Language-Grounded Indoor 3D Semantic Segmentation in the Wild

David Rozenberszki[1], Or Litany[2], and Angela Dai[1]

[1]Technical University of Munich [2]NVIDIA

https://rozdavid.github.io/scannet200

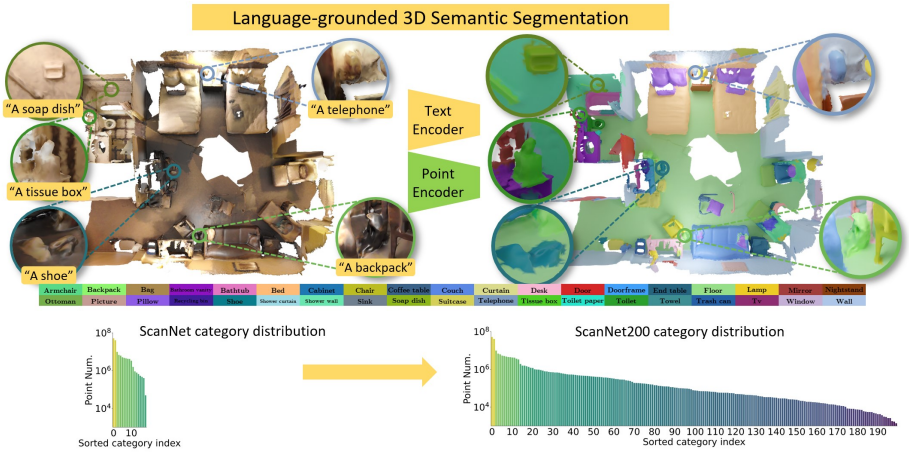


Fig. 1: We present the ScanNet200 benchmark, which studies 200-class 3D semantic segmentation – an order of magnitude more categories than previous 3D scene understanding benchmarks. To address this challenging task, we propose to guide 3D feature learning by anchoring it to the richly-structured text embedding space of CLIP for the semantic class labels. This results in improved 3D semantic segmentation across the large set of class categories.

Abstract. Recent advances in 3D semantic segmentation with deep neural networks have shown remarkable success, with rapid performance increase on available datasets. However, current 3D semantic segmentation benchmarks contain only a small number of categories – less than 30 for ScanNet and SemanticKITTI, for instance, which are not enough to reflect the diversity of real environments (e.g., semantic image understanding covers hundreds to thousands of classes). Thus, we propose to study a larger vocabulary for 3D semantic segmentation with a new extended benchmark on ScanNet data with 200 class categories, an order of magnitude more than previously studied. This large number of class categories also induces a large natural class imbalance, both of which are

challenging for existing 3D semantic segmentation methods. To learn more robust 3D features in this context, we propose a language-driven pre-training method to encourage learned 3D features that might have limited training examples to lie close to their pre-trained text embeddings. Extensive experiments show that our approach consistently outperforms state-of-the-art 3D pre-training for 3D semantic segmentation on our proposed benchmark (+9% relative mIoU), including limited-data scenarios with +25% relative mIoU using only 5% annotations.

Keywords: 3D semantic scene understanding, 3D semantic segmentation, 3D representation learning, language + 3D vision

1 Introduction

In recent years, remarkable advances have been made in 3D semantic segmentation as a core task underlying 3D perception for myriad applications, including robotics, autonomous navigation, and mixed reality. The introduction of several large-scale real-world 3D datasets [10,4,1] has led to rapid developments in data-driven 3D deep learning techniques, with an emphasis on point- and sparse-voxel-based approaches [13,44,9,47,16]. However, popular benchmarks such as ScanNet [10] or SemanticKITTI [1] focus on a limited number of class categories (20 and 28 classes, respectively), and thus these label sets do not well-represent the diversity and complexity of real scene content that would be encountered in the wild. In contrast, common image segmentation benchmarks [12,27] contain over 80 annotated class labels, with recent large-vocabulary image challenges [15] presenting over 1000 categories for recognition tasks.

Thus, we propose to address a larger-vocabulary setting for 3D semantic segmentation. In particular, we focus on the indoor domain and consider 3D scans of ScanNet [10] where a variety of different object categories are seen in the RGB-D scans despite its benchmark evaluating on only 20 classes. We present ScanNet200, a 200-class 3D semantic segmentation benchmark, considering an order of magnitude more class annotations than previously considered. This new set of classes includes both finer-grained categories of previous classes as well as a large number of previously unaddressed classes. This induces a much more challenging setting reflecting the naturally observed semantic classes already seen in the raw ScanNet RGB-D observations, where the data also reflects naturally encountered class imbalances (e.g., walls and floors are seen much more often than nightstands, which are also seen far more often than fire extinguishers). In addition considering the setting where all dense annotations are available for train scenes for the 200 classes, we also consider limited annotation scenarios with only sparse annotations per scene, given the expense of 3D data annotation.

In order to address this challenging new benchmark for 3D semantic segmentation, we explore standard techniques for data and loss balancing for the much larger number of class categories. In combination with the most effective techniques, we further observe that, unlike the limited, imbalanced geometric

content, state-of-the-art language models have observed and attained rich representations of all categories, and so can induce a better structure onto learned 3D embeddings. Thus, we propose to ground 3D feature learning with strong pre-trained CLIP text features to construct a richly-structured 3D feature representation space. To this end, we formulate a language-grounded pre-training by mapping learned 3D features to pre-trained language embeddings with a contrastive loss. This enables a more robust 3D representation learning under imbalanced and limited 3D observations. Experiments on our ScanNet200 semantic segmentation as well as semantic segmentation in the limited data regime demonstrate the effectiveness of our language-grounded 3D semantic segmentation.
In summary, our contributions are:

- We propose a new 200-class 3D semantic segmentation benchmark on real-world 3D ScanNet scene data, considering an order of magnitude more category annotation labels than existing 3D semantic segmentation benchmarks.
- In order to guide the construction of robust 3D semantic feature representations for this challenging task, we propose to align geometric feature extraction to the category embedding of the CLIP pretrained language model. This results in improved performance both overall and in the rarely seen, including in the limited-data regime.

2 Related Work

3D Semantic Segmentation. With the introduction of large-scale annotated real-world 3D datasets [10,4,1], 3D semantic segmentation has seen significant focus in recent years with various deep learning-based methods developed around different 3D representations. Early works tackled 3D semantic segmentation on dense volumetric grids [10,11], but were limited in cubic growth in memory and compute. The introduction of PointNet [38] presented a point-based alternative with strong memory efficiency by operating on unstructured point clouds, with various methods introducing local operators to better learn neighborhood structures [39,44,48]. Hierarchical grid structures such as octrees provided a more structured alternative for grid-based reasoning without dense memory consumption [42]. Recently, the introduction of sparse 3D convolutions [13,9] enabled significant performance improvements by leveraging a structured space representation in a sparse fashion to operate efficiently at high resolutions. In this work, we also adopt a sparse 3D convolutional backbone to explore language-guided pre-training for larger-vocabulary semantic segmentation.

3D Representation Learning. Inspired by the success of contrastive frameworks for 2D image representation learning [35,5,17,6], 3D representation learning has begun to see exploration in unsupervised contrastive pre-training. PointContrast [46] demonstrated the effectiveness of unsupervised contrastive pre-training for 3D scene understanding tasks, with various methods introducing augmentation alternatives for 3D pre-training [41,20,7,52]. Contrastive Scene

Contexts [18] introduced an unsupervised contrastive pre-training in the context of data-efficient 3D scene understanding with limited reconstruction and limited annotations available. In contrast to these 3D pre-training methods, we propose a supervised multi-modal 3D representation learning guided by text encoded features to learn a more robust feature representation space covering significantly more class categories than previously studied for 3D. Inspired by the data-efficient scene understanding of [18], we additionally study a limited annotations scenario for our ScanNet200 benchmark.

Additionally, Mix3D [34] presented a data augmentation scheme to mix multiple 3D scenes together to generate semantic segmentation that is more robust against undesired context biases. Our instance-based sampling when fine-tuning the learned language-guided 3D features is inspired by the scene mixing, but operates at an instance level to help mitigate class imbalances. Previous methods have also leveraged text embeddings in 3D learning for zero-shot pointcloud segmentation [31,8] and classification [51]. More recently, CLIP [40] was shown as a powerful conditioner for generative 3D models [43,30]. We also aim to leverage powerful CLIP text embeddings for robust 3D semantic pre-training.

3D Scene Understanding Benchmarks. Recently, various large-scale real-world 3D scene understanding benchmarks have been established. Early benchmarks such as the NYUv2 dataset [33] introduced RGB-D frame-based annotations on a limited number of frames (e.g., 1449 for NYUv2). ScanNet [10] presented a much larger-scale RGB-D dataset and benchmark with 1513 densely-annotated reconstructed 3D scans. While it contains hundreds of raw annotated label data, the ScanNet benchmark evaluates only 20 class categories for its 3D scene understanding tasks. Similarly, Matterport3D [4] presents a large-scale RGB-D dataset with a 20-class semantic segmentation evaluation. Additionally, SemanticKITTI [1] established an outdoor 3D dataset and benchmark for LiDAR-based scene understanding with 28 class category annotations. We present our ScanNet200 benchmark based on ScanNet scene data with an order of magnitude more classes than previous benchmarks.

Class Imbalance. Real-world dataset annotations tend to contain natural class imbalances which can lead to skewed learning of more frequently observed class categories. Despite the lack of study on mitigating class imbalances in 3D, various methods have been presented to address them in 2D image understanding tasks.

In particular, class imbalance in image classification problems is often addressed by oversampling underrepresented categories with strong data augmentation techniques to obtain an evenly-distributed dataset. Various methods have been introduced towards data-sampling-based re-balancing, for instance random oversampling of underrepresented classes [3,49,45], sampling novel poses of known categories [29], undersampling overrepresented classes [32], frequency-based sampling [22], as well as feature-based or generative sampling [37,36,50]. Inspired by such approaches, we propose a 3D instance-based sampling to mitigate class imbalances for 3D semantic segmentation.

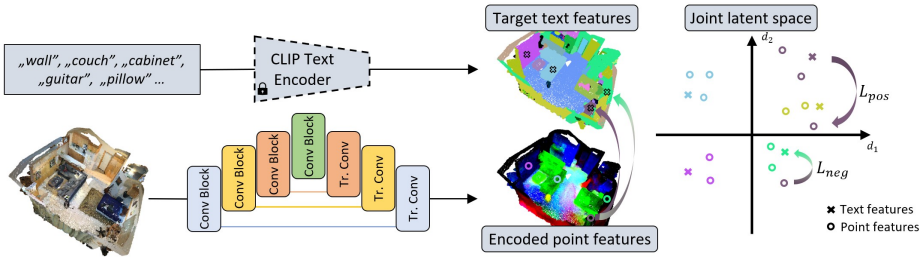


Fig. 2: During pre-training, we guide 3D feature learning by mapping learned features to text encoded anchors of the corresponding semantic labels, constructed by a constrastive loss between text and 3D. This establishes a more robust 3D feature representation space guided by the rich structure of the text embeddings.

Alternative methods have been proposed to re-balance the loss for image understanding tasks [25,19,26]. In particular, the focal loss [26] has been shown to be effective for 2D object detection and semantic segmentation by focusing the training on hard examples or to instance contours [2]. We also study the effect of focal loss balancing for the 3D semantic segmentation task.

3 Method

Our approach tackles the 200-class 3D semantic segmentation task on ScanNet [10] data, exploiting well-structured language models that have trained on rich observations across all category labels. In particular, we leverage pre-trained text embeddings from CLIP [40] as anchors to which we learn to map geometric features during the pre-training step. We then use these language-grounded features for fine-tuning downstream 3D semantic segmentation. During fine-tuning, we further address the class imbalance by instance-based augmentation as well as focal loss-based class-balancing for the downstream loss.

3.1 Language-Grounded 3D Feature Learning

As training data for language-based models are available in far greater quantities than 3D semantic annotations, we propose to ground 3D feature learning to well-structured, pre-trained text encodings. This enables a more robust construction of a learned feature space guided towards a highly-structured, rich text feature space, to support downstream 3D semantic segmentation. An overview of our language-grounded 3D pre-training is shown in Figure 2.

Text Encoder. We leverage a pre-trained CLIP [40] to map semantic labels to text features. Note that our approach is agnostic to the specific language model used, but we found CLIP's multi-modal training is well-suited to our

language-3D pre-training. We refer to the supplemental for additional analysis on alternative text embeddings.

During pre-training, the text encoder is kept fixed, and takes the N_{class} =200 target semantic labels in their text form, tokenizes them, and encodes them to their text encodings to $f_1^t, ..., f_{N_{\text{class}}}^t \in \mathbb{R}^D$, where D is the dimensionality of the text representation space. We leverage the text features f_i^t to anchor learning of 3D features such that learned 3D features will lie close to text encodings if they represent the same semantic class.

3D Encoder. For 3D feature extraction, we employ a state-of-the-art sparse 3D convolutional U-Net [9]. Our 3D encoder backbone takes as input a sparse voxelized 3D scan $\mathcal{S}$, with RGB color as input features, and produces for each sparse voxel location a 3D feature $f_i^s \in \mathbb{R}^D$.

Text-supervised Contrastive optimization. We then train the 3D feature encoder to map to the well-structured space of the text model by formulating a contrastive objective to bring together the different data modalities. For a 3D scan $\mathcal{S}$ with all N_p sparse voxel locations in the current batch, we map together 3D features f_i^s to text features $f_{h(i)}^t$ representing the semantic label text:

$$\mathcal{L}_{pos} = \sum_{i=1}^{N_p} max\left(0, 1 - \frac{f_i^s \cdot f_{h(i)}^t}{|f_i^s| \cdot |f_{h(i)}^t|} - t_{pos}\right), \tag{1}$$

where $h(i)$ is the semantic text label for location i, and t_{pos} is a threshold value for gradient clipping.

Similarly, multiple non-matching semantic text features, sampled from all text semantic labels, are pushed away from the learned features as negatives:

$$\mathcal{L}_{neg} = \sum_{i=1}^{N_p} \frac{1}{|M|} \sum_{j \in M} max\left(0, t_{neg} - 1 + \frac{f_i^s \cdot f_j^t}{|f_i^s| \cdot |f_j^t|}\right), \tag{2}$$

where $M \in N_{\text{class}}$ are a set of semantic label encodings different from i, f_j^t is the corresponding text feature, and t_{neg} is a threshold value for gradient clipping.

We found that a cosine distance between features empirically produced the best results compared to alternative distance measures such as ℓ_1, ℓ_2, or MSE. This allows for more flexibility in the feature learning by constraining only vector directions, and is similarly reflected in CLIP-driven image classification [24,14].

The final language-3D pre-training objective is then:

$$\mathcal{L} = \mathcal{L}_{pos} + \lambda \mathcal{L}_{neg} \tag{3}$$

where λ weights the effect of the multiple negatives with the positive loss. We found empirically that negative sampling was necessary for effective 3D representation learning, rather than employing positive text associations only. During

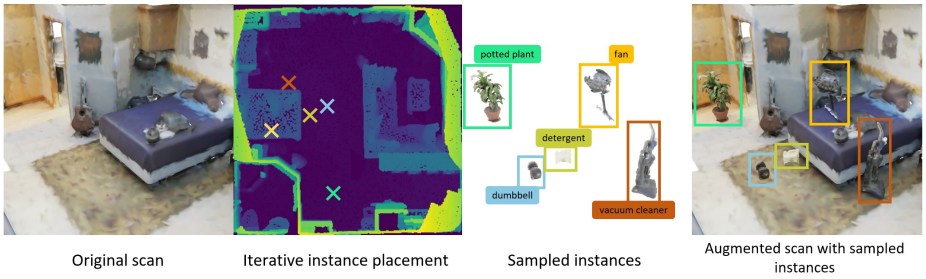


Fig. 3: Our instance sampling augments scenes during training with by placing rarely-seen class category instances into them, breaking unduly specific context dependencies that can be easily learned from only a few examples.

optimization, multiple possible point feature trajectories are converging to the target anchors, and we encourage the solutions that maximize cluster separation at all times (see Sec. 5 for additional analysis). Additionally, as we sample target feature anchors from the complete set of categories, we are able to maximize cluster separation within categories rarely appearing together in the same scenes, in contrast to unsupervised algorithms.

3.2 3D Semantic Segmentation Fine-tuning

We use the language-grounded pre-trained 3D features for fine-tuning for 3D semantic segmentation. Here, we also directly address the inherent class imbalance due to the natural long-tail distribution of the class categories in densely-annotated 3D scenes (e.g., far more walls and floors than lamps or dumbbells). In particular, we address this through data augmentation for class balancing as well as a class-balanced loss.

Class re-balancing by instance sampling. We observe that since rare classes are not only infrequently observed but are often small objects and thus represented by smaller sets of points or voxels, they often overfit to recognizing both the surrounding context and the object. We thus propose to augment scenes by placing instances of infrequently seen class categories in them and breaking overly specific context dependencies for recognition.

An overview of our instance sampling is shown in Figure 3. We obtain instances from ScanNet200 semantic instance annotations, and sample from instances of rare class categories from train scenes. We note here, that we relied on the available ScanNet instance annotations, but since we are augmenting long tail categories only, sparsely appearing in all scenes, the conversion from semantic to instance segmentations comes essentially free with surface label clustering. We place these sampled instances in potentially physically valid locations in a new scene. To this end, we compute a height map of the scene in which the object is to be inserted and iteratively sample instance centroid candidates where the

new object can be placed. Any sampled object center where the inserted object would collide with existing objects, based on bounding box overlap, is discarded. For all accepted placements we update the height map and continue with the iterations until the condition on the number of samples is met. This enables class re-balancing by over-sampling rare categories and breaking unduly specific context dependencies for recognition. For additional implementation details please refer to Section 8 in our supplemental material.

Class-balanced loss. As instance sampling-based data augmentation will not fully balance classes (e.g., walls, floors, and other frequently seen categories still dominate), we also consider the class balancing of the loss function. Rather than a standard cross entropy loss for semantic segmentation, we adapt a focal loss [26] which was shown to be effective in mitigating class imbalance effects for 2D object detection. The focal loss applies a dynamic weighting factor based on the usefulness of a given sample to re-weight the cross entropy, focusing on difficult-to-classify examples.
In particular, the focal loss proposes a modulating factor for a cross entropy loss:

$$\mathcal{L}_{\text{focal}}(p_t) = -(1-p_t)^{\gamma} log(p_t), \tag{4}$$

where p_t is the point prediction probability for the respective target label and $\gamma \geq 0$ is focusing the modulating factor $(1-p_t)^{\gamma}$.

In practice, we did not see a direct improvement over cross entropy training by applying a focal loss directly, so we additionally re-balance the loss based on the class imbalance of the train set:

$$FL(p_t) = -\alpha(1-p_t)^{\gamma} log(p_t), \qquad \alpha_i = \frac{log(n_i)}{\sum_{j=1}^{N_{\text{class}}} log(n_j)} \tag{5}$$

By explicitly considering category imbalances, we found this to provide improved performance over both a standard focal loss or direct category-balanced cross entropy (c.f. Sec 5 for more analysis).

3.3 Implementation Details

During pre-training, we use a sparse 3D U-Net backbone for 3D feature encoding, implemented with the MinkowskiEngine [9]. We adapt the MinkUNet34 to output feature dimension maps of size $D = 512$ to match the dimensionality of the pre-trained text encoding from CLIP [40]. For additional details on optimization please refer our supplemental at Section 7. We follow a two stage training with pretraining and fine-tuning for both semantic and instance segmentation, where for all comparisons we use the same 3D backbone architecture.

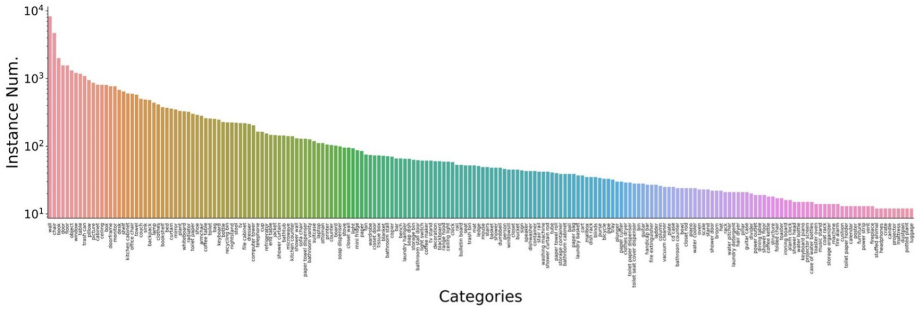


Fig. 4: Class category distribution for our ScanNet200 Benchmark showing number of instances per category; note that the frequencies are given on log-scale and ordered by number of instances per category.

4 ScanNet200 Benchmark

The ScanNet Benchmark[1] has provided an active online benchmark evaluation for 3D semantic segmentation, but only considers 20 class categories, which is insufficient to capture the diversity of many real-world environments. We thus present the ScanNet200 Benchmark for 3D semantic segmentation with 200 class categories, an order of magnitude more than previous. We follow the original train/val/test split of ScanNet [10], while training and evaluating over significantly more class categories. Figure 4 shows the class category distribution for ScanNet200 over the number of annotated instances and the number of annotated surface points per category in the train set.

To obtain the 200 class categories, we considered the raw semantic label annotations provided by ScanNet [10], which contains 607 raw categories. After merging near-duplicate labels, this resulted in 550 unique semantic classes, from which we selected the 200-most represented categories by the number of instances, forming ScanNet200. The 200-class selection enables enforcing a minimum of 10 samples from all categories.

In order to better understand performance under the natural class imbalance of the ScanNet200 benchmark, we further split the 200 categories into sets of 66, 68 and 66 categories, based on the frequency of number of labeled surface points in the train set: *head*, *common* and *tail* respectively. Evaluation over all categories as well as for the head, common, and tail splits enables a more precise understanding of segmentation performance.

Limited Annotation Task. We additionally study semantic segmentation performance on ScanNet200 in the limited annotation regime, as dense 3D annotations are expensive to acquire. In the limited annotation setting, we emulate annotations queried from annotators with a randomly sampled annotated point per

[1] http://kaldir.vc.in.tum.de/scannet_benchmark/

object, and any additional points annotated based on farthest point sampling, similar to settings of weakly-supervised methods [28]. We consider scenarios of (5%, 10%, 50%) of annotated surface points provided, where all scene geometry is available (but unlabeled for surface points without annotations).

Instance Segmentation Task. In addition to 3D semantic segmentation, we also evaluate 3D instance segmentation on ScanNet200. We evaluate methods by mean Average Precision (mAP) at IoU of (25%, 50%) and averaged over all overlaps between $[50\%, 95\%]$ at 5% steps, following the original [10] benchmark.

Evaluation metrics. To evaluate semantic segmentation, we consider several evaluation metrics. The primary evaluation metric is the category-level mean intersection-over-union ($mIoU$) score as $tp/(tp+fp+fn)$, as a commonly adopted segmentation measure. Additionally, we evaluate *precision* as $tp/(tp+fp)$ and *recall* as $tp/(tp+fn)$, to provide further insight towards over-prediction and under-prediction, respectively. All evaluation metrics are measured across head, common, and tail splits as well as globally across all categories, in order to consider performance for more and less frequently seen class categories.

5 Experiments

We evaluate our approach for language-grounded pre-training with state-of-the-art alternatives for 3D semantic segmentation on our ScanNet200 benchmark. For our method and all baselines, we use the same 80M parameter sparse 3D U-Net backbone implemented with MinkowskiNet [9].

	mIoU				**Precision**				**Recall**			
	Head	Common	Tail	All	Head	Common	Tail	All	Head	Common	Tail	All
Scratch	48.29	19.08	7.86	25.02	68.81	66.29	39.88	58.32	60.45	25.50	15.06	33.67
Ins. samp.	48.46	18.97	9.22	25.49	70.04	62.98	49.41	60.81	59.64	24.66	19.25	34.52
C-Focal	48.10	20.28	9.38	25.86	68.10	65.64	47.43	60.39	60.08	26.28	19.14	35.48
SupCon [23]	48.55	19.17	10.34	26.02	69.52	65.42	40.62	58.52	60.27	26.28	19.14	35.23
CSC [18]	49.43	19.52	10.28	26.41	70.00	67.75	40.78	59.51	61.01	25.75	17.62	34.79
Ours (CLIP only)	50.39	**22.84**	10.10	27.73	71.64	**69.72**	44.47	61.94	62.20	**29.37**	17.35	36.16
Ours	**51.51**	22.68	**12.41**	**28.87**	**72.72**	66.69	**58.30**	**65.90**	**62.50**	29.09	**26.61**	**39.40**

Table 1: Comparison to state of the art on ScanNet200. Our language-grounded 3D feature learning enables improved performance across frequent and infrequently seen categories in comparison with pure data augmentation or loss balancing techniques as well as state-of-the-art 3D pre-training. Our approach achieves over 5% mIoU performance over training from scratch, more than double the performance improvement of CSC [18].

Comparison to the state of the art. We compare with a state-of-the-art pre-training approaches Contrastive Scene Contexts (CSC) [18] and Supervised

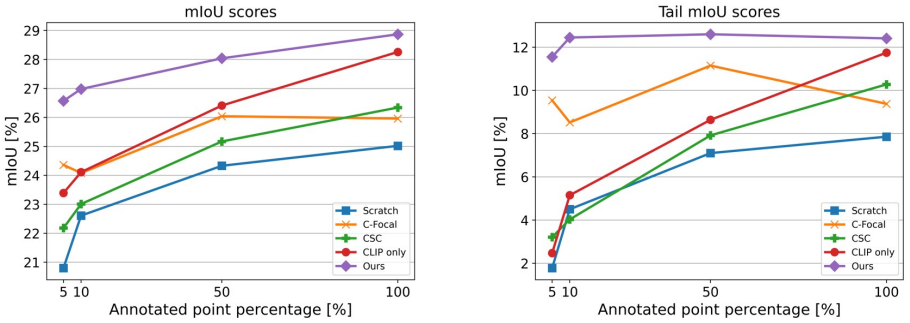


Fig. 5: 3D semantic segmentation under varying amounts of limited annotations. Even when considering only a small number of annotated surface points for our supervised language-guided 3D pre-training, our approach improves notably over the state-of-the-art 3D pre-training of CSC [18].

Contrastive Learning (SupCon) [23], along with our instance-based data balancing and focal loss [26] training in Table 1. For CSC, we use the same pre-training experimental setup as proposed by the authors for our 3D backbone. For SupCon, we sample 5 positive and 5 negative candidates from the training scene for each source point and train it for 300 epochs with the same optimization parameters as our method. Our instance sampling, as well as focal loss, individually help to improve performance, particularly for lesser-seen class categories. Additionally, all pre-training approaches improve performance over training from scratch, while our language-grounded feature learning enables more effective semantic reasoning with consistent improvements over baselines and across common and rarely seen class categories.

Limited annotation semantic segmentation. As data annotations remain expensive to acquire, we additionally evaluate our approach in comparison with state of the art in the limited annotation scenario of our ScanNet200 Benchmark described in Sec. 4. Figure 5 shows performance over varying amounts of labeled annotation data available (5%, 10%, 50%, 100%). Note that since our pre-training leverages text labels to guide pre-training, we only pre-train with the available annotations, whereas CSC is pre-trained with all geometric data available for the train scenes and fine-tuned with the limited annotation data. Our approach enables more robust semantic segmentation on this challenging benchmark, consistently improving and recovering the performance of training from scratch with only 5% of the annotations. Moreover, in the very low annotation regime, we see significant improvements on tail categories, with an increase of +8 mIoU from the state-of-the-art 3D pre-training of CSC with 5% of annotations available.

How much does a class-balanced focal loss help? We evaluate the effect of our class-balanced focal loss [26] variant (*C-Focal*) in Table 1, which helps to improve performance over training from scratch with a standard cross entropy

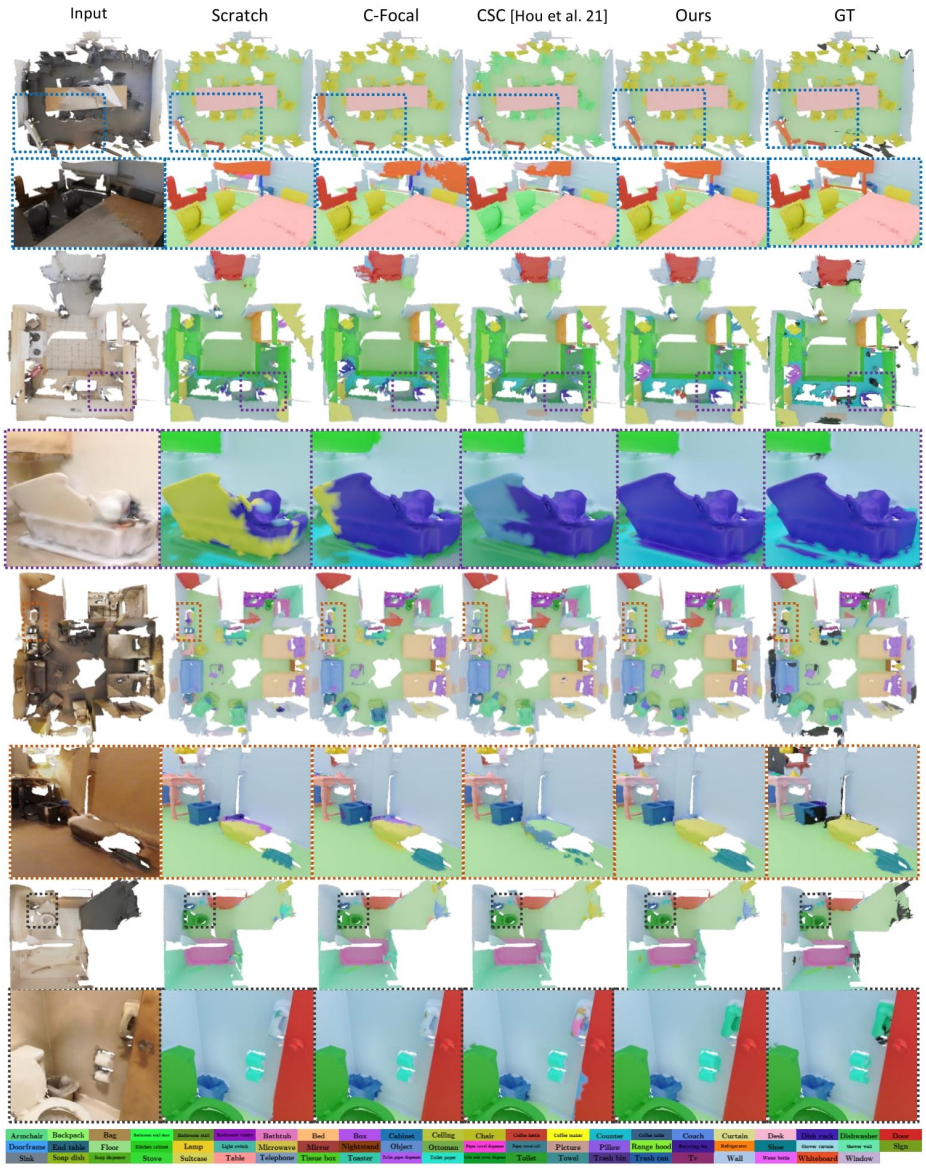


Fig. 6: Qualitative semantic segmentation results on ScanNet [10] scenes. In comparison to training from scratch, class-balance focal loss, and the 3D pre-training of CSC [18], our language-grounded 3D feature learning enables more consistent and accurate semantic segmentation, even for challenging less frequently seen class categories (e.g., *"dish rack"* in row 4, *"telephone"* in the last row).

loss. Additionally, we see a consistent improvement with a smaller 3D backbone model in Table 3 in supplementary material, particularly for tail categories. We note that the class-balanced focal loss improves notably over both the original focal loss formulation (both using $\gamma = 2$), as well as a class-balanced cross entropy.

What is the impact of data balancing with instance sampling? We additionally evaluate the effect of applying data balancing by our instance sampling during training in Table 1 (*Ins. samp*) as well as for a smaller 3D backbone in supplemental Table 3. We find that this instance sampling consistently provides a small improvement in performance across common and rare class categories.

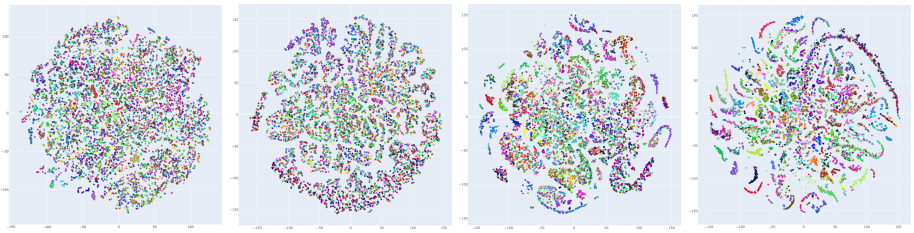
What is the effect of our language-grounded pre-training? Table 1 shows that our language-grounded pretraining to text-based CLIP [40] embeddings without focal loss or instance sampling already improves over all baselines. Our full approach with focal loss and instance sampling in addition to text-anchored pre-training enables consistent, effective improvement in comparison to alternative approaches.

3D instance segmentation task. In addition to 3D semantic segmentation, we also analyze a 3D instance segmentation task in Table 2, showing that our approach generalizes across multiple downstream tasks with consistent performance improvement. We use the same pre-trained 3D backbones and fine-tune them for instance segmentation by predicting an offset vector for every scene point as a voting mechanism together with the semantic labels. These directional distance vectors are optimized during train time, while the clustering of the instances is calculated only at test time. For the task and clustering algorithm, we adopt the paradigms of [21,18] to our ScanNet200 benchmark. For this task, we train our models with a batch size of 8 for 300 epochs and momentum SGD optimizer with the same parameters as in the semantic segmentation experiments, except for a smaller starting learning rate of 0.02.

	Precision	**mIoU**	**mAP@0.5**
Scratch	61.04	25.37	24.47
CSC [18]	63.13	25.92	25.24
CLIP only	64.24	27.58	**27.91**
Ours	**65.32**	**27.72**	26.09

Table 2: 3D instance segmentation, in comparison with training from scratch and state-of-the-art 3D pre-training approach CSC [18]. Our language-grounded pre-training improves over both baselines.

Learned feature representation space. We analyze the pre-trained representation spaces by visualizing a t-SNE projection of the learned features in Figure 7. By anchoring 3D feature learning to a richly-structured text embedding space, we can learn a more structured 3D feature representation space.



(a) CSC [18] (b) SupCon [23] (c) Ours (only pos.) (d) Ours

Fig. 7: We show a comparison with the representation learned by CSC [18], SupCon [23], as well as our approach when training with only positive samples. Our full language-grounded pre-training results in a more structured feature representation space with improved semantic segmentation performance.

Limitations and Future Work. We believe our language-grounded 3D feature learning provides a promising step towards more robust and general 3D scene understanding, though several important limitations remain. It is often the case that infrequently observed objects are small and their geometric resolution is limited, so while tail category performance has improved using only geometric input, there is still much room for improvement. In particular, we note that color image observations could provide significantly higher resolution signals to explore for more accurate tail category recognition. Additionally, text encodings are used to anchor learned 3D feature representations, but currently, only the semantic labels of each object are considered, whereas text caption or object attribute descriptions could potentially provide a richer signal.

6 Conclusion

We have presented ScanNet200, a new benchmark for 3D semantic segmentation with an order of magnitude more class categories, along with a new approach for language-grounded pre-training to address 3D semantic feature learning under imbalanced and limited data. Our approach demonstrates robust feature learning by anchoring learned features to richly-structured CLIP text embeddings, demonstrating consistent improvements over strong baselines on our challenging ScanNet200 Benchmark and under limited annotation scenarios. We believe that this makes an important step towards 3D semantic scene understanding in the wild, and forms a basis for future multi-modal exploration for a variety of 3D perception tasks.

Acknowledgements

This project is funded by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt).

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In: Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV) (2019) 2, 3, 4
2. Biasutti, P., Lepetit, V., Aujol, J.F., Brédif, M., Bugeau, A.: Lu-net: An efficient network for 3d lidar point cloud semantic segmentation based on end-to-end-learned 3d features and u-net. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 5
3. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249–259 (2018) 4
4. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017) 2, 3, 4
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 3
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021) 3
7. Chen, Y., Nießner, M., Dai, A.: 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. arXiv preprint arXiv:2112.02990 (2021) 3
8. Cheraghian, A., Rahman, S., Chowdhury, T.F., Campbell, D., Petersson, L.: Zero-shot learning on 3d point cloud objects and beyond. CoRR **abs/2104.04980** (2021), https://arxiv.org/abs/2104.04980 4
9. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019) 2, 3, 6, 8, 10
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 2, 3, 4, 5, 9, 10, 12
11. Dai, A., Nießner, M.: 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 452–468 (2018) 3
12. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010) 2
13. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. CVPR (2018) 2, 3
14. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Zero-shot detection via vision and language knowledge distillation. arXiv e-prints pp. arXiv–2104 (2021) 6
15. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 2
16. Han, L., Zheng, T., Xu, L., Fang, L.: Occuseg: Occupancy-aware 3d instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2940–2949 (2020) 2

17. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 3
18. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring data-efficient 3d scene understanding with contrastive scene contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15587–15597 (2021) 3, 4, 10, 11, 12, 13, 14
19. Hsieh, T.I., Robb, E., Chen, H.T., Huang, J.B.: Droploss for long-tail instance segmentation. In: AAAI. vol. 3, p. 15 (2021) 5
20. Huang, S., Xie, Y., Zhu, S.C., Zhu, Y.: Spatio-temporal self-supervised representation learning for 3d point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6535–6545 (2021) 3
21. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: Pointgroup: Dual-set point grouping for 3d instance segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 13
22. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: Eighth International Conference on Learning Representations (ICLR) (2020) 4
23. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020) 10, 11, 14
24. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022) 6
25. Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., Feng, J.: Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10991–11000 (2020) 5
26. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 5, 8, 11
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 2
28. Liu, Z., Qi, X., Fu, C.W.: One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1726–1736 (2021) 10
29. Manhardt, F., Kehl, W., Gaidon, A.: Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2069–2078 (2019) 4
30. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. arXiv preprint arXiv:2112.03221 (2021) 4
31. Michele, B., Boulch, A., Puy, G., Bucher, M., Marlet, R.: Generative zero-shot learning for semantic segmentation of 3d point cloud. CoRR **abs/2108.06230** (2021), https://arxiv.org/abs/2108.06230 4
32. More, A.: Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:1608.06048 (2016) 4
33. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012) 4
34. Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F.: Mix3D: Out-of-Context Data Augmentation for 3D Scenes. In: International Conference on 3D Vision (3DV) (2021) 4

35. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018) 3
36. Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y.G., Ding, K., Chen, Z.: Trainable undersampling for class-imbalance learning. In: AAAI (2019) 4
37. Perez-Ortiz, M., Tiňo, P., Mantiuk, R., Hervás-Martínez, C.: Exploiting synthetically generated data with semi-supervised learning for small and imbalanced datasets. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 4715–4722 (2019) 4
38. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017) 3
39. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017) 3
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 4, 5, 8, 13
41. Rao, Y., Liu, B., Wei, Y., Lu, J., Hsieh, C.J., Zhou, J.: Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3283–3292 (2021) 3
42. Riegler, G., Osman Ulusoy, A., Geiger, A.: Octnet: Learning deep 3d representations at high resolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3577–3586 (2017) 3
43. Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C., Fumero, M.: Clip-forge: Towards zero-shot text-to-shape generation. CoRR **abs/2110.02624** (2021), https://arxiv.org/abs/2110.02624 4
44. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6411–6420 (2019) 2, 3
45. Wang, C., Ma, C., Zhu, M., Yang, X.: Pointaugmenting: Cross-modal augmentation for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11794–11803 (2021) 4
46. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O.: Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: European conference on computer vision. pp. 574–591. Springer (2020) 3
47. Xu, J., Zhang, R., Dou, J., Zhu, Y., Sun, J., Pu, S.: Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16024–16033 (2021) 2
48. Xu, Y., Fan, T., Xu, M., Zeng, L., Qiao, Y.: Spidercnn: Deep learning on point sets with parameterized convolutional filters. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 87–102 (2018) 3
49. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors (Basel, Switzerland) **18** (2018) 4
50. Yan, Y., Tan, M., Xu, Y., Cao, J., Ng, M.K., Min, H., Wu, Q.: Oversampling for imbalanced data via optimal transport. In: AAAI (2019) 4

51. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by CLIP. CoRR **abs/2112.02413** (2021), https://arxiv.org/abs/2112.02413 4
52. Zhang, Z., Girdhar, R., Joulin, A., Misra, I.: Self-supervised pretraining of 3d features on any point-cloud. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10252–10263 (2021) 3