vocabulary instance segmentation task.

2 Related Work

2.1 3D semantic and instance segmentation.

Previous works[4, 14, 15, 16, 19, 20, 22, 30, 32, 40] have utilized large-scale 3D annotated data as supervision and employed deep learning with neural networks to achieve these objectives. On the ScanNet200 instance segmentation benchmark[3, 27], Mask3D achieved outstanding instance segmentation performance by utilizing Transformer-based segmentation networks[26]. TD3D achieved good results through a simple and fully data-driven approach from top to bottom[12]. LGround guided the learning of semantic category labels by anchoring 3D feature to the text embedding space of CLIP[24]. In addition, some methods based on superpoint[13, 28] represent the entire 3D scene by constructing superpoint graphs and employ graph neural networks to perform segmentation. Some 2D-Guided methods[37] utilize 2D segmentation models to achieve segmentation by projecting the camera poses to obtain 3D results.

2.2 Zero-shot and open-vocabulary 3D scene understanding.

Zero-shot 3D scene understanding is a relatively new research task with limited related studies. Currently, the main research still involves some pre-trained 3D models[18, 29]. However, with the development of 2D visual backbone models, the Segment Anything Model(SAM)[11], has made zero-shot object recognition possible. SAM is trained on the SA-1B dataset, acquiring extensive prior knowledge that enables effective segmentation of unfamiliar images without further training. Similarly, in indoor specific scenes, Cropformer can obtain more comprehensive 2D masks[21].

Recent studies are making efforts to apply these 2D segmentation models to 3D domain[6, 36, 37]. SAM3D performs segmentation by projecting 3D points onto 2D images as prompts for SAM, then back-projecting to obtain instance masks in 3D[37]. To address the consistency issue in SAM3D, SAMPro3D designs a filtering mechanism for masks filtering and fusion. SAM-Graph takes a graph neural network perspective, combine SAM to construct node and edge weights, and employs graph segmentation methods to segment scenes[36].

For open-vocabulary 3D scene understanding, OpenScene utilizes pixel-wise features extracted from posed images of scenes to obtain scene representations[18]. OpenMask3D has achieved open-vocabulary scene understanding in the 3D domain by combining CLIP features with pre-trained point cloud segmentation models[29]. OpenMask3D has also established a new benchmark on ScanNet200 dataset. Based on these, OpenIns3D has designed a module to generate images from point clouds cleverly eliminating the need for 2D image inputs[8]. Open3DIS also promotes research in open vocabulary scene understanding by aggregating 2D masks and mapping them to geometrically consistent point clouds[17].

3 Methodology

3.1 Problem Definition

The objective of point cloud semantic segmentation is to assign a label to each point in the point cloud that belongs to a specific category. Instance segmentation extends this further, as it not only provides the label for each point but also distinguishes between different individual instances. The Open-Vocabulary task requires us to be able to query the corresponding point cloud described by a given text prompt.

Specifically, our pipeline requires the input scene that includes: the point cloud P which contains N points, and the corresponding posed RGB-D frames of the point cloud. We denote the camera intrinsic as K and the number of RGB-D frames as T. For the certain frame t, its RGB image is denoted as F_t, depth image as D_t, and camera extrinsic as R_t. From the camera intrinsic, we can obtain the camera focal lengths (fx, fy), the principal point (cx, cy), and the radial distortion coefficients (bx, by).

We preprocess all frames of the RGB-D images using the 2D pre-trained model to extract all instance-level masks which are denoted as $\mathbf{M} = \{M_1, M_2, ..., M_T\}$. For the certain frame t, there are m_t 2D instance masks on the frame. On each mask map, each pixel is assigned a corresponding instance ID, which ranges from $[0, m_t]$. The instance ID of 0 is denoted as the meaningless background class.

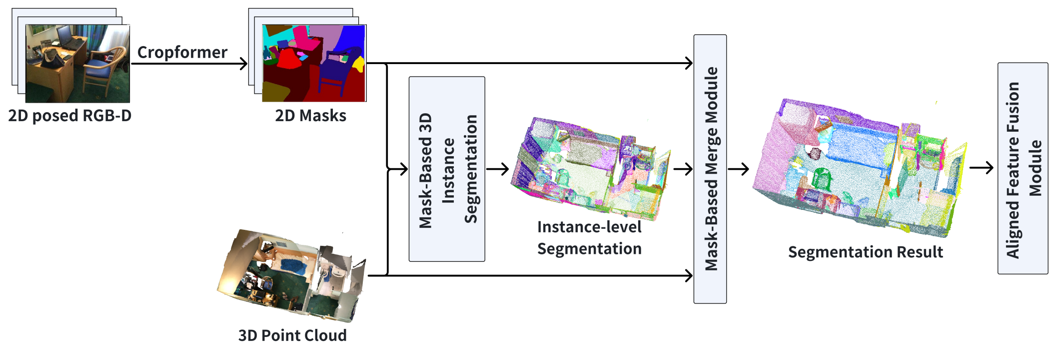

Figure 2: **Main pipeline of RE0**. We utilize the Cropformer to obtain 2D masks. For all frames, we project 3D point clouds on the masks and generate instance-level segmentation by Mask-Based 3D Instance Segmentation Module. Then, 2D masks and projection relationship are conducted to merge small-scale instances. Finally, we add CLIP semantic feature in Aligned Feature Fusion Module.

Notably, the 2D pre-trained model is replaceable. Since SAM[11] tends to segment indoor scenes with excessive fine granularity, we have chosen the Cropformer model[21], which provides a more complete segmentation results for indoor scenes.

3.2 Mask-based 3D Instance Segmentation

Projection. For a single frame F_t, we can establish a 3D-to-2D projection correspondence at this viewpoint. The points successfully projected onto the mask map are assigned the instance label of the corresponding pixel.

After projection, we obtain the segmentation state $S_t \in \mathbb{R}^N$ of the point cloud. Points projected onto the mask map receive the same instance label s as the corresponding pixel, where $s \in [1, m_t]$. Points that cannot be projected are labeled as 0, indicating an invalid label.

For the certain 3D point p_{3D}, in the designated camera coordinate system with intrinsic K and extrinsic R_t, its coordinate is (x, y, z). We can get the corresponding 2D pixel $p_{2D}(u, v)$ by following the equation below:

$$
\begin{aligned}
u &= \frac{(x - bx) \cdot fx}{z} + cx, \\
v &= \frac{(y - by) \cdot fy}{z} + cy,
\end{aligned} \tag{1}
$$

where, (fx, fy) is the camera focal lengths, (cx, cy) is the the principal point , and (bx, by) is the radial distortion coefficients. Note that not all points are valid projections. We will compare the estimated depth of the actual projections with the depth map D_t to filter out the valid points.

Alignment. After projection, we obtain the set of segmentation state $\mathbf{S} = \{S_1, S_2, ..., S_T\}$, where $S_t \in \mathbb{R}^N$. However, due to the lack of consistency in instance labels between different frames, the results in the instance labels between point cloud states not being aligned in 3D space. We propose a strategy for aligning two point cloud segmentation states S_{t_1} and S_{t_2}. The detailed algorithm is shown in Alg. 1.

Segmentation. In the Segmentation step, we set the final segmentation state as $S_{final} = \mathbf{0} \in \mathbb{R}^N$ firstly, and we iterate through all frames to add the final segmentation result. For the same point, we choose the instance label that appears most frequently. We denote the Alg. 1 as function $align(\cdot, \cdot)$, denote the operation of add segmentation state as function $add(\cdot, \cdot)$, the formula is followed:

$$
S_{final} = add(S_{final}, align(S_{final}, S_t)), t \in [1, T]. \tag{2}
$$

Algorithm 1 Aligning Strategy of Point Cloud Segmentation States

procedure ALIGN(S_{t_1}, S_{t_2}) ▷ Two segmentation states of the point cloud, $S_{t_1}, S_{t_2} \in \mathbb{R}^N$
  $s_{new} \leftarrow \max(S_{t_1}) + 1$
  for $s \leftarrow 1$ **to** $\max(S_{t_2})$ **do** ▷ Traverse all instance label in S_{t_1}
    $cluster_j \leftarrow S_{t_2}[S_{t_2} == s]$ ▷ Get point cluster in S_{t_2} with the same instance label s
    $cluster_i \leftarrow S_{t_1}[cluster_j]$ ▷ Get point cluster in S_{t_1} with the same index of $cluster_j$
    $cnt \leftarrow cluster_i.value\_count()$ ▷ Count the number of different label
    $max\_label, max\_num \leftarrow cnt[0]$ ▷ Get the label with the maximum count
    if $max\_num/len(cluster_j) > k_{align}$ **then**
      $S_{t_2}[S_{t_2} == s] \leftarrow max\_label$ ▷ Set the label to the aligned label
    else
      $S_{t_2}[S_{t_2} == s] \leftarrow s_{new}$ ▷ Set the label to the new label
      $s_{new} \leftarrow s_{new} + 1$ ▷ Update the new label
    end if
  end for
  return S_{t_2} ▷ The segmentation state S_{t_2} aligned with S_{t_1}
end procedure

3.3 Mask-based Merge Module

In Sec 3.2, we obtain a complete instance-level segmented point cloud state S_{final} which achieves instance consistency across 2D frames. However, due to the limitations of the projection perspective, the same mask may correspond to multiple local point clouds in 3D space. In this module, we achieve the generation of the segmented point cloud through Projection Merge.

Given two point cloud instance Ins_{i1}, Ins_{i2}, Mask-based Merge Module is used to determine whether or not these two instance should be merge based on the frame t.

First, we need to consider the efficacy of each point cloud instance.For the frame t and the labeled point cloud instance Ins_i with a point count of N^i, we set a projection score α. The formula is followed:

$$\alpha = \frac{V_t^i}{N^i}, \tag{3}$$

where V_t^i is the number of valid points which are projected on frame t by Ins_i. For Ins_i, if most points are valid($\alpha > k_{proj}$) on frame t, we consider Ins_i is a valid instance on frame t. Only when two instance is valid on frame t, we can continue to next step.

Although the instance Ins_i is valid on frame t, it may correspond to multiple different masks after projection. To measure this situation, we set the mask score β using the following formula:

$$\beta_t^i = \frac{\max_{j=1}^{m_t} c_i^j}{V_t^i} \tag{4}$$

where c_i^j denotes the number of valid points for Ins_i on the 2D mask j of frame t. We can also obtain the related mask label $Ins\_mask_i^t = max_{j=1}^{m_t} c_i^j$ of Ins_i. The core idea of Merge Module is that, if two point cloud instance can be merged, they should mostly be projected onto the same mask at frame t. Therefore, there are two conditions to merge Ins_{i1} and Ins_{i2}:

$$\begin{gathered} Ins\_mask_{i_1}^t = Ins\_mask_{i_2}^t \\ \beta_t^{i_1}, \beta_t^{i_2} > k_{mask} \end{gathered} \tag{5}$$

We follow the above operation to traverse all point cloud instance and frames to complete the merge stage.

3.4 Aligned Feature Fusion Module

Adding accurate features in a reasonable manner is a key step. For each point cloud instance Ins_i, we extract its CLIP semantic features for every frame. We reuse the projection mentioned in Sec. 3.2 and the projection score mentioned in Sec. 3.3. The whole module can be seen as Fig. 3.

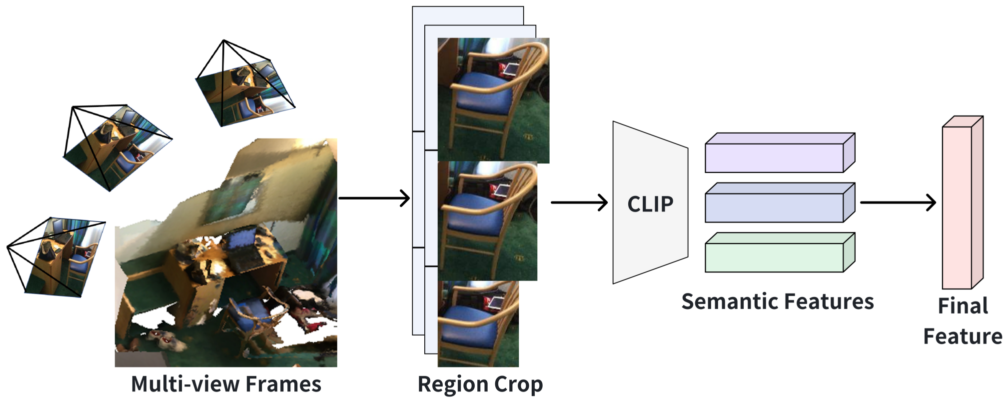

Figure 3: **Aligned Feature Fusion Module**. For selected instance Ins_i, we choose Top-K$_{scale}$ frames based on α and β. Then we crop the region three times and send them into CLIP to obtain semantic features. Finally, we calculate the average K$_{scale}$ $\times$ 3 features to generate the final feature of Ins_i.

If Ins_i is not a valid point cloud instance in frame t, the corresponding CLIP semantic features for that frame are set to **0**. Otherwise, through the distribution of the projected points, we can obtain the 2D mask area Rot_t^i. We feed Roi_t^i to CLIP to extract the semantic feature. We record the semantic features of all frames and obtain the Top-K$_{scale}$ CLIP semantic features with the largest weight proportions by sorting the weights w_t^i. The weights is calculated by following formula:

$$w_t^i = Softmax(\beta_t^i), \tag{6}$$

where β_t^i is the mask score for Ins_i on frame t. It is our contention that the more points on the corresponding mask area, the more accurate the semantics are represented.

In the context of the open-vocabulary task, it can be reasonably assumed that the instances have been segmented with a high degree of accuracy. Consequently, it is advisable to add CLIP semantic feature with precision. In this part, the Roi_t^i formula is followed.

$$Roi_t^i = [\min_{j=1}^{N_i} u_j + \lambda, \min_{j=1}^{N_i} v_j + \lambda, \max_{j=1}^{N_i} u_j - \lambda, \max_{j=1}^{N_i} v_j - \lambda], \tag{7}$$

where the N_i denotes the point count of instance Ins_i, (u, v) denotes the 2D points on frame t projected by instance Ins_i and λ is a hyper-parameter to control the scales of Roi_t^i. λ has 3 different scales to obtain multi-level semantic features.

4 Experiments

4.1 Experimental Details

4.1.1 Settings

We utilize the ScanNet200[25] dataset, which provides extensive annotations for 200 classes based on the RGB-D data of ScanNet[3]. The dataset offers an extremely challenging task for zero-shot 3D indoor scene segmentation. We validated our framework on the scannet200 validation set which contains 312 different indoor scenes. To expedite testing and conduct quantitative experimental analysis with previous zero-shot methods, we set the RGB-D frames to 240×320. The information about CLIP and Cropformer are provided in the supplementary material. Experimental results showcase that the entire framework's GPU usage does not exceed 10G, and that testing was conducted testing on a single RTX2080.

4.1.2 Metrics

Due to the particularity of zero-shot 3D instance segmentation, the segmented point cloud instances lack semantic labels. Consequently, traditional evaluation metrics are challenging to measure the accuracy of the work. As a result, we evaluate our framework by two different metrics.

For the first metric **mAP**, we follow the setting of OpenMask3D[29]. By matching the segmented point clouds with CLIP feature against the dataset's vocabulary, we select the label that is closest in semantic features to the point cloud instance as its label. This approach assesses the association from an open vocabulary of semantics to the closed set of class labels in the dataset. We compare our framework with OpenMask3D[29]. As shown in the supplementary material, our segmentation method segment the scene in more detail than GT, so we cannot segment some objects presented by ScanNet200. Following previous standard is unfair to us. Therefore, we adopted the method of calculating the mAP value of each scene separately and then averaging the scenes.

For the second metric **mAP**$_{GT}$, we follow the setting of SAMPro3D[36]. The segmented point cloud instances are compared with the ground truth points, and then a voting mechanism is used to select the most frequent ground truth label among the points in the segmented point cloud instances as the semantic label for this instance. Although the calculation of mAP$_{GT}$ is unfair, we believe it is a relatively reasonable method to describe the qualitative effects of zero-shot segmentation. Moreover, under this evaluation metric, we only compare with other zero-shot segmentation methods[36, 37].

More details about the evaluation metrics can be found in the supplementary material.

4.2 Experimental Results

4.2.1 Quantitative Results

As the Tab. 1 shows, for the open-vocabulary 3D instance segmentation on the ScanNet200 benchmark, a higher mAP indicates that the point clouds are more similar to the set of point clouds represented by the corresponding vocabulary in the validation set. Although our mAP is not good enough, our mAP$_{50\%}$ and mAP$_{25\%}$ have surpassed the OpenMask3D. The lack of control over the granularity of the zero-shot method makes it challenging for zero-shot methods to implement it as required for closed datasets.

Table 1: **Results(%) on ScanNet200**. The **bolder number** is the best and the <u>underline number</u> is the second best result. Methods with * means that this method validated on mAP$_{GT}$.

Method	mAP	mAP$_{50\%}$	mAP$_{25\%}$
OpenMask3D	**10.84**	<u>13.52</u>	<u>14.95</u>
Ours	<u>6.27</u>	**14.58**	**23.09**
SAM*	9.03	22.24	39.21
SAMPro3D*	<u>11.15</u>	<u>28.47</u>	<u>55.53</u>
Ours*	**15.76**	**37.16**	**61.22**

In our metric mAP$_{GT}$, our framework has achieved the state-of-the-art(SOTA) result on the ScanNet200 benchmark under zero-shot 3D segmentation methods. A higher mAP$_{GT}$ indicates that the segmented point clouds are more similar to the ground truth point clouds in terms of location. That is, at the positions where the ground truth point clouds exist, we have an equivalent amount of segmented instance-level point clouds present.

4.2.2 Qualitative Results

Zero-shot 3D instance segmentation. In Fig. 4, we present a qualitative result about zero-shot task. We compare GT, SAM3D and SAMPro3D. The highlighted visualization results help us prove that our method has stronger versatility compared to SAM3D and SAMPro3D. For specific objects or as a whole, corresponding point clouds can be segmented.

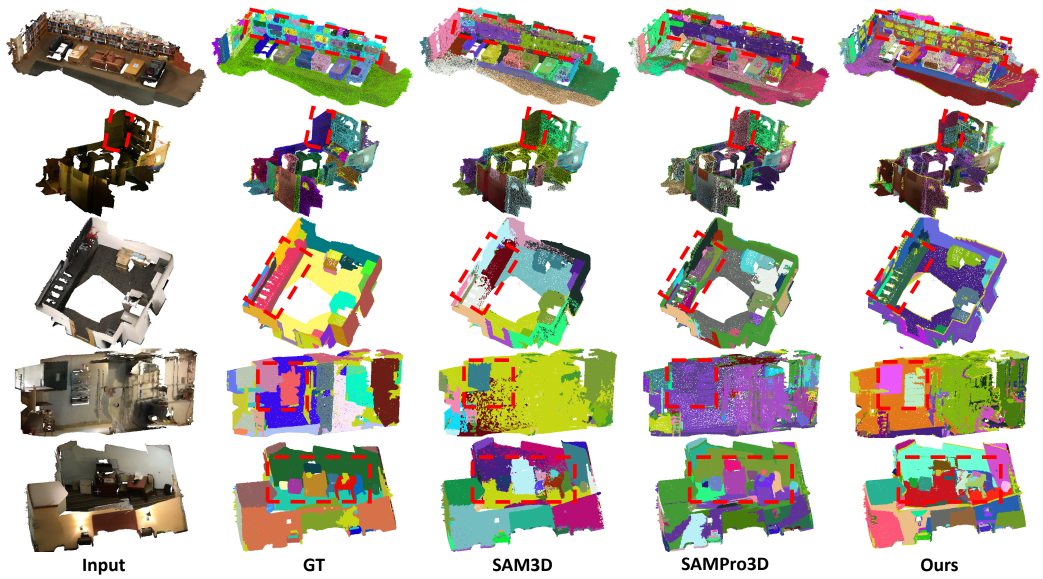

Figure 4: **The qualitative comparison of GT, SAM3D, SAMPro3D and Our Method.** The highlighted areas demonstrate the superiority of our method.

Open-vocabulary 3D instance segmentation. In Fig.5, we present a qualitative result about open-vocabulary task. RE0 is able to segment a corresponding object based on given query. It can be observed that RE0 can effectively segment the objects themselves for large-scale objects(like dresser, chair). Similarly, RE0 can also focus well on their geometric structures for small-scale objects(like light switch, toilet paper holder) .

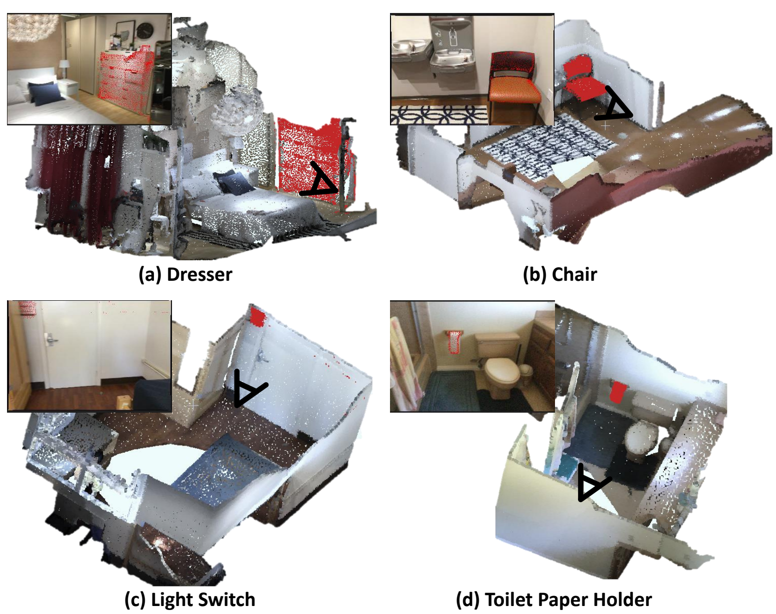

Figure 5: **Qualitative results of open-vocabulary tasks.** Our open-vocabulary instance segmentation is able to handle different queries. For each query, a corresponding 3D point cloud and a 2D image are provided. The segmented parts are marked in red.

4.3 Ablation Study

Ablation of Modules. In this work, we proposed two modules for 3D point cloud segmentation. Mask-based Merge Module(M3) is a interchangeable module after Mask-based Segmentation. As Fig. 6 shows that, the Mask-based Merge Module takes the responsibility for mergence of small-scale instances.

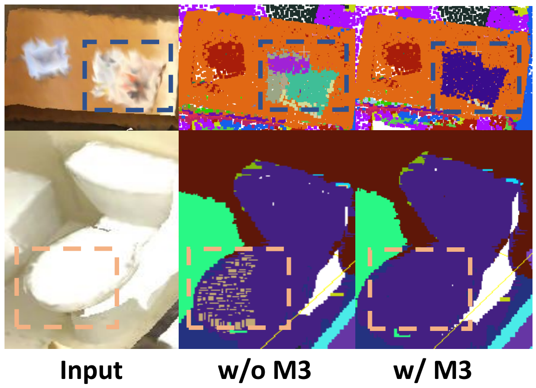

Figure 6: **Qualitative results of ablation studies.** The highlighted area has been effectively merged by the M3 module, filtering out fine noise.

Ablation of Hyperparameters. Due to the writing limitations, only the most important hyper-parameters related to projection are presented here. k_{proj} denotes that valid points after projection as a proportion of total points and k_{mask} proportion of valid points on a mask after projection. As the Tab. 2 shows that we decide the final $k_{proj} = 0.4$ and the final $k_{mask} = 0.6$.

Table 2: **Ablation study of hyperparameters.** mAP results(%) on randomly selected 20% of the 312 scenes in ScanNet200. The **bolder number** is the best and the <u>underline number</u> is the second best result.

k_{proj}	k_{mask}	mAP	mAP$_{50\%}$	mAP$_{25\%}$
0.3	0.5	5.49	13.11	21.30
0.3	0.7	<u>5.86</u>	13.92	22.47
0.4	**0.6**	5.68	**14.61**	**23.08**
0.4	0.8	**5.87**	<u>14.12</u>	<u>22.94</u>

5 Conclusion

Conclusion. In summary, we propose a novel framework **RE0** for 3D zero-shot open-vocabulary instance segmentation. The proposed framework utilizes the 2D mask extracted by Cropformer[21] and utilizes the projection relationship to achieve the mask-based segmentation. By combining with the 3D geometry position and CLIP[23] semantic feature, our approach can achieve the fusion and filtration of the 3D instances to generate the trustworthy 3D instance segmentation results.

Limitations and future works. The results of our approach are rely on the 2D pre-trained model. While we have selected the Cropformer[21] in our experiments, other 2D segmentation models such as SAM[11], MobileSAM[38], and EfficientSAM[34] can also be connected to our framework easily. Furthermore, in some scenes, we believe that the current segmentation granularity is not very satisfactory. For example, it is difficult to say whether the keycaps on the keyboard should be separated into instances or not. In the future, the potential for zero-shot segmentation to create a method like Garfiled[10] that can freely control the scale represents an exciting avenue for further research.

References

[1] Yu Cao, Yancheng Wang, Yifei Xue, Huiqing Zhang, and Yizhen Lao. Fec: fast euclidean clustering for point cloud segmentation. *Drones*, 6(11):325, 2022.

[2] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019.

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[4] Xin Deng, WenYu Zhang, Qing Ding, and XinMing Zhang. Pointvector: a vector representation in point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9455–9465, 2023.

[5] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2023.

[6] Haoyu Guo, He Zhu, Sida Peng, Yuang Wang, Yujun Shen, Ruizhen Hu, and Xiaowei Zhou. Sam-guided graph cut for 3d instance segmentation. *arXiv preprint arXiv:2312.08372*, 2023.

[7] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021.

[8] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. *arXiv preprint arXiv:2309.00616*, 2023.

[9] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020.

[10] Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. *arXiv preprint arXiv:2401.09419*, 2024.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[12] Maksim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Top-down beats bottom-up in 3d instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3566–3574, 2024.

[13] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018.

[14] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[16] Yan Lu and Christopher Rasmussen. Simplified markov random fields for efficient semantic labeling of 3d point clouds. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2690–2697. IEEE, 2012.

[17] Phuc DA Nguyen, Tuan Duc Ngo, Chuang Gan, Evangelos Kalogerakis, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. *arXiv preprint arXiv:2312.10671*, 2023.

[18] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023.

[19] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[20] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[21] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022.

[22] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35:23192–23204, 2022.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[24] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022.

[25] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022.

[26] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023.

[27] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663*, 2022.

[28] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2393–2401, 2023.

[29] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

[30] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017.

[31] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019.

[32] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022.

[33] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020.

[34] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *arXiv preprint arXiv:2312.00863*, 2023.

[35] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021.

[36] Mutian Xu, Xingyilang Yin, Lingteng Qiu, Yang Liu, Xin Tong, and Xiaoguang Han. Sampro3d: Locating sam prompts in 3d for zero-shot scene segmentation. *arXiv preprint arXiv:2311.17707*, 2023.

[37] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023.

[38] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[39] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[40] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[41] Chengjie Zong, Hao Wang, et al. An improved 3d point cloud instance segmentation method for overhead catenary height detection. *Computers & electrical engineering*, 98:107685, 2022.

A Appendix / supplemental material

A.1 More Information.

The discussion about the metrics.

We want to discuss the issue of evaluation metrics for zero-shot 3D instance segmentation.

Since the inception of the SAM3D method, evaluating these approaches fairly has become a challenging task. Traditional evaluation methods are not suitable for this task, because we only obtain segmented point clouds without knowing their semantic labels. SAM3D does not address this issue. The evaluation metric mIoU in SAMPro3D allocates scores based on the intersection between the segmented point cloud and the ground truth (GT), which tends to yield high scores when the point cloud scene is fragmented. This is due to the fact that the intersection of the fragmented point clouds with the complete GT is always the fragmented point cloud itself, which results in the segmentation of excessively fragmented data sets being assigned inflated scores.

We followed the idea of SAMPro3D and designed a corresponding mAP$_{GT}$ to solve this issue. It also allocates labels based on the intersection between the segmented point cloud and GT. Because the ScanNet200 benchmark calculates mAP by considering the respective positional intersections, it partially mitigates the problem of fragmented point cloud segmentation receiving higher scores.

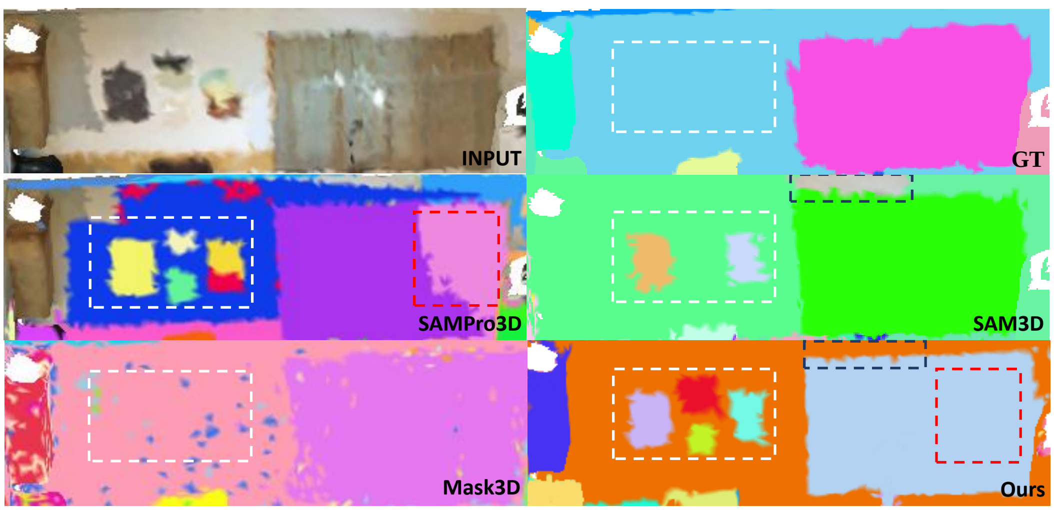

Figure 7: Comparison on scene0000_00.

It is evident that the core issue lies in the process of attaching semantics to segmented point cloud instances. If semantics can be attached to each point cloud instance, the problem of fair quantitative evaluation of zero-shot segmentation can be addressed. The recently introduced 3D open-vocabulary task by OpenMask3D seems to align well with this objective.

However, we found that this approach is not entirely fair either in practice. This is because the vocabulary provided by ScanNet200 does not cover all terms and there may be ambiguity for the same object. This is not a problem for training-based methods because they are specifically trained on the dataset, so the segmented shapes tend to correspond more closely to the evaluation metric categories. In contrast, zero-shot methods may have disadvantages because they are better suited for showcasing fine-grained results, and their overall segmentation performance may be comparatively weaker. Additionally, some fine-grained objects are not annotated in the dataset, which causes zero-shot methods to lose their inherent advantages.

To address this issue, we modified the traditional category-based mAP to a scene quantity-based mAP, which helps to alleviate the problem to some extent.

The settings of experiments.

Table 3: The settings of experiments.

Devices/Hyper-parameters	Versions/Numbers
k_{scale}	3
k_{proj}	0.4
k_{mask}	0.6
λ	0.1, 0.2, 0.3
Confidence of Cropformer	0.25
Jump Frame	10
2D RGB-D Scale	240×320
GPU Device	GTX3090 24G

408 ### A.2 More Experiments.

409 Some experiments have followed and more experiments are shown in our anonymous project page.

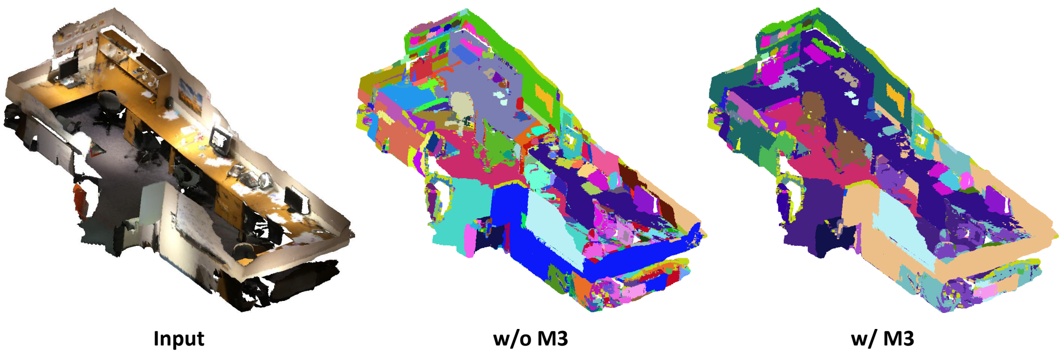

Figure 8: Ablation on Scene0131_00.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim our contributions and scope in the last paragraph of introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the conclusion chapter.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper is mainly discuss the experiments, and does not include theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We open our code with the anonymous url in abstract, we display our experiment setting in Sec. 4 and hyperparameters are presented in supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we open our code with the anonymous url in abstract and our data is based on ScanNet200 which is an open-source dataset.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We display our experiment setting in Sec. 4 and hyperparameters are presented in supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We discuss the metrics which may bring errors on Supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss the settings in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, this paper conducted with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our research task is a basic 3D segmentation task which has no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not publish any new models and we just use the previous models to solve 3D instance segmentation task. So this paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We claim the previous works on the position where we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release our code on an anonymized URL.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.