OpenMask3D++: Scaling Open-Vocabulary 3D Instance Segmentation to Fine-Grained Semantic Annotations

Michael Neumayr[1] Michal Stary[1] Simon Pannek[1] Daniel Korth[1]
{michael.neumayr, michal.stary, simon.pannek, daniel.korth}@tum.de
[1]Technical University of Munich (TUM)

Abstract

Achieving thorough 3D scene comprehension is crucial for emerging technologies like augmented reality and robotics. In this study, we enhance OpenMask3D, a leading open vocabulary 3D instance segmentation framework. Our approach involves diverse view selection, outlier removal using DBSCAN, and upgraded multimodal embeddings (SigLIP and DINOv2). While our quantitative results exhibit modest improvements, we propose a future exploration of alternative backbones, such as Segment3D, or finetuning Mask3D on the ScanNet++ dataset. Our contributions aim to elevate the capabilities of OpenMask3D in navigating dynamic and diverse real-world environments. The code is available at https://github.com/danielkorth/openmask3d/

1. Introduction

The proliferation of augmented reality and robotics has highlighted the importance of achieving comprehensive 3D scene understanding, pivotal for applications navigating complex and dynamic environments. This understanding encompasses the capability to predict an object mask and its semantic category for any object, even those not included in the model's training dataset. Such advanced comprehension enhances application functionality and user experience by allowing for the recognition of a wide array of real-world objects not limited to the training classes.

In our study, we build upon OpenMask3D [7], a state-of-the-art approach to open vocabulary instance segmentation (see Figure 1 for description). We aim to accomplish two main goals: extended evaluation to simulate truly open-world situations and improvements to the OpenMask3D pipeline. Firstly, we evaluate the performance of OpenMask3D in a more diverse and open environment. Following a shortcoming noted by the authors [1] – a limitation in what top k views can see; we secondly try to alleviate this problem by introducing several enhancements to the 2D segmentation and semantic matching pipelines to leverage more diverse views from different angles and remove any outliers that result from the view diversification. Thirdly, we upgrade the CLIP feature extraction [4]. We switch to the state-of-the-art SigLIP [9] model for text-to-image similarity. We also leverage DINOv2 [3] encodings to allow for querying the scene via images instead of text.

We present qualitative and quantitative analyses of our modifications, benchmarking against the original OpenMask3D setup using ScanNet200 and against the more comprehensive ScanNet++ dataset [5, 8].

We refrain from finetuning the Mask3D [6] checkpoints, which form the foundation for OpenMask3D's class-agnostic 3D instance mask proposals, on the ScanNet++ dataset. This decision allows us to simulate a truly open vocabulary scene understanding assessment of OpenMask3D when leveraging the unseen ScanNet++ dataset. For a more detailed analysis of 3D mask proposals on ScanNet++, we refer to the concurrent work Segment3D [2], which offers a detailed comparison of the latest 3D mask proposal networks, including Mask3D. Our contributions are as follows:

- More diverse view selection for improved scene visibility
- View outlier removal with DBSCAN [1] on image embeddings
- Upgrade to latest text to image and image to image embedding methods: SigLIP [9], and DINOv2 [3]

2. Related Work

2.1. Instance Segmentation

Mask3D [6] leverages a transformer architecture for 3D mask proposals and semantic labeling, achieving significant performance on the ScanNet200 benchmark. Its limitation, however, lies in its adherence to a closed vocabulary dictated by dataset annotations.

Segment3D [2] concurrently explores a new direction by utilizing automatically generated labels to overcome the re-

[1]"since the per-mask features originate from images, they can only encode scene context visible in the camera frustum". [7]

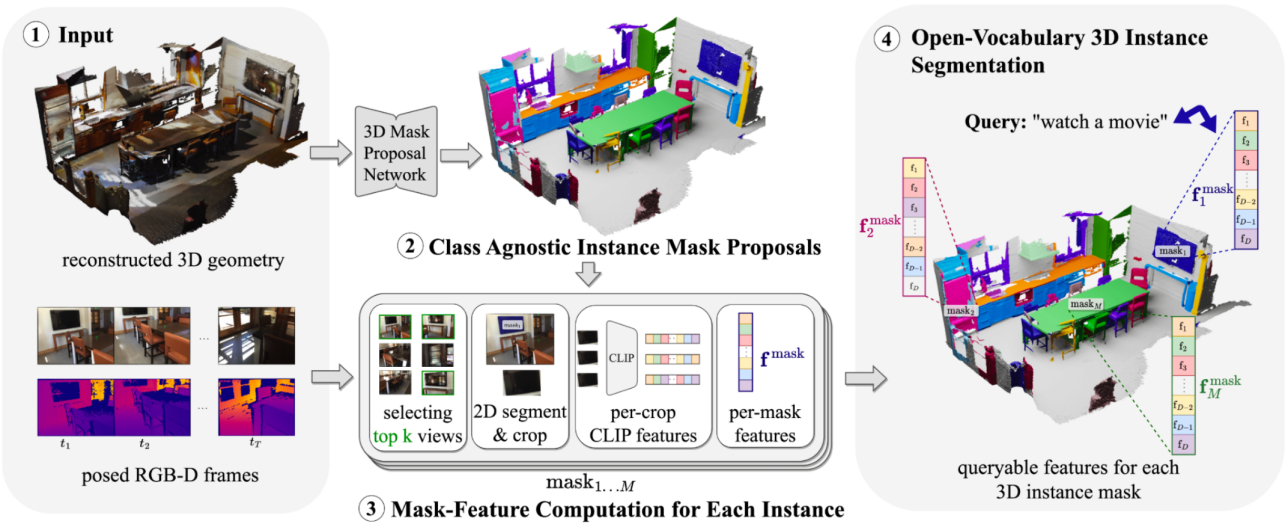

Figure 1. The original pipeline of OpenMask3D. (1) The input comprises a point cloud and a sequence of posed RGB-D frames. (2) The point cloud is processed by an adjusted Mask3D pipeline to extract class-agnostic 3D mask proposals. (3) Each mask proposal is projected to input views, and the best views are selected. The particular region of the object is obtained by running SAM segmentation conditioned on the projected points, this region is cropped out from the view. CLIP model computes a semantic embedding for each crop. Finally, the embeddings for each instance are averaged. (4) To segment 3D instances in an open-vocabulary manner, the CLIP embedding of arbitrary text query is computed and compared by cosine similarity with each instance's embedding.

liance on manual, costly ground-truth annotations tied to predefined object classes. This shift towards automation and generalization significantly improves understanding of truly open vocabulary scenes.

2.2. Foundation Models

Multiple foundation models for vision tasks have been developed and made publicly available in recent years. Operating in the pure image domain, DINOv2 [3] is a self-supervised pre-trained model that can be used as a feature extractor for further finetuning on downstream tasks. Contrary to a classification-based pretraining schema, the self-distillation on image crops used in DINOv2 has been shown to lead to emerging properties in extracted features, such as understanding object masks, thus making it more suitable for zero-shot similarity queries.

CLIP [4] was a pioneering work from the multimodal family that leveraged large-scale pretraining to jointly learn text and image embeddings. By computing the embedding similarity such joint representations allow for zero-shot, open-vocabulary, recognition.

SigLIP [9] is a follow-up work that replaces the softmax loss with sigmoid and introduces learnable bias to mitigate the issues in initial optimization caused by the dominance of negative samples. It has been reported to surpass CLIP on multiple benchmarks with identical backbones.

3. Method

In this section we motivate and explain our proposed modifications to the OpenMask3D [7] pipeline.

3.1. View Diversification

First, we explored a significant enhancement to the original view selection methodology of the paper. The goal of the view selection is to select the top k best-suited video image frames for each mask. We can then use those images to extract all relevant semantic details for that particular mask.

Initially, the process was geared towards selecting views based on the quantity of points from the original scan visible in a given perspective. This method aimed to identify and isolate the optimal visual representations of an object by prioritizing views where a substantial portion of the object's points were discernible, thus minimizing occlusions as verified through depth maps. The original researchers verified through ablation that a selection of five views ($k = 5$) yielded the most effective results.

However, our investigation identified a critical limitation in this approach, particularly in its ability to comprehensively capture the semantic nuances of particular objects. For instance, when considering a flat object like a door, which may possess crucial features on its narrow side, we noticed that the prevailing method of view selection would invariably overlook relevant perspectives from the side due to the minimal point visibility from these angles. To address

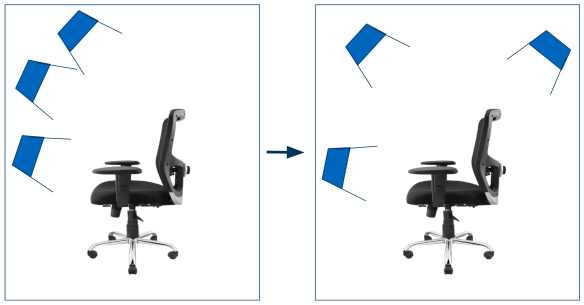

Figure 2. Our approach for view diversification replaces multiple views from a similar perspective with more diverse views, prioritizing angular disparity over maximizing object visibility.

this shortfall, we introduce the concept of view diversification.

Our proposed methodology augments the traditional selection process by not solely relying on point visibility. Instead, we commence by pre-selecting a broader set of candidate views based on visibility ($n = 10$) and subsequently refine this selection to identify the k most diverse views based on view angle. This approach ensures a more holistic representation of an object's semantics by incorporating views that, while not maximizing point visibility, offer critical spatial and contextual insights. To balance the importance of visibility with our aim for diversification, the most visible view is always included as the primary selection, followed by the subsequent $k-1$ views that exhibit the highest angular disparity from this initial choice.

3.2. Outlier Removal

While diversifying views based on angular disparity offers a more comprehensive semantic representation of objects, it simultaneously presents the risk of including perspectives that minimally showcase the object. This issue is particularly pronounced in scenarios with a limited selection of frames, leading to the potential inclusion of angles that offer bad visibility of the object of interest.

Additionally, our reliance on box crops to isolate object views introduces another layer of complexity. Depending on the chosen angle, these crops may inadvertently encapsulate additional semantic elements alongside the target object. For example, a view of a chair from an elevated angle might also capture a significant portion of the floor, thereby diluting the focus on the chair itself. When using those images for extracting features for a given mask, this might lead to false positive on queries relating to the additional captures semantic elements.

To mitigate these issues and ensure the integrity of our semantic analysis, we have integrated a spatial clustering mechanism using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [1]. DBSCAN excels in identifying and segregating outliers within a dataset by analyzing the spatial proximity of data points (or, in our case, images). Through this process, we can effectively isolate and eliminate images that, due to their distant placement from the cluster of more relevant views, are deemed outliers. We hope that these images are typically characterized as outliers by their failure to adequately capture the intended semantics of the object due to their marginal visibility or the inclusion of unrelated semantic content. By applying DBSCAN for outlier removal, we refine our model to include only those views that significantly contribute to a holistic and accurate semantic representation.

We also consider a more naive approach by pooling the features for each image via a max pooling operation instead of mean, as proposed by OpenMask3D [7]. Intuitively, this allows to extract the most prominent features out of the different views that were captured and preprocessed, which might lead to sharper similarities between images and text.

3.3. Multimodal Embeddings

In the original work, raw CLIP [4] method with `ViT-L/14@336px` backbone was used for computing the embedding for each crop. Following the recent research, we switched to a different model SigLIP [9] which is state-of-the-art (see Section 2.2 for details). In particular, we used the SigLIP model with `ViT-SO400M@384px` backbone.

3.4. Image Queries

Next, we expand our model's capabilities beyond text-based queries to include image-based queries, addressing a significant limitation in the existing framework. Previously, the model relied solely on text descriptions to identify and match objects within its database. However, accurately describing certain objects through text alone can be challenging, especially when specific visual features are difficult to verbalize. This limitation becomes apparent in use cases where the user must describe unique or complex objects, where textual descriptions may not capture the entirety of the object's characteristics.

To overcome this challenge and enhance the model's versatility, we have incorporated DINOv2, a model that leverages image similarity during its training process [3]. By utilizing DINOv2 to extract image features for every mask, our system now supports direct image queries, enabling users to search for objects by providing an image instead of a textual description.

4. Experimental Setup

Datasets. We assess model performance on two datasets for indoor scene segmentation, ScanNet200 and ScanNet++ [5, 8].

ScanNet200	AP	AP50%	AP25%
Base	11.1	13.5	13.8
Base + max aggr.	**12.5**	**15.4**	**15.8**
Base + diverse views	9.1	10.8	11.3
Base + outlier removal	11.0	13.8	14.4
Base + SigLIP emb.	11.0	13.8	14.4

Table 1. Comparison of modifications applied *independently*. Top 1 AP, AP25% and AP50% evaluated on ScanNet200 benchmark of first 40 scenes of validation split.

ScanNet++	AP	AP50%	AP25%
Base	7.6	11.2	14.6
Base + max aggr.	7.4	11.0	13.8
Base + diverse views	7.8	11.6	14.4
Base + outlier removal	8.1	11.6	14.8
Base + SigLIP emb.	**11.2**	**16.3**	**20.1**

Table 2. Comparison of modifications applied *independently*. Top 5 AP, AP25% and AP50% evaluated on the first 40 scenes of ScanNet++ validation split.

ScanNet200 extends the original ScanNet dataset to include 200 semantic class annotations, enhancing object recognition diversity in over 1,500 indoor scenes. With a pre-trained Mask3D checkpoint, it serves as the baseline to validate or reject our modifications.

ScanNet++ elevates the challenge with 460 scenes and over 1.6k distinct classes, providing a robust platform for open vocabulary instance segmentation. It features high-resolution 3D scans and extensive annotations of varied-sized objects, testing our model's adaptability to complex and previously unseen environments. By comparing the model's performance on ScanNet200 and ScanNet++, we test its generalization capability and the effectiveness of the pre-trained Mask3D component in navigating the complexity of ScanNet++'s diverse indoor settings which it has never seen before. Due to more than 8-fold increase in the number of classes of ScanNet++, we report top 5 matches of classes.

Evaluation Metrics. To evaluate open vocabulary 3D instance segmentation, we compute the Average Precision (AP) scores at IoU thresholds of 25% (AP25%), 50% (AP50%), and average across thresholds from $50\% - 95\%$ in 5% increments (AP). This aligns with standard practices within the field, ensuring a comprehensive evaluation of segmentation precision relative to ground-truth masks. To navigate the trade-off of computational constraints and meaningful comparison, we select the first 40 scenes of the validation set of each dataset. For both experiments, we use Mask3D Instance Segmentation trained on ScanNet200.

4.1. Results

Quantitatively, we observe that switching from mean to max aggregation of the 5 most visible views improves performance by over 10% for ScanNet200, a non-trivial increase. However, the other proposed modifications do not improve the quantitative evaluation significantly - in the case of diversifying the views, it even decreases performance by quite a bit.

In the case of ScanNet++, the SigLIP embeddings yield the biggest performance increase, improving the average precision by almost 50% to 11.2% AP. The other modifications, however, do not lead to any significant performance increases.

One major limitation for operating on ScanNet++ dataset is the 3D mask proposal backbone Mask3D. We have evaluated this backbone pre-trained on ScanNet200, but during experiments, we observed that the proposed 3D masks sometimes overlap and morph object parts together that should not be. This, of course, sets an upper bound to the performance of our modified 2D segmentation and matching pipeline.

We also explored combining the proposed modifications to enhance the overall performance. However, this did not lead not lead to any significant or consistent performance increases over the baseline method.

5. Conclusion

In this report we explored shortcomings and modifications to the current OpenMask3D pipeline. We focused on the mask feature computation for each instance, diversifying selected views, making them more robust to outliers and replacing the multimodal embedding model CLIP. While conceptually sound, our modifications did not yield consistent performance improvements over OpenMask3D across multiple settings. The overall approach especially struggled to transfer to the more complex ScanNet++ dataset with an order of magnitude more instance classes. One of the main limitations is the 3D mask proposal backbone Mask3D, which was not able to accurately segment instances.

We propose two alternatives for future work. Firstly, one can follow the open vocabulary philosophy and switch to the concurrent backbone Segment3D that does not need for closed vocabulary annotations but outperforms Mask3D. Secondly, one can finetune Mask3D trained on ScanNet with the more complex ScanNet++ dataset. Although, this needs closed vocabulary annotations, it is still worth testing whether fine-tuning the instance proposal to new scenes will make the instance masks more robust.

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996. 1, 3

[2] Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. *arXiv*, 2023. 1

[3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1, 2, 3

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2 2021. 1, 2, 3

[5] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 3

[6] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. 2023. 1

[7] 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. 1, 2, 3

[8] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 1, 3

[9] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 1, 2, 3