Open-Set 3D Scene Segmentation with Rendered Novel Views

Francis Engelmann
ETH Zurich, Google

Fabian Manhardt
Google

Michael Niemeyer
Google

Keisuke Tateno
Google

Marc Pollefeys
ETH Zurich

Federico Tombari
Google

Abstract

Recently, large visual-language models (VLMs), like CLIP, enabled open-set image segmentation to segment arbitrary concepts from an image in a zero-shot manner. This goes beyond the traditional closed-set assumption, *i.e.*,where models are limited to segment only those classes which appear in a pre-defined training set. Very recently the first works on exploring open-set segmentation in 3D scenes have appeared in the literature. These methods are heavily influenced by closed-set 3D convolutional approaches that process point clouds or polygon meshes. However, these 3D scene representations do not align well with the image-based nature of the visual-language models. Indeed, point cloud and 3D meshes typically have a lower resolution than images and the reconstructed 3D scene geometry might not project well to the underlying 2D image sequences used to compute pixel-aligned CLIP features. To address these challenges, we propose an approach based on neural radiance fields (NeRF), that naturally operates on posed images and directly encodes the VLM features within the NeRF. Our experiments demonstrate that NeRF representations improve the quality of zero-shot segmentation over point cloud-based methods. Our approach, called OpenReNo, further leverages NeRF's ability to *re*nder *no*vel views and extract open-set features from areas that are not well observed in the initial image sequence. In experiments, OpenReNo significantly outperforms the recent open-vocabulary method of OpenScene by $+4.9$ mIoU on the challenging Replica dataset.

1 Introduction

The 3D semantic segmentation of a scene is the task of estimating, for each 3D point of a scene, the category that it belongs to. Being able to accurately estimate the scene semantics is of high importance as it empowers many applications, including robotics [39] and autonomous driving [16], since they all require having a very detailed understanding of the environment. While the domain of 3D scene segmentation has recently made a lot of progress [5, 21, 24, 28, 32], these methods are exclusively trained in a supervised manner on a (*closed*-)set of semantic categories, rendering them impractical for many real world applications as the models lack the flexibility to continuously adapt to new concepts/classes in the scene [2, 4, 6, 25, 30]. Therefore, in this work we aim at tackling the problem of open-set 3D scene segmentation. The main idea of *open*-set scene segmentation is that arbitrary concepts can be segmented, independent of any pre-defined closed set of classes. Specifically, given an *arbitrary* query – for example, a textual description or an image of an object – the goal is to segment those parts in the 3D scene that are described by the query. Such general unconstrained functionality can be crucial for helping robots interact with previously unseen environments, or applications on AR/VR devices in complex indoor scenes, especially for queries where annotated training labels are scarce or not available at all [1].

Preprint. Under review.

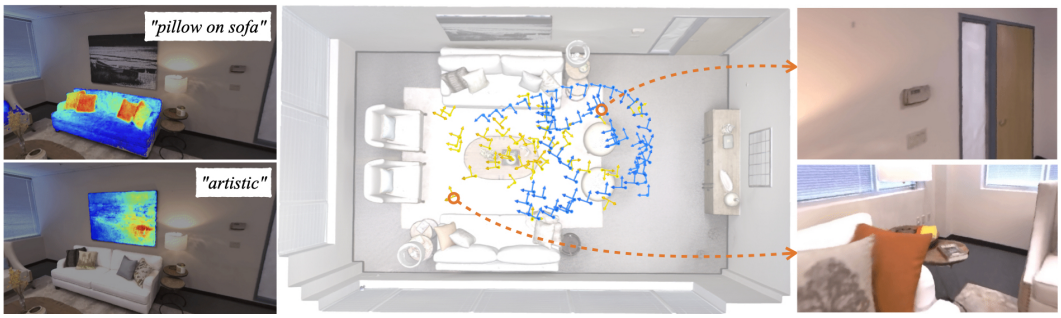

Figure 1: We propose OpenReNo, an approach for open-set 3D scene understanding based on neural radiance fields. Arbitrary concepts can be queried from our representation *(left)*. As the original camera trajectory (shown in blue, middle) might not capture all interesting scene details, we use NeRFs ability to render novel views and propose a mechanism to obtain relevant novel camera poses (shown in yellow, middle) that focus on scene details from which we extract open-scene features improving the overall open-set scene representation.

On another note, vision-language models (VLMs) (such as CLIP [34] or ALIGN [13]) have shown impressive performance on open-set image classification. Trained on Internet-scale image-caption pairs, these models learn a joint embedding space mapping text and image inputs to the same (or different) embedding vector depending on whether they describe similar (or different) concepts. More recently, these powerful concepts were even applied to dense pixel-level tasks, enabling open-set 2D image segmentation [8, 17, 26]. On the other hand, using VLMs in combination with 3D point clouds is to date mostly unexplored. Similar to [17], Rozenberszki *et al.* [27] train a 3D convolutional network that predicts per-point CLIP features in a fully-supervised manner based on CLIP-text encodings from the annotated per-point class names. Such fine-tuning on densely annotated datasets works well on the labeled training set classes but has limited generalization to novel unseen classes which hinders open-set scene understanding abilities. OpenScene [23] has instead recently demonstrated the first exciting results on open-vocabulary 3D scene understanding. In line with [27], given a reconstructed 3D point cloud as input, a 3D convolutional network predicts CLIP features for each point. However, unlike [27], the model is trained on projected CLIP-image features from posed 2D images, preserving the generalization ability of the pixel-aligned visual-language image features.

This initial work, however, necessitates 3D scene reconstruction in the form of a polygon mesh (or a point cloud sampled from the mesh surface) that usually stems from a recorded sequences of RGB-D frames. The pixel-aligned CLIP features are extracted from the accompanying RGB frames. This can lead to two drawbacks: First, some parts of the scene might not be well captured by the provided camera trajectory, *e.g.*, larger objects might be cropped, or lacking sufficient context when the camera is too close. In addition, smaller objects might cover only a few pixels in the image, making it hard to obtain meaningful pixel-aligned CLIP features. Second, mesh-based approaches are inherently limited by the mesh resolution and might be unable to represent smaller objects.

In this work, we address both aforementioned limitations and propose OpenReNo, the first Neural Radiance Fields (NeRFs) [20] based approach for open-set 3D scene segmentation. As neural representation, NeRFs have inherently unlimited resolution and, more importantly, they also provide an intuitive mechanism for rendering novel views from arbitrary camera positions. We leverage this ability to extract additional visual-language features from novel views leading to a large boost in the model's segmentation performance. In particular, this mechanism enables the extraction of pixel-aligned CLIP features from novel view points that focus on interesting areas of the 3D scene, even if they significantly deviate from the real camera trajectory used during scene capture. One key challenge is to determine the relevant parts of the scene requiring further attention. We identify the disagreement from multiple views as a powerful signal and propose a probabilistic approach to generate novel view points.

We experimentally demonstrate that NeRF representations exceed point cloud based representations, despite not even requiring depth information, *i.e.* using posed RGB images only. We further show improved segmentation performance by incorporating pixel-aligned CLIP features from novel views. We identify the Replica dataset as a great candidate to evaluate open-set 3D semantic segmentation since, unlike ScanNet or Matterport, it comes with very accurate mesh reconstruction, per-point semantic labels as well as a long-tail class distribution (see Figure 3).

In summary, the contributions of this work are as follows:

- We propose OpenReNo, the first approach for open-set 3D semantic scene understanding based on neural radiance fields (NeRF).
- We propose a mechanism that relies on NeRFs ability to render novel views to extract additional visual-language features, leading to improved segmentation performance.
- OpenReNo significantly outperforms the current state-of-the-art for open-vocabulary 3D segmentation with an $+4.5$ mIoU gain on the challenging Replica dataset.

2 Related Work

2D Visual-Language Features. CLIP [34] is a large-scale visual-language model trained on Internet-scale image-caption pairs. It consists of an image-encoder and a text-encoder that map the respective inputs into a shared embedding space. Both encoders are trained in a contrastive manner, such that they map images and captions to the same location in the embedding space if the caption describes the image, and different locations in the opposite case. While the CLIP image-encoder yields a single global feature vector per image, LSeg [17] extends this idea and predicts pixel-level features which enables dense image segmentation tasks. Pixel-aligned CLIP features are obtained via fine-tuning on a fully-annotated 2D semantic segmentation dataset. This works well for the semantic classes present in the fully-annotated dataset, however, the pixel-aligned features generalize less well to novel concepts outside of the training classes. OpenSeg [8] further improves on these aspects and proposes a class-agnostic fine-tuning to obtain pixel-aligned features. OVSeg [19] is the latest development that proposes to fine-tune a CLIP-like model on cropped objects. In this work, we use the pixel-wise features from OpenSeg [8] which enables a direct and fair comparison with the publicly available models of OpenScene [23]. Further, as also noted by [23], OpenSeg improves over LSeg [17] on long-tail classes that were not seen during the training of LSeg.

Neural Radiance Fields. Since the introduction of Neural Radiance Fields (NeRFs) [20] for view synthesis, they have been adopted as scene representation for various tasks ranging from 3D reconstruction [22, 35, 37, 38] to semantic segmentation [39] due their simplicity and state-of-the-art performance. Next to impressive view synthesis results, NeRFs also offer a flexible way of fusing 2D-based information in 3D. A series of works have explored this property in the context of 3D semantic scene understanding. In PanopticLifting [29], predicted 2D semantic maps are utilized to obtain a 3D semantic and instance-segmented representation of the scene. In [15], 2D semantic features are incorporated and fused during the NeRF optimization which are shown to allow for localized edits for input text prompts. Neural Feature Fusion Fields [33] investigates the NeRF-based fusion of 2D semantic features in the context of 3D distillation and shows superior performance to 2D distillation. Finally, in LERF [14], NeRF-based 3D CLIP-feature fields are optimized via multi-scale 2D supervision to obtain scene representations that allow for rendering response maps for long-tail open-vocabulary queries. While all of the above achieve impressive 3D fusion results, we propose to investigate NeRF-based feature fusing in the context of open-set 3D scene segmentation. This does not only require to solve additional challenges, such as detecting relevant parts of the scene, but also enables more rigorous evaluation and comparison to other fusion approaches.

3D Open-Set Scene Understanding While most approaches for 3D scene understanding utilize 3D supervision [10, 11, 18], a recent line of works investigates how 2D-based information can be lifted to 3D. In Semantic Abstraction [9], 2D-based CLIP features are projected to 3D space via relevancy maps, which are extracted from an input RGB-D stream. While achieving promising results, their 3D reasoning is coarse and thus limited. Similarly, in ConceptFusion [12] multi-modal 3D semantic representations are inferred from an input RGB-D stream by leveraging 2D foundation models. They use point clouds as 3D representation. ScanNet200 [27] uses CLIP to investigate and develop a novel 200-class 3D semantic segmentation benchmark. Most similar to our approach is OpenScene [23], which is the only existing method for open-set 3D semantic segmentation of point clouds. Unlike [23] and the other above-mentioned methods, our approach does not require pre-computed 3D polygon meshes or point clouds and can be used with input RGB or RGB-D data. Further, our NeRF-based scene representation offers high-quality view synthesis capabilities that we use to generate novel renderings of interesting parts of the scene improving segmentation performance.

3 Method

Given a set of posed RGB input images and corresponding pixel-aligned open-set image features, we want to to obtain a continuous volumetric 3D scene representation that enables to query the scene for any arbitrary concept. Our approach, called OpenReNo, enables open-set 3D scene understanding to address a wide variety of different tasks, such as material and property understanding as well as object localization. By means of leveraging the normalized cosine similarity between the encoded queries and the rendered open-set features, we are capable of exploring various different concepts (see Figure 6). Alternatively, our approach can by seen as a method for unsupervised 3D semantic segmentation (Figure 5) where we assign to each point the semantic class with the highest similarity.

3.1 Open-Set Radiance Fields

A radiance field is a continuous mapping that predicts a volume density $\sigma \in [0, \infty]$ and a RGB color $\mathbf{c} \in [0, 1]^3$ for a given input 3D point $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{S}^2$. Mildenhall et. al. [20] propose to parameterize this function with a neural network (NeRF) using a multi-layer perceptron (MLP) where the weights of this MLP are optimized to fit a set of input images of a scene. To enable higher-frequency modeling, the input 3D point as well as the viewing direction are first passed to a predefined positional encoding [20, 31] before feeding them into the network. Building on this representation, we additionally assign an open-set feature $\mathbf{o} \in \mathbb{R}^D$ to each 3D point:

$$f_\theta(\mathbf{x}, \mathbf{d}) \mapsto (\sigma, \mathbf{c}, \mathbf{o}) \tag{1}$$

where θ indicates the trainable network weights. We base our scene representation on Mip-NeRF [3] for both the appearance and density, which we extent with another MLP head to model the open-set field. Essentially, the rendering of color, density and the open-set feature maps is conducted following the volumetric rendering paradigm via integrating over sampled point positions along a ray r.

3.2 Training Objective

Appearance Loss. Following standard procedure, we optimize the appearance using the Euclidean distance between the rendered color $\hat{\mathbf{c}}_r$ and the ground truth color $\mathbf{c}_r$ over a set of sampled rays $\mathcal{R}$ within a training batch:

$$\mathcal{L}_{\text{RGB}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} ||\mathbf{c}_r - \hat{\mathbf{c}}_r||^2. \tag{2}$$

Open-Set Loss. Similarly, the open-set features are supervised via pre-computed 2D open-scene feature maps from OpenSeg. However, in contrast to before, we instead maximize the cosine similarity between both via:

$$\mathcal{L}_{\text{open}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} -\frac{\mathbf{o}_r}{||\mathbf{o}_r||_2} \cdot \frac{\hat{\mathbf{o}}_r}{||\hat{\mathbf{o}}_r||_2}. \tag{3}$$

Since the 2D open-set feature maps are generally not multi-view consistent, we do not backpropagate from the open-set feature branch to the density branch [15, 29]. Additionally, the 2D open-set maps from OpenSeg [8] contain many artifacts near the image border. Therefore, we avoid sampling rays within a margin of 10 pixels from the image border during training.

Depth Loss. While our method is able to be trained from RGB data alone, we can still leverage depth data if available to further improve the results. To this end, we supervise the density for those pixels where ground truth depth information is available (see Table 2 for an analysis). We compute the mean depth from the densities for each ray r and supervise them using the smooth L_1 (Huber) loss:

$$\mathcal{L}_{\text{depth}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} ||d_r - \hat{d}_r||. \tag{4}$$

To summarize, our total loss is defined as $\mathcal{L} = \mathcal{L}_{\text{RGB}} + \lambda_{\text{open}} \cdot \mathcal{L}_{\text{open}} + \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{depth}}$.

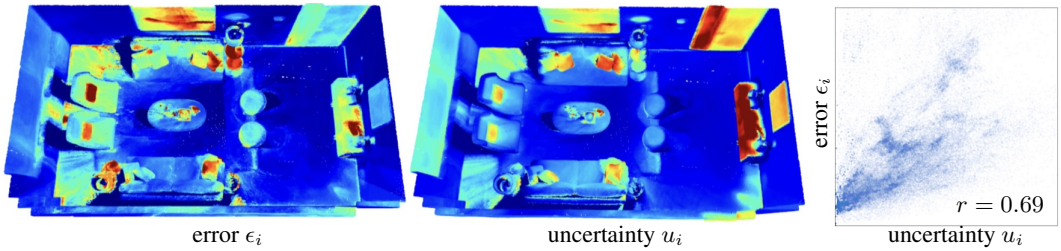

Figure 2: Confidence Estimation. The error e_i *(left)* correlates well with the estimated uncertainty u_i *(center)*. Our mechanism for selecting novel view points is based on the estimated uncertainty. The plot *(right)* shows the error-uncertainty correlation r for `room0` of the Replica [30] dataset.

3.3 Rendering Novel Views

A key advantage of NeRF-based representations is their ability to render photo-realistic novel views. These rendered novel views can naturally be used to extract 2D open-set features. We would like to use this ability to obtain improved open-set features for those parts of the scene where we have low confidence in the existing open-set features. A key challenge, however, is to first identify the parts of the scene that exhibit low confidence features and would thus benefit from rendering novel views.

Confidence Estimation. We identify the uncertainty $u_i \in \mathbb{R}$ over multiple projected open-set features as a surprisingly strong signal. Specifically, starting with the original color images, we obtain their corresponding open-set feature maps with OpenSeg. We then project each feature map onto the scene point cloud and compute for each point i the mean $\mu_i \in \mathbb{R}^D$ and covariance $\Sigma_i \in \mathbb{R}^{D \times D}$ over the per-point projected features. As per-point uncertainty measure $u_i \in \mathbb{R}$ we compute the *generalized* variance [36] defined as $u_i = \det(\Sigma_i)$. Those parts of the scene that exhibit a large uncertainty intuitively correspond to areas where the open-set features from multiple viewpoints disagree on. They hence deserve further investigation by means of re-rendering from more suitable viewpoints. In practice, calculating the variance over a large number of high-dimensional features can be computational challenging. For a memory efficient and numerically stable implementation, we thus rely on Welford's online algorithm [7] to compute the variance in a single pass.

To confirm that idea, we compute the correlation between the per-point uncertainty u_i and the per-point error ϵ_i. We define the per-point error ϵ_i as the Euclidean distance between the ground truth open-set vector $\mathbf{o}_i^{\text{gt}}$ (*i.e.*, the CLIP text-encoding of the annotated class name) and the per-point mean open-set vector μ_i. Indeed, measured over all Replica scenes, we observe a strong positive correlation ($\bar{r} = 0.653$) between u_i and ϵ_i. See Figure 2 for an illustration.

Novel Camera View Selection. To generate novel camera poses we compute the `lookat` matrix based on a target $\mathbf{t} \in \mathbb{R}^3$ and camera position $\mathbf{p} \in \mathbb{R}^3$. We then sample target candidates from the scene surface based on farthest point sampling (FPS) and accept candidates based on their uncertainty u_i if $x < u_i$ with $x \sim \texttt{Uniform}[0, 1]$. The camera position is placed at a random offset from the target position inside the scene. Importantly, we sample the density of the NeRF at the camera position to make sure it does not collide with the scene geometry.

3.4 Implementation and Training Details

Our model is implement in Jax based on Mip-NeRF [3]. We train each scene representation on 8 NVIDIA V100 in parallel for 3000 iterations ($\sim$ 10 mins / scene). The pixel-aligned opens-set features are computed using OpenSeg[8] resulting in $640 \times 360 \times D$ dimensional feature maps with $D = 768$. For memory efficiency reasons, we convert them to float16 values. For querying, we use the pre-trained CLIP text-encoder based on the ViT-144@336 model [34].

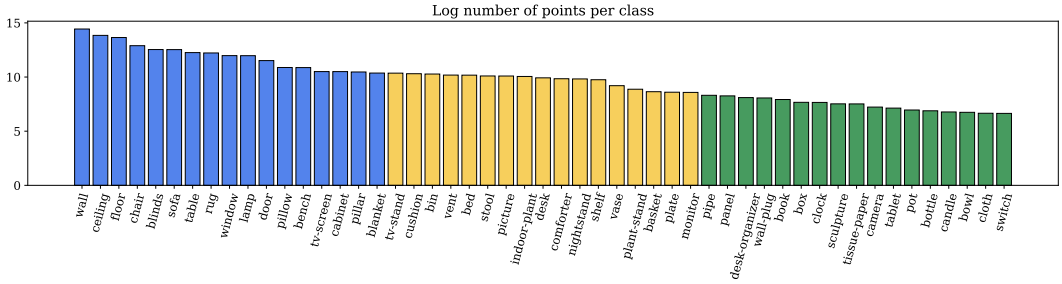

Figure 3: Class Frequency Distribution of the Replica [30] Dataset. We show the number of point annotations for each category. The colors indicate the separation in *head* (blue), *common* (yellow) and *tail* (green) classes from left to right in decreasing order. Note that the plot is shown at log-scale.

4 Experiments

4.1 Dataset and Metrics

We evaluate our approach on the Replica [30] dataset. Replica consists of high quality 3D reconstructions of a variety of real-world indoor spaces with photo-realistic textures. Unlike other popular 3D semantic segmentation datasets, such as Scannet [2, 6] or Matterport [25], Replica is particularly well suited to evaluate open-set 3D scene understanding as it contains both a long-tail class distribution and carefully-annotated ground-truth semantic labels, including very small objects such as switches and wall-plugs. All experiments are evaluated on the commonly-used [23, 39, 40] 8 scenes `{office0-4, room0-2}`, using the camera poses and RGB-D frame sequences from Nice-SLAM [40]. Each RGB-D video sequence consists of 2000 frames which are scaled to a resolution of 640×360 pixels. Following OpenScene [23], we use only every 10-th frame for training such that, for each scene, we have 200 posed RGB-D frames.

Labels. Overall, the 3D scene reconstructions are annotated with 51 different semantic class categories. To enable a more detailed analysis of the experiments, we further split the original categories into three equally sized subsets (*head, common, tail*) based on the number of annotated points where each subset contains 17 classes (see Figure 3). Note that the ground-truth semantic labels, however, are only used for evaluation and not for training or optimization of the 3D scene representations.

Metrics. In the evaluation, we compare 3D semantic segmentation performance on the provided scene reconstructions. In particular, we follow [23] and measure the accuracy of the predicted semantic labels using the mean intersection over union (mIoU) and mean accuracy (mAcc) over all ground truth annotated semantic classes.

4.2 Methods in comparison.

We compare our approach to the recently proposed OpenScene [23], which is currently the only existing method for open-world 3D semantic segmentation. The OpenScene model is a sparse 3D convolutional network that consumes a 3D point cloud and predicts for each point an open-set feature. The model is trained on large-scale 3D point cloud datasets and supervised with multi-view fused CLIP features. We compare with the two variations of their trained model, the 3D distilled model (Distilled), which directly predicts per-point features, and the improved 2D-3D ensemble model (Ensemble), which additionally combines the predicted per-point features with fused pixel features. For a fair comparison, we follow their experimental setup, using the same pixel-aligned visual-language feature extractor OpenSeg [8]. We further use their publicly available source code and the Matterport [25] pre-trained model as suggested in the official code repository [1].

[1] https://github.com/pengsongyou/openscene

	All		*Head*		*Common*		*Tail*	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
OpenScene [23] *(Distilled)*	14.8	23.0	30.2	41.1	12.8	21.3	1.4	6.7
OpenScene [23] *(Ensemble)*	15.9	24.6	31.7	44.8	14.5	22.6	1.5	6.3
OpenReNo (Ours)	**20.4** ±0.25	**31.7** ±0.37	**35.4** ±0.39	**46.2** ±0.56	**20.1** ±0.43	**31.3** ±0.55	**5.8** ±0.11	**17.6** ±0.14

Table 1: 3D Semantic Segmentation Scores on Replica [30]. All results are obtained from image resolution of 640×360 pixels, using the OpenSeg image encoder and the CLIP text encoder based on ViT-144@336 and averaged over three runs. The reported OpenScene scores are obtained with their provided pre-trained models. Note that the OpenScene models additionally profit from pre-training on the large-scale Matterport [25] dataset.

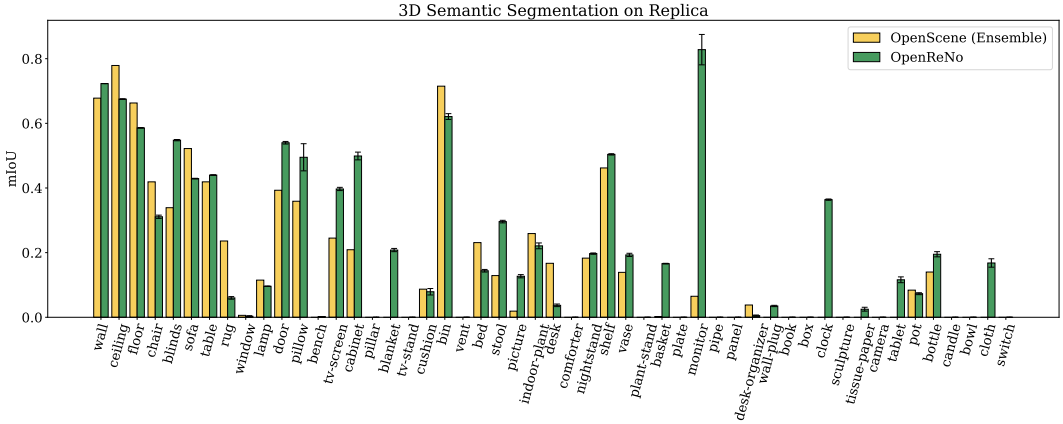

Figure 4: Per-Class Scores for 3D Semantic Segmentation on Replica [30]. We compare our model to the best-performing OpenScene variant (Ensemble). For our model, we show the mean over three runs and the error bars indicate the standard deviation. OpenScene provides only one pre-trained model.

4.3 Results on 3D Semantic Segmentation

We show the 3D semantic segmentation scores for all three subsets in Table 1. We further plot the per-class scores in Figure 2. The 3D scene segmentations are obtained by querying the 3D scene representations for each one of the annotated ground truth semantic classes and assigning the class with the highest similarity score to each 3D point. Querying is performed via correlation of the 3D point cloud features with the embedding from large language models, specifically the CLIP-text encoder (ViT-L14@336). For OpenScene, we observe a similar trend as reported in [23], where the *Ensemble* model improves over the *Distilled* model ($+1.1$ mIoU).

Our results significantly exceed OpenScene by $+4.5$ mIoU over all classes, and by $+3.7$ (head), $+5.6$ (common) and $+4.3$ (tail) for each subset. We achieve the smallest improvement on the head classes. We attribute this to the fact that OpenScene can benefit from the pre-training on large-scale 3D datasets in this setting, enabling OpenScene to learn geometric priors of more popular classes (wall, ceiling, floor, chair). However, it is interesting to note that our approach is able to detect semantic categories that are not recognized at all by OpenScene (wall-plug, clock, tissue-paper, tablet, cloth). This is an important aspect, as these are often exactly the classes that are most relevant for an autonomous agent to interact with. Nevertheless, we also observe that numerous long-tail classes are not detected at all by neither method. This highlights that open-scene segmentation is a difficult problem, and also shows that Replica is a challenging dataset for benchmarking.

4.4 Analysis Experiments

Sampling or rendering? Unlike OpenScene, which directly predicts per-point features for each point of a given 3D point cloud, our NeRF representation is more versatile. We can either directly *sample* ① the open-set features at a specified 3D position from the NeRF representation, or we can first *render* and then *project* ② the open-set features onto the 3D point cloud for evaluation. When multiple open-set features are projected to the same 3D point we take the average over all points. Note that in both cases the 3D point cloud is only required for evaluation and, in contrast to OpenScene, not necessary to obtain the NeRF-based scene representation, which only relies on posed RGB(-D) images. Table 2 shows that the projection approach ② improves over sampling ①, which can be a direct consequence of the volumetric rendering within NeRFs that accumulates multiple samples along each ray compared to a single sample at a given 3D point. Note that the sampling ① already improves over [23].

	mIoU	mAcc
OpenScene [23] *(Ensemble)*	15.09	24.6
① Sampled	16.5	29.1
② Render & Project	18.5	29.8
③ + depth supervision	19.4	30.8
④ + rendered novel views	**20.4**	**31.7**

Table 2: Analysis Experiments.

Impact of Depth Supervision. We proceed with a second experiment to analyze the importance of depth as additional supervision signal. We implement an additional regression loss using the Huber (smooth-L_1) loss between the rendered average distance and the ground truth depth from each training pose. Table 2, ③ shows that, perhaps not too surprising, this improves the open-set feature field since the additional depth supervision has a direct impact on the volumetric reconstruction quality.

Impact of Novel Views. Next, we analyze the contribution of the rendered novel views from the generated view points using the approach described in Section 3.3. Table 2, ④ clearly demonstrates the increased performance from rendered novel views. To disseminate whether the improvement comes from the novel views or simply from additional views, we also compare with views generated along the same camera trajectory as the original views which yielded comparable results as ③. We then placed additional cameras into the scene volume to render novel views from random position. This results in a drastic performance drop (15.4 mIoU) due to numerous frames that are either inside the scene geometry or showing no meaningful context leading to deteriorated open-set features.

4.5 Qualitative Results for 3D Scene Segmentation and Open-Set Applications

In Figure 5, we compare qualitative semantic segmentation results of OpenScene and our OpenReNo. The white dashed circles highlight the different predictions of both methods. In contrast to OpenScene, our approach is able to correctly segment the wall-plugs as well as the blanket on the bed (Figure 5, *left*). Our approach also properly detects the basket, the blinds and produces less noisy results on the door (Figure 5, *right*).

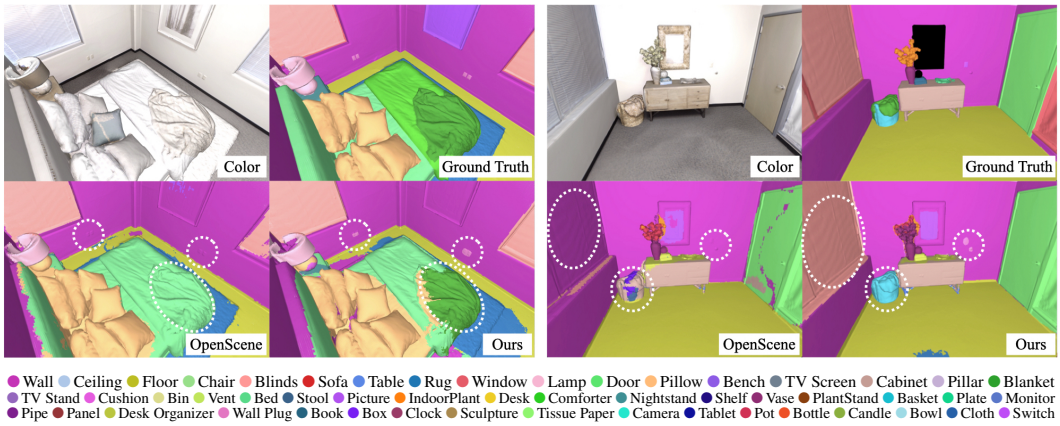

Figure 5: Qualitative 3D Segmentation Results and Comparison with OpenScene [23]. The white dashed circles indicate the most noticeable differences between both approaches. Color and ground truth are shown for reference only. Overall, our approach produces less noisy segmentation masks.

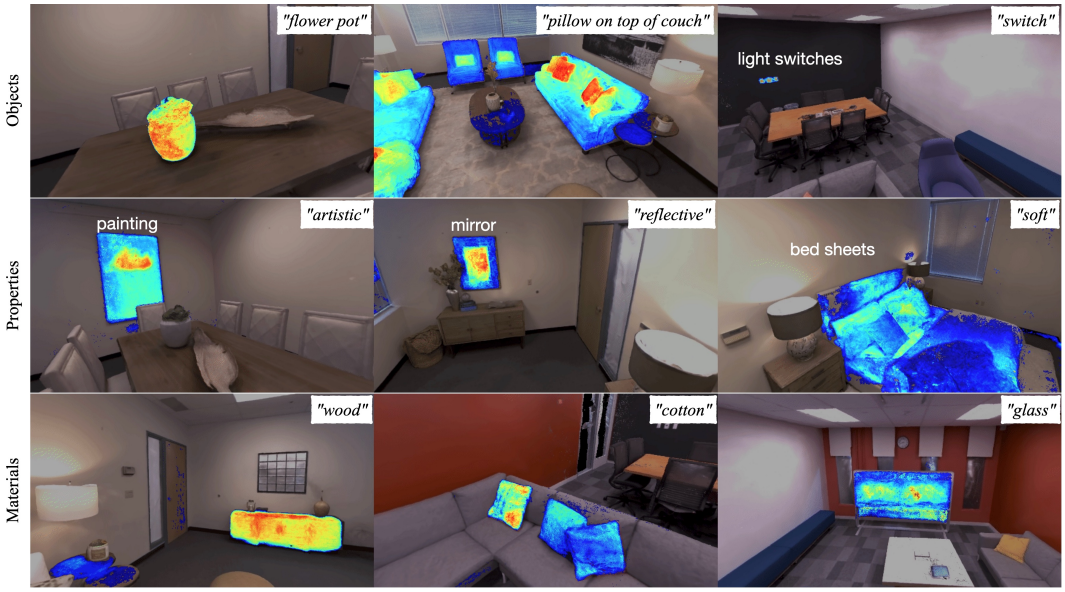

Figure 6: Open-Set Scene Exploration. We visualize the normalized cosine similarity between the rendered open-set scene features and the encoded text queries shown at the top-right of each example. Examples cover a broad range of concepts. Going beyond specific objects (*top*), we show scene properties (*middle*) and various materials (*bottom*). **Red** is the highest relevancy, **green** is middle, **blue** the lowest. Uncolored means the similarity values are under 0.5.

4.6 Open-Set 3D Scene Exploration

The more interesting aspect of open-set scene representation is that they can be queried for arbitrary concepts. In Figure 6, we show the response of open-set queries. For each example, we provide the text query as label. We observe that our method can be used to not only query for classes, but also for concepts like object properties or material types.

5 Limitations and Broader Impact

Broader Impact. Our proposed system allows for open set unsupervised 3D semantic scene understanding. As a result, our system is not restricted by design, which could lead to unintended use cases, such as the recovery or identification of sensitive data. Further, our method utilizes 2D foundation models to obtain features, inheriting biases as well as limitations from these models. Finally, similar to other deep-learning based approaches, our system requires powerful GPU hardware and thus has a high energy consumption.

Limitations. A potential limitation of NeRF scene representation is that they can be limited to room/house-scale scenes, while more elaborated approaches are required to represent larger scenes as for example on a city-scale. Moreover, the quality of our CLIP features from novel views is inherently limited by the rendering quality of the trained radiance field.

6 Conclusion

We presented OpenReNo, a NeRF based scene representation for open set 3D semantic scene understanding. We demonstrate the potential of NeRFs as a powerful scene representations compared to explicit mesh representations, specifically on the task of unsupervised 3D semantic segmentation scene segmentation. We further exploit the novel view synthesis capabilities of NeRF to compute additional views of the scene from which we can extract open-set features. This enables us to focus in greater detail on areas that remain underexplored. To that end, we proposed a mechanism to identify regions for which novel views should be generated. Our experiments show that our NeRF-based representation outperforms other mesh-based methods such as OpenScene.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as I Can, Not as I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes–a diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] E Knuth Donald et al. The Art of Computer Programming. *Sorting and searching*, 1999.

[8] 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 (ECCV)*, 2022.

[9] Huy Ha and Shuran Song. Semantic Abstraction: Open-World 3D Scene Understanding from 2D Vision-Language Models. In *Conference on Robot Learning*, 2022.

[10] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2937–2946. Computer Vision Foundation / IEEE, 2020.

[11] Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In *International Conference on Computer Vision (ICCV)*, pages 15468–15478. IEEE, 2021.

[12] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. ConceptFusion: Open-Set Multimodal 3D Mapping. *arXiv preprint arXiv:2302.07241*, 2023.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[14] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language Embedded Radiance Fields. *arXiv preprint arXiv:2303.09553*, 2023.

[15] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing NeRF for Editing via Feature Field Distillation. 2022.

[16] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022.

[17] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-Driven Semantic Segmentation. *ICLR*, 2022.

[18] Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11799–11808. IEEE, 2022.

[19] 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. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] 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. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, 2020.

[21] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-Context Data Augmentation for 3D Scenes. In *International Conference on 3D Vision (3DV)*, 2021.

[22] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021.

[23] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. OpenScene: 3D Scene Understanding with Open Vocabularies. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[24] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. 2017.

[25] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport

3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[26] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseClip: Language-Guided Dense Prediction with Context-Aware Prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[27] David Rozenberszki, Or Litany, and Angela Dai. Language-Grounded Indoor 3D Semantic Segmentation in the Wild. In *European Conference on Computer Vision (ECCV)*, 2022.

[28] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Instance Segmentation. *International Conference on Robotics and Automation (ICRA)*, 2023.

[29] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic Lifting for 3D Scene Understanding with Neural Fields. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[30] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[31] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. 2020.

[32] 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 *International Conference on Computer Vision (ICCV)*, 2019.

[33] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representations. In *International Conference on 3D Vision (3DV)*, 2022.

[34] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[35] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. 2021.

[36] Samuel S Wilks. Certain Generalizations in the Analysis of Variance. *Biometrika*, 1932.

[37] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. 2021.

[38] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. 2022.

[39] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place Scene Labeling and Understanding with Implicit Scene Representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[40] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-SLAM: Neural Implicit Scalable Encoding for SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.