DCSEG: Decoupled 3D Open-Set Segmentation using Gaussian Splatting

Luis Wiedmann* Luca Wiehe* David Rozenberszki
Technical University of Munich
{luis.wiedmann, luca.wiehe, david.rozenberszki}@tum.de

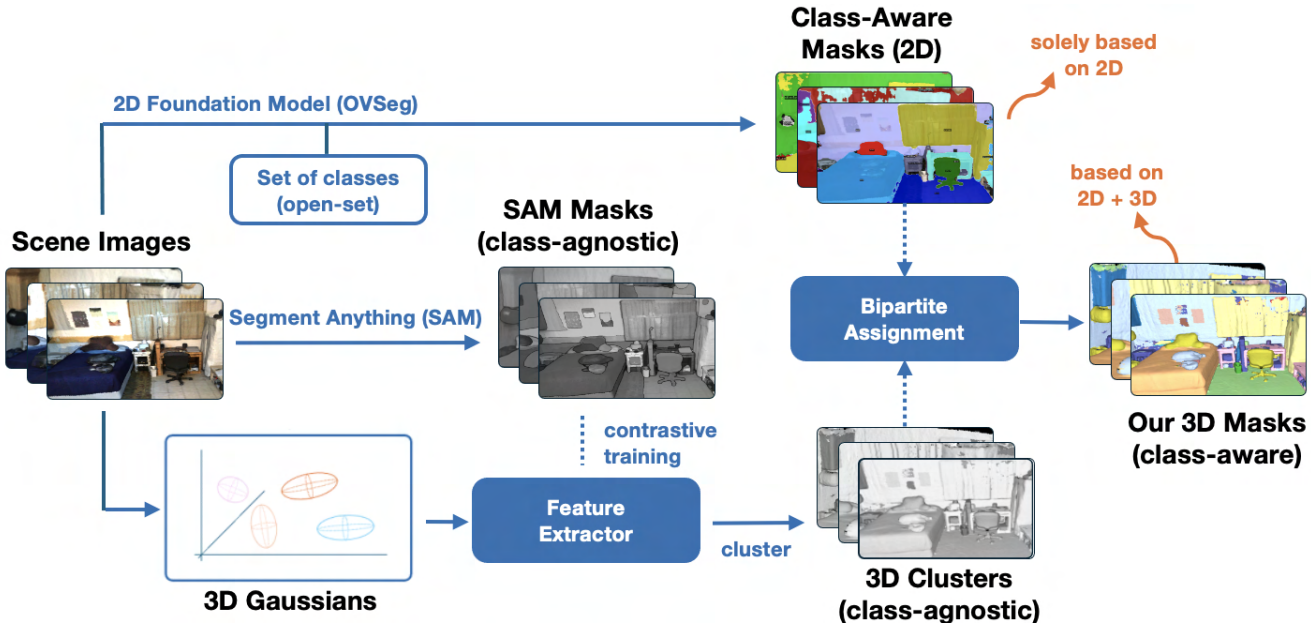

Figure 1. **Decoupling the semantic segmentation pipeline.** We present DCSEG, a holistic 3D reconstruction and scene understanding method. In the core of our method we leverage pretrained 2D foundation models such as SAM [1] to recognize uniform semantic concepts in 2D images of 3D scenes, and use these predicted masks as contrastive optimization targets from multi-view images to class-agnostic 3D instances and object parts. These features are then used to cluster the Gaussians in 3D with hierarchical clustering methods. Simultaneously, we use a 2D semantic segmentation network to obtain class-aware masks and aggregate class-agnostic parts into meaningful semantic instances. As a results we obtain 2D/3D instance and semantic segmentation on synthetic and real world scenes.

Abstract

Open-set 3D segmentation represents a major point of interest for multiple downstream robotics and augmented/virtual reality applications. Recent advances introduce 3D Gaussian Splatting as a computationally efficient representation of the underlying scene. They enable the rendering of novel views while achieving real-time display rates and matching the quality of computationally far more expensive methods. We present a decoupled 3D segmentation pipeline to ensure modularity and adaptability to novel 3D representations and semantic segmentation foundation models. The pipeline proposes class-agnostic masks based on a 3D reconstruction of the scene. Given the resulting class-agnostic masks, we use a class-aware 2D foundation model to add class annotations to the 3D masks. We test this pipeline with 3D Gaussian Splatting and different 2D segmentation models and achieve better performance than more tailored approaches while also significantly increasing the modularity.

1. Introduction

Examples of common 3D representations include Point Clouds, Voxel Grids, Meshes, and Neural Radiance Fields (NeRFs) [2]. NeRFs have become particularly prominent in the scientific community in recent years. This approach does not utilize explicit geometry to model the scene; instead, the scene is represented by an MLP. There have been

several adaptations of NeRFs, but the basic principle remains the same. 3D Gaussian Splatting [3] is a fundamentally different approach compared to NeRFs. Instead of representing the scene implicitly in an MLP, Gaussian Splatting uses 3D Gaussians as an explicit representation. Given a set of images, we can initialize and optimize a set of 3D Gaussians for every scene. The key advantage is that this explicit representation doesn't need volumetric rendering to project the 3D reconstruction onto 2D space. Instead, we can directly rasterize our explicit representation onto the image plane to render an image much faster.

NeRFs and Gaussian Splatting have already been used as an underlying representation for scene understanding. NeRF-based approaches include LeRF, Panoptic NeRF, and OpenNeRF [4–6]. OpenNeRF is built on the concepts of these latest methods. It speeds up standard pipelines by directly integrating the segmentation head onto the NeRF used to represent the scene. This integration enables rendering and segmentation in a single pass, and when combined with NeRF's novel view synthesis capabilities, it achieves dominant performance.

Recent research efforts, including ZegFormer [7] and DeOP [8], have focused on decoupling the mask proposal from the mask classification in the context of 2D image analysis. This decoupling is particularly advantageous in 3D space, where the diversity of 3D representations necessitates a more generalizable approach. The need for flexibility in 3D representations has led to a line of research that similarly separates mask proposal from mask classification, particularly in the realm of 3D instance segmentation, as demonstrated by works like OpenMask3D [9] and SAI3D [10]. However, it is worth noting that these approaches do not utilize dense geometry but point clouds to perform segmentation, making them less adaptable for general use compared to 3D Gaussians. There are also some works that utilize 3D Gaussian Splatting as a basis for 3D semantic segmentation [11–13]. However, compared to our approach, they do not decouple mask proposal and mask classification, making them unadaptable to recent advances in mask classification architectures.

Motivated by this, we add modularity to class-aware segmentation in 3D by decoupling the 3D semantic segmentation pipeline. Our contributions can be summarized as follows:

- We utilize 3D Gaussian Splatting as an underlying representation for class-aware open-vocabulary semantic scene segmentation
- We demonstrate that Gaussian Splatting can outperform comparable SOTA NeRF-based architectures for 3D semantic segmentation while being more modular
- We present an architecture that can identify 3D instances and event parts without needing to train an instance-segmentation network

2. Related Work

3D Gaussian Splatting. Gaussian Splatting uses 3D Gaussians as an explicit geometry to represent a 3D scene. Given classical structure from motion (SfM) approaches, they take a sparse point cloud that is used to initialize a set of 3D Gaussians $\mathcal{G} = \{\mathbf{g_i}\}_{i=1..k}$ where k is the total number of Gaussians. These 3D Gaussians can be efficiently rasterized to a given 2D image frame. This allows comparing reconstructed images to a set of ground truth images $\mathcal{I}$ without requiring an expensive neural rendering approach as NeRFs do.

An important feature of 3D Gaussian splatting is that the underlying rasterization approach is differentiable, enabling direct optimization of the 3D Gaussians based on the aforementioned comparison. This gradient-driven optimization is crucial for fine-tuning the Gaussians to reconstruct the scene accurately. Another advantage is the flexibility in the representation process: if a Gaussian is too large and causes a loss of detail, it can be split into smaller Gaussians for finer granularity. Conversely, if a single Gaussian is not enough to represent a structure, it can be duplicated to better capture the scene's intricacies. Thus, the adoption of 3D Gaussians in our semantic segmentation pipeline is well-motivated by significant efficiency gains and flexibility.

3D Segmentation. Given a 3D reconstruction, semantic segmentation remains a challenging task (in particular, in an open-set fashion). Since there is way less training data available in 3D than in 2D, it is difficult to train an accurate segmentation network directly on 3D data. A common approach is integrating 2D foundation models into the training pipeline to leverage their more diverse and accurate training data set. SAGA uses this idea to segment 3D Gaussians. In essence, features are attached to 3D Gaussians for segmentation in 3D space. These features are then iteratively improved in a contrastive fashion utilizing a loss contrasting SAM masks [1] and masks from 3D rendered onto the 2D image plane.

Open Vocabulary Image Segmentation. Recent advances in Vision-Language Models (e.g., CLIP [14]) achieve impressive open vocabulary image-level classification performance. However, these models are unable to perform accurate pixel-wise segmentation. OpenSeg [15] addresses these shortcomings by adding a layer that performs visual grouping of various concepts in an image before performing the final segmentation. OVSeg [16] tackles the same issue by identifying the pre-trained CLIP model as the segmentation bottleneck before fine-tuning this on masked images. Both approaches provide an open-vocabulary image segmentation model that can provide class-aware segmentation masks.

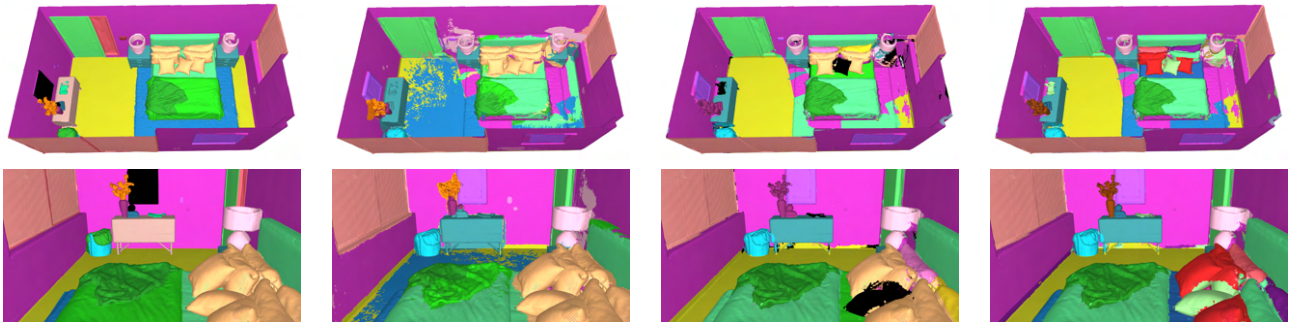

Figure 2. **Segmentation results of our method compared to the ground truth and OpenNeRF.** Our segmentation masks can detect boundaries more accurately e.g. the blanket/pillows or the wall behind the bed-lamps. Large uniform areas, such as the floor, can be detected with significantly less noise. Switching between Openseg and OVSeg can be done without retraining and demonstrates adaptability with respect to foundation models.

Decoupled 3D Instance Segmentation. OpenMask3D [9] is an architecture geared toward 3D semantic instance segmentation. It operates by first proposing 3D masks, then extracting features from these masks, and finally assigning a class label to each of the masks. As a 3D representation, it processes high-density point clouds to propose masks based on them. However, point clouds inherently do not represent volumes and surfaces continuously, making them less suitable for applications that require detailed volumetric and surface information. Follow-up work [10, 17, 18] has not addressed the limitations associated with the discontinuous representation of volumes and surfaces in point clouds.

3. Method

Each of the approaches above faces at least one of the following weaknesses: the inability to perform class-aware segmentation, the inability to incorporate (dense) 3D information, the inability to distinguish instances or the coupling between semantic segmentation and 3D reconstruction. We aim to compensate for all these weaknesses and develop a robust and modular approach to perform 3D open-set segmentation in a class-aware fashion. We aim to achieve this in a decoupled fashion, allowing us to interchange the underlying 3D Representation and the semantic feature extraction with any other pipeline that can provide class-agnostic 3D clustering and class-aware 2D segmentation. Our pipeline consists of two essential stages:

1. Propose class-agnostic segmentation masks that are based on a 3D representation
2. Classify these class-agnostic masks by establishing correspondence with multiple-view class-aware 2D segmentation masks

Stage 1: Class-Agnostic Mask Proposal. Given a set $\mathcal{I}$ of posed RGB-D input images of a 3D scene, we start by obtaining a 3D reconstruction using Gaussian Splatting. This results in a set of k Gaussians $\mathcal{G} = \{\mathbf{g}_i\}_{i=1..k}$ representing the scene. Inspired by SAGA [21], we then use a scale-aware contrastive learning strategy to attach a set of Gaussian affinity features $\mathcal{F} = \{\mathbf{f_{g_i}} \mid \mathbf{f_{g_i}} \in \mathbb{R}^n\}_{i=1..k}$ to every Gaussian. Let $\mathbf{p_1}$ and $\mathbf{p_2}$ be two corresponding pixels from a given image $\mathbf{I} \in \mathcal{I}$, then the loss function is given by:

$$\mathcal{L} = \sum_{\mathbf{p_1},\mathbf{p_2}} \mathcal{L}_{corr}(s, \mathbf{p_1}, \mathbf{p_2}) + \frac{1}{h \cdot w} \sum_{\mathbf{p}} \mathcal{L}_{norm}(\mathbf{p})$$

This loss contains two main components: A correspondence distillation loss $\mathcal{L}_{corr}$ and a feature normalization loss $\mathcal{L}_{norm}$. The correspondence distillation loss resembles the optimization target that two pixels $\mathbf{p_1}, \mathbf{p_2}$ from a given image $\mathbf{I} \in \mathcal{I}$ should have similar features if and only if they belong to the same SAM mask. Note that these features are conditioned on a scale hyperparameter s. This hyperparameter is geared towards preserving SAM's granularity. This allows us to adjust the level of detail that is supposed to be captured without the need to rerun the feature extraction. The normalization loss aims to prevent misalignment between the 2D projected features and the original 3D features. It achieves this by imposing a constraint on the norm of the feature vector. For further details regarding the loss formulation and refinement, refer to [21].

Once each Gaussian $\mathbf{g_i} \in \mathcal{G}$ has a corresponding feature $\mathbf{f_{g_i}}$ attached to it, we can use these features as a foundation for clustering. We apply a density-based hierarchical clustering algorithm (HDBScan) [22] that can be formally described as a function $f(\mathbf{f_{g_i}}) \rightarrow \{1, 2, ..., M\}$ where M describes the total number of clusters identified by HDBScan. In anticipation of the mask classification stage, we rasterize these clusters back onto 2D to obtain binary 2D

	Total		*Head*		*Common*		*Tail*	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
LERF [4]	10.5	25.8	19.2	28.1	10.1	31.2	2.3	17.6
OpenScene [19]	15.9	24.6	31.7	44.8	14.5	22.6	1.5	6.3
OpenNeRF [6]	19.1	32.1	30.5	44.2	**20.2**	33.5	6.6	18.6
Ours	**19.9**	**33.1**	**38.1**	**47.6**	16.1	**34.4**	**6.7**	**19.3**

Table 1. **3D Semantic Segmentation scores on Replica [20] with reproducible results from LERF, OpenScene, and OpenNeRF.** The *Total* is over all 51 classes, with the *Head*, *Common*, and *Tail* splits defined following OpenNeRF, each consisting of one-third of the total labels with 17 classes each.

segmentation masks. These frames are rasterized from the same perspective as the set of input images $\mathcal{I}$. As a result, we obtain the set of masks $\mathcal{M}_a \in \{0,1\}^{M\times h\times w}$ for every input image $\mathbf{I} \in \mathcal{I}$, consisting of M class-agnostic binary masks.

Stage 2: Mask Classification. Once we have obtained the set of projected 3D-based class-agnostic masks $\mathcal{M}_a$, we need to assign a semantic class label to each of the 3D clusters. We do this using a simple yet effective assignment method. We utilize the OVSeg [16] 2D foundation model for mask classification of the N classes, generating a set of class-aware masks $\mathcal{M}_b \in \{0,1\}^{N\times h\times w}$ in 2D space. Each projected 3D mask $m_a \in \mathcal{M}_a$ should be assigned to the semantic label of the 2D mask $m_b \in \mathcal{M}_b$ with the highest correspondence.

An intuitive approach to associating the sets $\mathcal{M}_a$ and $\mathcal{M}_b$ is to apply a weighted bipartite matching algorithm. Given one mask from each bipartite set $m_a \in \mathcal{M}_a, m_b \in \mathcal{M}_b$, their weight is given by the inverse of the Jaccard Index [23]:

$$w(m_a, m_b) = \sum_{i}^{h}\sum_{j}^{w} \frac{|m_{a,ij} \cup m_{b,ij}|}{|m_{a,ij} \cap m_{b,ij}|}$$

However, we observe that the Segment Anything Model (SAM) primarily proposes masks for instances rather than semantic classes. This means the 3D masks in $\mathcal{M}_a$ often represent multiple instances or parts of the same class in $\mathcal{M}_b$. This difference in the nature of masks introduces a mismatch in the cardinality of the sets $\mathcal{M}_a$ and $\mathcal{M}_b$, as there are generally several instances of each semantic class. Since bipartite matching can only efficiently assign each mask once, this mismatch complicates the process.

Switching to a generalized assignment problem (GAP) would allow multiple assignments but is known to be NP-hard [24], therefore posing significant computational challenges. In contrast, bipartite matching can be efficiently solved using the Hungarian Algorithm [25], which has cubic time complexity. Therefore, we opted not to switch to GAP to maintain computational efficiency. Instead, we replicated the vertices in $\mathcal{M}_b$ corresponding to the number of instances per class to match the instance-level correspondence required. This approach is solvable by the Jonker-Volgenant variant of the Hungarian Algorithm [26, 27], a version for non-square cost-matrices, ensuring a fast and effective assignment of semantic labels to our 3D-based class-agnostic masks.

A significant benefit of our approach is the ability to easily swap the open-vocabulary 2D segmentation model. This can be done without retraining the 3D scene representation since we only utilize the segmentation masks to match fixed 3D mask proposals. This means no retraining is needed if our class-aware model is changed. We test both OpenSeg and OVSeg as the class-aware segmentation models (see Tab. 2). In a similar fashion, one can swap out the mask assignment mechanism with mask classification mechanisms that extract features from 3D mask proposals such as in OpenMask3D [9]. For the sake of computational speed and memory efficiency, we decided to stick to a bipartite assignment mechanism.

3.1. Implementation Details

Our method is implemented in Pytorch and runs on a single Nvidia RTX A5000 GPU with 24GB of memory. Due to the decoupled nature of our method and depending on the available setup and resources, multiple steps (e.g. training of the 3D Gaussian Representation and generation of the 2D segmentation masks) can easily be executed in parallel. The best-performing 2D segmentation model is OVSeg's biggest available model (Swin-Base + CLIP-ViT-L/14), which we utilize for inference only. Regarding the 3D Gaussian Spatting Reconstruction, we closely follow SAGA's approach with slight modifications to the clustering and scale parameters.

ScanNet Ground Truth | Colored Reconstruction | Segmentation (Ours)

Figure 3. **Shortcomings of the ScanNet GT**. Our Method accurately recognizes and segments the posters on the wall, but they are not represented in the provided ScanNet Ground Truth files, therefore lowering our IoU despite a more accurate segmentation of the scene.

4. Experiments

4.1. Datasets

We evaluate our method both on synthetic data with the Replica Dataset [20] as well as real-world data with the ScanNet Dataset [28]. Replica consists of high-quality scenes with realistic textures. It is well-suited for 3D open vocabulary semantic segmentation since it entails a long-tail distribution of small objects and very accurate semantic labels. We evaluate on the commonly used 8 scenes (*office0, office1, office2, office3, office4, room0, room1, room2*). To ensure comparability to the baseline methods, we only evaluate on a subset of 200 of the original posed RGB-D images. The annotations consist of 51 distinct class labels, and we follow OpenNeRF and split them further into (*head, common, tail*) subsets, each consisting of 17 classes. ScanNet consists of high-quality scans of indoor spaces, including significantly larger scenes than Replica. For evaluation, we use the 20-class subset of the NYUv2 40-label set since this is the setting in which the ground truth is given. Note that our method does not use any of the provided ground truth semantic labels for training and is is not bound to the evaluation classes but able to segment any object or concept.

4.2. Metrics

For a quantitative evaluation of our method, we project our semantic predictions back to the given annotated point clouds and follow OpenNeRF and ScanNet to report the *mean intersection over union (mIoU)* and *mean accuracy (mAcc)* for the whole scene as well as the subsets if applicable.

4.3. Synthetic Data: Replica Dataset

When comparing our results to pipelines based only on 2D class-aware segmentation features (e.g. OpenNeRF), we see that our masks are more accurate. This happens, in particular, if the scene has some shadows. This improvement can likely be accounted for by the additional availability of 3D geometry, making classification easier. Compared to OpenNeRF, we can observe that our method achieves less scattered results in large areas. These artifacts are part of the MLP and the rendering function which is based on ray-tracing. In contrast, the explicit geometry in Gaussian Splatting as an underlying representation ensures consistency for these areas. Borders of smaller objects, such as pillows and blankets, are sharper compared to OpenNeRF. Note that decoupling the 3D segmentation proposals from the class-aware segmentation masks allows us to simultaneously perform instance segmentation. Each pillow was assigned ”pillow” as a label, but the clusters were identified separately before the assignment (see Fig. 2).

Our method outperforms the NeRF-based baseline, OpenNeRF, in all but one subfield (*common mIoU*) despite using a completely different architecture that significantly increases the modularity. The effect of differing open-vocabulary segmentation models is apparent when comparing OpenSeg to OVSeg, which offers a notable difference in tail-class performance. This means the segmentation performance is still heavily influenced by the underlying 2D Segmentation Foundation Model, further reinforcing our approach of decoupling the 3D segmentation pipeline to ensure modularity and adaptability to this fast-evolving field.

4.4. Real-World Data: ScanNet v2

OpenNeRF does not report any quantitative measures on real-world data. To validate our performance from synthetic data on real-world data, we evaluate both our method and OpenNeRF on four scenes from ScanNet v2, the commonly used *scene0000_00* from the category *Apartment* as well as one randomly picked scene from *Classroom* (*scene0030_01*) and two from *Bathroom* (*scene0062_00* and *scene0100_01*). It is important to note that these scenes initially contain 5578, 1648, 730, and 1120 posed RGB-D images. To challenge the effectiveness of our method and compare it to synthetic data, we only utilized 200 images for reconstruction and segmentation, meaning only a fraction of the available data for each scene. Additionally, we also don’t utilize any of the available ScanNet annotations

	Total		*Head*		*Common*		*Tail*	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
OpenSeg + Matching	16.17	29.61	31.89	43.76	14.63	32.50	2.93	14.28
OVSeg + Matching	17.96	32.41	35.35	43.41	15.48	31.61	4.11	**24.10**
OpenSeg + Assignment	17.10	27.96	29.87	42.86	**18.95**	33.61	3.47	9.05
OVSeg + Assignment	**19.91**	**33.11**	**38.08**	**47.61**	16.14	**34.37**	**6.69**	19.31

Table 2. **Ablation Study on Replica.** Effect of different segmentation models and bipartite matching vs assignment (see Sec. 4.5)

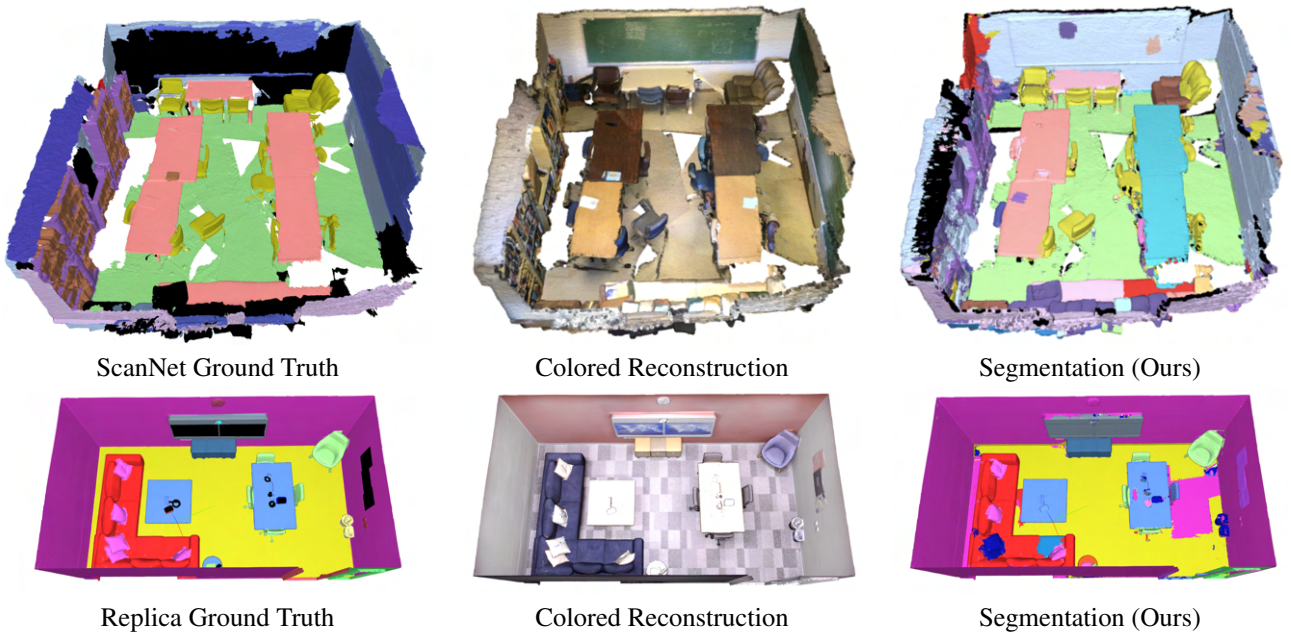

Figure 4. **Further Results on ScanNet (*scene0030_01*) and Replica (*office2*)**

for training but rather perform our segmentation in a zero-shot manner. Qualitative Results are visible in Fig. 3 and Fig. 4.

We observe that there are some areas where segmentation masks are accurate but the assignment of the correct label is unsuccessful. This indicates that our mask proposal is successful, but the underlying 2D foundation model may be unable to assess a given object accurately. A strength that we can observe is that even with limited data, our method is able to pick up long-tail classes that are not even represented in the ground-truth annotations very accurately, as seen with the posters on the wall (see Fig. 3). Keeping the amount of data equal, we continue outperforming OpenNeRF on ScanNet data. While our performance is very competitive with respect to mIoU, we are slightly inferior in terms of mAcc with respect to MinkowskiNet and OpenScene. A likely cause of this observation is that we only utilize a subset of the given images to evaluate a scene.

	mIoU	mAcc
OpenNeRF [6]	49.5	62.7
Ours	**55.1**	**63.5**
MinkowskiNet [29]	69.0	77.5
OpenScene (LSeg) [19]	54.2	66.6
OpenScene (OpenSeg) [19]	47.5	70.7

Table 3. **Semantic segmentation results on *scene0000_00*, *scene0030_01*, *scene0062_00* and *scene0100_01* of ScanNet v2 on 200 images, a fraction of the original amount.** The grey values provide a reference; MinkowskiNet is one of the strongest fully-supervised approaches. OpenScene is zero-shot and utilizes point clouds, i.e. sparse geometry.

4.5. Ablation Study

Effect of different segmentation models. Due to our modular architecture, we can easily swap between differ-

ent 2D Foundation Segmentation Models. We conducted a small ablation study utilizing both OpenSeg and OVSeg. A notable difference in tail-class performance is apparent. Furthermore, our pipeline can be adapted to different tasks by switching the underlying 2D segmentation model to fit the user's specific needs.

Bipartite Matching vs. Bipartite Assignment. As mentioned previously, a bipartite matching formulation faces the challenge that every class can only be assigned once. We hypothesized that relaxing the bipartite matching formulation by introducing duplicates of semantic masks would mitigate the risk of misalignment between instances in the 3D clustering and classes in our 2D foundation model. To test this hypothesis, we have tested both 2D foundation models with a bipartite matching and the relaxed assignment formulation. Our ablation study (as shown in Tab. 2) confirms our hypothesis and indicates that OVSeg, in combination with bipartite assignment, is the most promising approach.

5. Discussion

5.1. Instance and Part Segmentation.

As mentioned previously and demonstrated in Figure 5, our approach predominantly learns masks for instances, enabling precise instance segmentation capabilities.

Figure 5. **Visualization of class-agnostic masks.** Our mask proposal tends to propose instances, as demonstrated by the three separately identified towels and two armchair instances.

Moreover, adjusting the scale parameter s based on the estimated number of visible objects in a scene is a significant advantage. This adaptability allows for the consideration of smaller objects without necessitating retraining of the model, thanks to scale-conditioned affinity features. The modular nature of our approach further enhances its utility, allowing for substituting the mask proposal network with one better suited for more fine-grained tasks, such as part segmentation. The underlying SAM masks that we use for our architectures are not geared toward a specific goal like instance or part segmentation but still demonstrate the ability to perform both tasks as demonstrated in Figure 6. Since our approach is not specialized to perform instance or part segmentation and we are mainly interested in the correct aggregation to class-level, metrics specialized on either of the two targets are not necessarily representative indicators of our approach's performance.

Figure 6. **Part Segmentation.** For objects with clearly separable parts, our approach tends to propose masks that correspond to part segmentation.

In conclusion, our methodology offers distinct advantages over our baselines by enabling instance or even part segmentation without needing network retraining or architectural redesign, thus providing a flexible and robust solution for diverse segmentation tasks.

5.2. Limitations

Bipartite Matching is not optimal. Matching multiple projected instance proposals to 2D segmentation masks ideally requires a generalized assignment instead of a bipartite matching. To account for this weakness, future work could replace the matching step and directly perform the classification step on the mask proposals. While such approaches exist in 2D [30], they must be trained on ground-truth data, requiring significant computational resources for training and large datasets to be generalizable across domains. In 3D, the high computational requirements of such approaches [9, 10] are a concern that needs to be addressed. Another approach that is left for future work and is in line with our architecture involves designing a more flexible assignment algorithm that adjusts to scene size and object counts both in general and in individual frames to further increase the robustness of the assignment formulation.

Tail-class performance is limited by the 2D foundation model. The above approach also addresses the reliance on an almost perfect match between our masks and the compared 2D masks. Classes that the 2D model fails to recognize but are identified by the 3D clustering cannot be accurately labeled. Thus, one of our key advantages—accurately

identifying long-tail classes using a combination of 3D geometry and segmentation features—is compromised if the class-aware foundation model underperforms.

Gaussians vs. Sharp Edges. Gaussians, due to their inherent spherical nature, sometimes struggle to accurately segment objects with sharp edges. This limitation leads to imprecise boundaries and overlaps in the segmentation output as seen in Figure 7. There are alternative approaches and modifications to Gaussian-based models to better handle complex geometries with sharp edges and mitigate this issue. For instance, Hu et al. [31] refine Gaussian segmentation by decomposing Gaussians to address this shortcoming, enabling Gaussian-based segmentation methods for domains in which accuracy is crucial.

Figure 7. The nature of 3D Gaussians prevents clear edges when segmenting (top edge). Additionally, we can observe low-connectivity clusters with few pixels between the two walls.

Outlier Clusters. The proposed feature extraction represents a rather novel method to leverage 3D Gaussians for segmentation mask proposal, offering a new perspective in the realm of 3D segmentation. However, an observed challenge (see Fig. 7) with this approach is its sensitivity to outlier clusters that exhibit low connectivity. These clusters, which do not conform to the expected connectivity patterns, can adversely affect the overall accuracy of the segmentation. To counteract this, we implement a post-processing step that systematically identifies and eliminates clusters with low connectivity. Despite this measure, challenges persist when clusters that marginally exceed the connectivity threshold remain in the dataset. Such clusters continue to pose a problem, indicating the need for more sophisticated strategies to ensure robust segmentation.

6. Conclusion

In this work, we presented DCSEG, a decoupled pipeline for open-vocabulary 3D semantic segmentation that is simultaneously able to segment parts and instances that can be aggregated to classes without the need for retraining. We utilize 3D Gaussian Splatting as an underlying scene representation. This alternative to NeRF-based approaches shows improved results while being computationally more efficient. Additionally, we provide a way to approximate a general assignment by matching clusters over multiple image pairs and propose a modular framework that can easily be adapted if novel methods for either 3D instance proposals or 2D open-vocabulary segmentation become available.

References

[1] 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. 1, 2

[2] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 2

[4] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 2, 4

[5] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022.

[6] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, Marc Pollefeys, and Federico Tombari. OpenNeRF: Open Set 3D Neural Scene Segmentation with Pixel-Wise Features and Rendered Novel Views. In *International Conference on Learning Representations*, 2024. 2, 4, 6

[7] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2

[8] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1086–1096, 2023. 2

[9] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 4, 7

[10] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3292–3302, 2024. 2, 3, 7

[11] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *arXiv preprint arXiv:2403.15624*, 2024. 2

[12] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024.

[13] Mingrui Li, Shuhong Liu, and Heng Zhou. Sgs-slam: Semantic gaussian splatting for neural dense slam. *arXiv preprint arXiv:2402.03246*, 2024. 2

[14] 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. 2

[15] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2

[16] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2, 4

[17] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4018–4028, 2024. 3

[18] Hanchen Tai, Qingdong He, Jiangning Zhang, Yijie Qian, Zhenyu Zhang, Xiaobin Hu, Yabiao Wang, and Yong Liu. Open-vocabulary sam3d: Understand any 3d scene. *ArXiv*, abs/2405.15580, 2024. 3

[19] 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. 4, 6

[20] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 4, 5

[21] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. *arXiv preprint arXiv:2312.00860*, 2023. 3

[22] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2013. 3

[23] Sam Fletcher, Md Zahidul Islam, et al. Comparing sets of patterns with the jaccard index. *Australasian Journal of Information Systems*, 22, 2018. 4

[24] Dirk G Cattrysse and Luk N Van Wassenhove. A survey of algorithms for the generalized assignment problem. *European journal of operational research*, 60(3):260–272, 1992. 4

[25] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[26] Roy Jonker and Ton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38:325–340, 1987. 4

[27] David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. 4

[28] 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. 5

[29] 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. 6

[30] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 7

[31] Xu Hu, Yuxi Wang, Lue Fan, Junsong Fan, Junran Peng, Zhen Lei, Qing Li, and Zhaoxiang Zhang. Sagd: Boundary-enhanced segment anything in 3d gaussian via gaussian decomposition, 2024. 8