Exploring Data-Efficient 3D Scene Understanding with Contrastive Scene Contexts
Supplementary Material

Ji Hou[1] Benjamin Graham[2] Matthiats Nießner[1] Saining Xie[2]
[1]Technical University of Munich [2]Facebook AI Research

In this supplemental document, we describe the details of our implementation in Section 1. We show more visualizations of our models on semantic segmentation and object detection tasks with extremely scarce data for training in Section 2. Detailed per-category results on data-efficient benchmark as well as on full data are showed in Section 3.

1. Implementation Details

Data Preprocessing. Following [7], we subsample the partial frames by every 25 frames. We find pairs of frames within each scene by computing their overlaps. In detail, every single frame is transformed to world coordinates. We iterate every pair of frames to calculate how many points are overlapped by 2.5cm threshold. For example, for each point in frame A, if we can find another point in frame B within 2.5cm in the transformed coordinate system (world), then those 2 points are stored as a correspondence pair. When 2 frames have at least 30% overlaps of points, those 2 frames are saved for training. We save and use both the xyz coordinates and rgb color for pre-training.

PointInfoNCE Loss. Here we explain the details of the PointInfoNCE loss (Equation 3 in the main paper).

$$\mathcal{L}_p = -\sum_{(i,j)\in M} \log \frac{\exp(\mathbf{f}_i^1 \cdot \mathbf{f}_j^2/\tau)}{\sum_{(\cdot,k)\in M, k\in par_p(i)} \exp(\mathbf{f}_i^1 \cdot \mathbf{f}_k^2/\tau)}$$

M denotes the set of all the corresponding matches from two frames. Denote the point features from two frames $\mathbf{f}^1$ and $\mathbf{f}^2$ respectively. In this formulation, we use the points that have at least one match as negative, and non-matched points are discarded. For a matched pair $(i,j) \in M$, point feature $\mathbf{f}_i^1$ serves as the query and $\mathbf{f}_j^2$ serves as the positive key. Point feature $\mathbf{f}_k^2$ where $\exists(\cdot,k) \in M$, $k \in par_p(i)$ and $k \neq j$ are used as the set of negative keys. In practice, we sample a subset of matched pairs from M for training.

Active Labelling. We first use our pre-trained network to make a forward pass on all the voxels of each scene in the training data, and save the 96-dim penultimate layer features at each voxel. Then we back-project the features at each voxel to the raw point cloud using nearest neighbour search. We run a k-means clustering algorithm on the features and xyz coordinates of the point cloud on each scene to get k centroids, where k is the number of points we propose to annotator to label. We run k-means for 50 iterations.

Clustering Algorithm in Instance Segmentation. We adapt the code of breadth first search from PointGroup [5]. Clustering only happens in the test time. In the test time, we cluster on points that are shifted by learned directional and distance vectors. Directional and distance vectors are learned by voting-center loss in the training time. We use 3cm-ball as threshold for every point to search its neighbouring points at each iteration. Within the ball, the points are grouped into one instance when they have the same semantic label. We don't use the ScoreNet proposed in PointGroup, so that we don't have additional network for training. We simply average the scores of semantic prediction of the points belonging to the same instance.

2. More Visualizations

We show more visualizations of semantic segmentation and object detection predictions from our model trained with extremely scarce annotations. We show the semantic segmentation on ScanNet validation set with our model trained on 20 labelled points per scene in Figure 2. We also demonstrate the object detection results on ScanNet validation set predicted by our model trained on 1 bounding box annotated per scene in Figure 1.

3. Per-Category Results

In this section, we demonstrate detailed per-category performance as supplement of data-efficient benchmark. In-

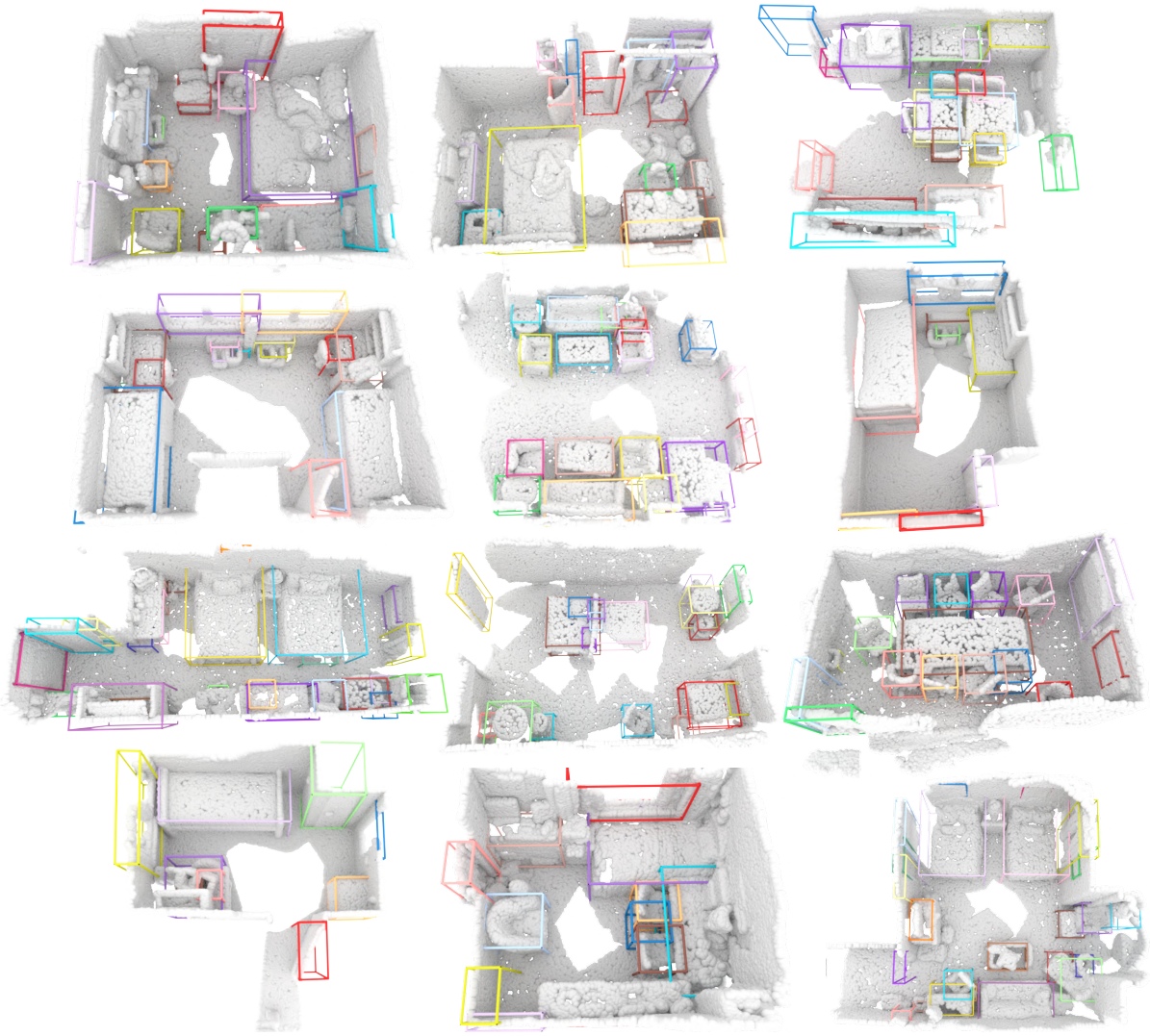

Figure 1: **Object Detection Results (Limited Bounding Box Annotations).** With our pre-trained model as initialization for fine-tuning, our approach generates high-quality detection predictions. Here our model is trained with 1 bounding box annotated per scene.

	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	**avg**
Scratch	31.8	72.4	56.0	52.7	**55.9**	36.6	25.3	47.6	14.7	11.3	10.1	**36.4**	34.5	57.5	90.0	33.7	**80.3**	35.8	43.5
PointContrast	39.2	71.2	**63.1**	**71.4**	48.4	36.9	20.5	45.2	18.2	8.1	13.9	32.4	31.5	**64.1**	97.0	42.3	54.9	**40.1**	44.5
Ours	**43.7**	**75.2**	62.9	65.7	50.5	**43.4**	**27.4**	**52.9**	**26.9**	**19.7**	**14.4**	34.4	**39.9**	61.9	**97.4**	**49.4**	75.3	39.0	**48.9**

Table 1: **Instance Segmentation with Limited Point Annotations (ScanNet-LA)**. We use mAP@0.5 as metric and demonstrate per-category performance over 18 classes on data-efficient benchmark (200 labelled points for training per scene).

	wall	floor	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	**avg**
Scratch	81.6	**96.1**	57.5	79.5	88.1	82.2	67.1	55.9	54.4	76.3	24.3	59.9	52.9	67.9	39.8	55.9	86.9	58.2	82.4	42.1	65.5
PointContrast	83.0	96.0	**61.1**	**79.5**	89.5	81.9	**71.6**	57.1	**57.0**	73.0	22.6	62.0	**58.8**	**69.1**	44.4	63.6	91.5	**59.4**	**85.2**	48.5	67.8
Ours	**84.0**	95.9	60.2	79.0	**89.5**	**83.8**	69.6	**60.2**	56.7	**80.6**	**26.1**	**63.9**	55.6	63.5	**45.1**	**63.7**	**91.9**	56.9	84.7	**52.6**	**68.2**

Table 2: **Semantic Segmentation with Limited Point Annotations (ScanNet-LA)**. We evaluate mean IoU over 20 classes on data-efficient benchmark (200 labelled points per scene for training).

Figure 2: **Semantic Segmentation Results (ScanNet-LA).** With our pre-trained model as initialization for fine-tuning, together with an active labeling process, our approach generates high-quality semantic segmentation predictions. Here our model is fine-tuned with 20 labeled points per scene.

	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	**avg**
Scratch	5.4	**71.9**	64.2	59.8	37.3	17.1	6.8	32.3	0.4	**16.8**	**33.7**	**26.7**	29.2	3.3	**87.9**	20.6	70.2	17.9	33.4
PointContrast	10.3	71.8	**71.1**	61.2	43.1	**21.6**	**9.4**	34.7	**2.3**	6.8	25.7	21.2	32.6	**17.1**	84.1	20.4	74.6	20.0	34.9
Ours	**10.9**	69.5	70.2	**62.1**	44.3	18.2	9.0	**39.8**	1.0	9.2	32.9	25.3	**35.6**	10.3	78.9	**26.5**	**81.0**	**21.2**	**35.9**

Table 3: **Object Detection with Limited Bounding Box Annotations** . We evaluate mAP@0.5 over 18 classes on data-efficient benchmark (7 annotated bounding boxes for training per scene).

	ceiling	floor	wall	beam	column	window	door	chair	table	bookcase	sofa	board	**avg**
Scratch	46.8	89.5	72.5	0.0	**38.2**	72.5	89.5	88.0	39.3	34.7	72.7	85.7	59.3
PointContrast	66.0	**93.0**	73.0	0.0	18.6	72.8	88.3	**91.4**	42.3	**29.5**	63.6	88.0	60.5
Ours	**74.4**	88.0	**76.5**	**0.0**	32.4	**74.6**	**96.4**	91.0	**45.0**	28.8	**63.6**	**90.5**	**63.4**

Table 4: **Instance Segmentation on Stanford Area 5 Test [1]**. We evaluate mAP@0.5 over 12 classes.

	ceiling	floor	wall	beam	column	window	door	chair	table	bookcase	sofa	board	clutter	**avg**
Scratch	91.5	98.6	84.1	0.0	33.0	56.9	63.9	90.1	81.7	72.5	76.5	77.9	59.6	68.2
PointContrast	93.3	**98.7**	85.6	**0.1**	**45.9**	54.4	67.9	91.6	80.1	**74.7**	78.2	81.5	62.3	70.3
Ours	**95.1**	98.4	**86.3**	0.0	40.7	**60.8**	**85.2**	**91.8**	**81.9**	73.9	**78.9**	**82.8**	**62.4**	**72.2**

Table 5: **Semantic Segmentation on Stanford Area 5 Test [1]**. We evaluate mIoU over 13 classes.

	bed	table	sofa	chair	toilet	desk	dresser	night stand	book	bathtub	**avg**
Scratch	47.8	19.6	48.1	54.6	60.0	6.3	15.8	27.3	5.4	32.1	31.7
PointContrast [7]	50.5	19.4	51.8	**54.9**	57.4	**7.5**	**16.2**	37.0	5.9	**47.6**	34.8
Ours	**55.3**	**20.3**	**53.8**	53.6	**65.9**	6.1	15.5	**38.0**	**9.1**	46.5	**36.4**

Table 6: **Object Detection on SUN RGB-D [6]**. We use mAP@0.5 as metric and show per-category AP@0.5 over 10 classes.

	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	**avg**
Scratch	49.0	70.0	87.4	66.5	71.1	47.4	39.6	53.0	30.8	**32.8**	30.8	41.7	48.6	60.1	**99.9**	68.4	75.3	52.4	56.9
PointContrast	49.4	72.1	87.2	**71.7**	67.0	**49.0**	40.7	**57.8**	**35.6**	24.0	30.2	**49.9**	**53.0**	65.2	98.3	61.7	80.5	50.8	58.0
Ours	**50.8**	**74.1**	**88.7**	61.4	**67.2**	48.0	**42.0**	57.0	33.8	32.5	**42.9**	47.4	49.5	**68.9**	98.2	**71.3**	**80.5**	**54.7**	**59.4**

Table 7: **Instance Segmentation on ScanNetV2 [3] Validation Set**. We evaluate the mean average precision with IoU threshold of 0.5 over 18 classes.

	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	**avg**
Scratch	9.9	70.5	70.0	60.5	43.4	**21.8**	10.5	**33.3**	0.8	15.4	33.3	26.6	**39.3**	9.7	74.7	23.7	75.8	18.1	35.4
PointContrast	13.1	**74.7**	**75.4**	**61.3**	44.8	19.8	12.9	32.0	0.9	**21.9**	31.9	**27.0**	32.6	17.5	**87.4**	23.2	80.8	**26.7**	38.0
Ours	**15.1**	74.3	71.9	60.2	**46.4**	21.2	**15.0**	32.5	**1.1**	9.4	**36.6**	21.3	37.3	**47.5**	84.3	**26.2**	**86.8**	21.2	**39.3**

Table 8: **Object Detection on ScanNetV2 Validation Set**. We use mAP@0.5 as metric and show per-category performance over 18 classes.

stance segmentation on ScanNet-LA (Limited Scene Annotations, 200 labelled points for training) is showed in Table 1; semantic segmentation of per-category performance on ScanNet-LA is showed in Table 2; object detection on Limited Bounding Boxes Annotations is showed in Table 3.

We further show the detailed per-category performance as supplement of Table. 6 in the main paper on full data. Instance segmentation and semantic segmentation results on S3DIS are showed in Table 4 and Table 5; object detection on SUN-RGBD result is showed in Table 6; instance segmentation and object detection on ScanNet validation set are showed in Table 7 and Table 8.

4. Different Backbones.

We use Sparse Residual U-Net (SR-UNet-34, also used in [2]) as backbone architecture. 3D-MPA also uses a Sparse Residual U-Net backbone, and the performance gap is due to the additional head modules (e.g., Proposal Consolidation) which is orthogonal to our pre-training method. To show our algorithm is generic and agnostic to the specific backbone, we perform experiments with different backbones, including SR-UNet-18A and PointNet++. Models pre-trained with our method yield significant better results; see Tab. 9.

	Task	Dataset	Backone	mAP@0.5
scratch	ins	S3DIS	SR-UNet-18A	58.6
ours (pre-trained)	ins	S3DIS	SR-UNet-18A	**62.8**
scratch	det	ScanNet	PointNet++	33.5
ours (pre-trained)	det	ScanNet	PointNet++	**39.2**

Table 9: Pre-training with different backbones; 100% of available train data is used; we would expect larger deltas with smaller train set.

5. ScanNet Benchmark

We report validation results to directly compare with PointContrast which also evaluates on the val set. Additionally, we submitted our model to the ScanNet Benchmark (test set); see Tab. 10. Our method significantly outperforms 3D-MPA, despite not leveraging the special 3D-MPA proposal module.

	AP	AP@50	AP@25
3D-MPA [4]	35.5	61.1	73.7
ours (pre-trained)	**40.5**	**64.8**	**79.1**

Table 10: ScanNet **test** set: similar to S3DIS, we outperform 3D-MPA.

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *ICCV*, 2016. 3, 4

[2] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 4

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 4

[4] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *CVPR*, 2020. 4

[5] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In *CVPR*, 2020. 1

[6] Shuran Song and Jianxiong Xiao. Sliding shapes for 3D object detection in depth images. In *ECCV*. 2014. 4

[7] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas J Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3D point cloud understanding. *ECCV*, 2020. 1, 4