Appendix

In this supplemental material, we show additional results on the downstream task of 3D instance segmentation and the extended benchmark of ScanNet200 in Sec. 9. We additionally provide further ablation analysis in Sec. 7 on the effect of the pretrained language model and distance metric used in the contrastive objective of the point-to-language features. Finally, we present a break-down of per-class IoU scores in Sec. 10.

7 Additional Ablations

Generalization across backbone sizes. In Table 3, we evaluate our approach and baselines using a smaller 20M parameter 3D backbone model. In this scenario, we do not map 3D features directly to their 512-dimensional text embeddings from CLIP but a 96-dimensional projection (obtained by PCA). We see consistent improvements from training from scratch as well as over state of the art while using a smaller-backbone architecture.

	mIoU				**Precision**				**Recall**			
	Head	Common	Tail	All	Head	Common	Tail	All	Head	Common	Tail	All
Scratch	45.50	13.64	3.41	20.78	66.89	55.69	23.30	48.62	57.59	19.12	5.83	27.78
Ins. samp.	47.70	14.40	5.33	22.48	69.01	58.71	36.33	54.68	59.95	51.75	13.07	31.59
Weighted CE	46.99	16.72	6.24	23.25	67.33	62.15	35.62	55.03	59.76	23.69	12.58	31.80
Focal [29]	48.31	15.43	4.61	22.78	66.92	62.68	24.95	51.52	58.22	22.86	9.87	30.31
C-Focal	46.47	17.15	**9.45**	24.29	66.27	60.24	38.25	54.92	59.47	23.25	16.27	32.90
CSC [20]	48.51	16.29	5.88	23.49	**70.99**	62.16	29.35	54.16	60.47	21.31	9.46	30.19
Ours	**48.88**	**19.97**	9.03	**25.93**	68.77	**65.36**	**42.33**	**58.82**	**61.34**	**27.75**	**19.11**	**35.49**

Table 3: Generalization across backbone sizes: 3D semantic segmentation with a 20M parameter 3D U-Net backbone on ScanNet200. Our approach maintains consistent improvements over state of the art with this smaller 3D backbone.

Comparison with point-based baselines As an additional ablation we compare out method with point-based state-of-the-art segmentation models capable of processing complete ScanNet scenes at once. We choose RandLa-Net [22] and SCF-Net [14] as baselines and use the official authors implementation and hyper-parameters for both methods. Our method outperforms both approaches with our language-guided pretraining on this challenging large-vocabulary task. The performance evaluated on ScanNet200 can be seen in Table 4.

Effect of contrastive distance metric. Table 5 additionally considers alternative distance metrics of ℓ_1 and ℓ_2 in comparison with our used cosine distance metric. The ℓ_1 and ℓ_2 distances were more challenging to optimize to align to corresponding text embeddings, with cosine distance producing the best performance.

	Head	Common	Tail	Mean
RandLA-Net	35.35	5.15	0.87	15.06
SCF-Net	35.99	5.97	0.38	15.45
Ours	**51.51**	**22.68**	**12.41**	**28.87**

Table 4: Comparison with point-based RandLA-Net and SCF-Net on ScanNet200 semantic segmentation (mIoU).

Effect of the pre-trained language model. In Table 5, we consider alternative language models to CLIP [43]; both BERT [12] and GPT2 [44] are also popular language models trained on large amounts of text data, rather than the image-text training of CLIP.

For text encodings, we use the BERT variant *bert_uncased_L-8_H-512_A-8* from [49], and project the 768-dimensional GPT2 encodings from the small GPT2 model to 512 dimensions by PCA. We find the rich embedding structure from the multi-modal nature of CLIP produces the best results.

	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	33.67	60.45	25.50	15.06	33.67
GPT2 [44]	45.70	19.07	9.73	24.78	69.42	66.18	48.76		56.86	25.132	16.75	32.41
BERT [12]	47.70	18.19	9.16	24.95	70.48	64.09	42.16		58.24	24.01	16.19	32.65
BERT ℓ_2	41.16	14.02	8.89	21.28	67.82	61.65	39.38		53.90	19.86	13.72	20.77
CLIP [43] ℓ_1	39.28	10.26	2.80	17.38	64.52	57.64	27.57		48.81	14.09	3.95	22.29
CLIP ℓ_2	43.48	16.97	8.87	23.04	67.15	64.64	41.54		54.50	22.07	14.77	29.92
CLIP	**50.39**	**22.84**	**10.10**	**27.73**	**71.64**	**69.72**	**44.47**	**42.67**	**62.20**	**29.37**	**17.35**	**36.16**

Table 5: Ablation study on different language models for generating the text anchors during the pre-training stage. We show that while the model benefited from pretraining guided by all language models, CLIP was found to be the most suitable for this task. We also show that more rigid loss distance metrics such as $l1$ or $l2$ can even significantly hinder the performance.

8 Implementation Details

Training parameters In the pretraining stage, we use a momentum SGD optimizer with batch size 8 and an initial learning rate of 0.05, decayed by a factor of 0.3 at epochs 150 and 250, a momentum of 0.9, and train for 400 epochs until convergence. We additionally use $\lambda = 1$, $t_{pos} = 0$, $t_{neg} = 0.6$ and $N_i = 3$ uniformly sampled from all ScanNet200 categories.

We then fine-tune the pre-trained 3D backbone for 3D semantic segmentation. We optimize with the same momentum SGD and batch size, with an initial learning rate of 0.05, decayed by a factor of 0.3, and train for 170 epochs until

convergence. For the instance sampling, we sampled from the 66 classes least frequently represented in the training set surface point annotations.

Instance Sampling For the instance sampling, we randomly sample from the 66 tail classes, and select them by probability computed from the inverse log frequencies of the train set histogram. Object centers are placed in the scene by randomly sampling (x, y) locations in the scene, with z determined by the max height of the scene geometry at (x, y) (such that the object sits on the scene geometry). Objects are then inserted with a random orientation around the up (z) axis. Any object insertion that induces a bounding box collision with scene objects or previously inserted objects are discarded. Placement determined by object center on the support plane encourages placement of instances with sufficient physical support in the scene. In combination with color and geometry augmentation, this helps to learn more robust features in our large-vocabulary setting.

We also note here that while we do not explicitly address lighting effects during instance augmentation, a minor appearance difference is noticeable on the original and sampled parts of the scene, we still achieve a clear effect with this augmentation technique. We hypothesize that at the 2cm resolution which our method (and state-of-the-art methods) use, lighting inconsistencies have a limited-to-negligible effect. In addition, we use random color jittering, which reduces the chance of learning from erroneous signal, and observe that the advantage of this geometric augmentation provides notably more benefit than color. We provide evidence of our reasoning with an experiment to train the same baseline sparse 20M parameter UNet model with and without voxel color signal, and observed that the final performance differs only in 1% mIoU.

9 3D Instance Segmentation Results

Figure 8 shows qualitative visualizations for the downstream task of 3D instance segmentation in comparison to training from scratch and CSC [20].

10 Breakdown of Class IoU scores

Table 6 shows the per-category IoU scores for 3D semantic segmentation on ScanNet200.

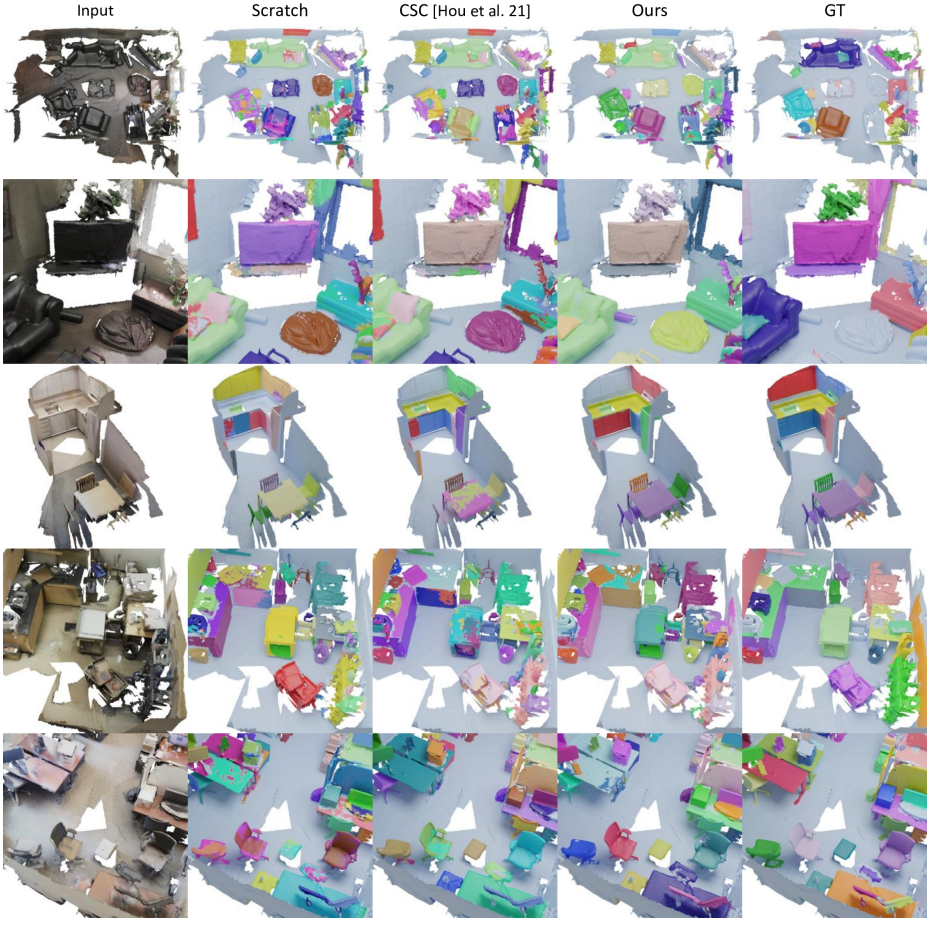

Fig. 8: Qualitative results for 3D semantic instance segmentation results on ScanNet [10] scenes. Our language-grounded pretraining together with class-balanced losses can also effectively improve performance in object recognition.

	Class IoU						
Categories	**Scratch**	**Inst. Sampl.**	**C-Focal**	**CSC**	**CSC + C-Focal**	**Ours - CE**	**Ours**
wall	**77.25**	**77.05**	**76.02**	**77.64**	**76.71**	**78.15**	**78.66**
chair	75.06	76.22	75.84	72.59	75.85	71.44	**76.54**
floor	95.41	95.24	95.19	**95.45**	95.25	95.41	95.42
table	64.21	**66.77**	65.32	65.23	65.67	64.68	64.92
door	62.57	63.52	59.91	62.15	59.11	**64.35**	62.78
couch	80.96	82.82	81.72	80.60	79.21	82.18	**83.23**
cabinet	31.73	34.56	32.82	34.78	32.84	**36.74**	34.24
shelf	**44.79**	37.31	39.48	38.87	39.40	43.37	38.30
desk	57.34	57.75	57.78	57.78	56.62	**59.78**	58.90
office chair	1.12	4.57	5.56	16.75	3.75	**18.43**	1.16
bed	79.04	78.60	78.79	77.80	78.97	**79.97**	79.78
pillow	49.36	48.88	49.70	52.24	51.47	**53.13**	51.93
sink	63.43	63.73	64.17	63.71	64.21	**64.80**	64.62
picture	29.53	27.49	25.53	31.03	32.20	30.71	**33.63**
window	54.05	55.95	53.69	54.91	55.66	**59.47**	59.40
toilet	88.24	88.92	90.20	89.53	89.86	**92.78**	92.26
bookshelf	**53.84**	48.79	50.93	51.05	48.23	50.32	52.04
monitor	80.20	79.93	80.85	83.71	81.96	84.94	**85.81**
curtain	70.93	66.02	68.70	**71.04**	68.47	70.23	69.90
book	37.53	**40.49**	39.39	20.42	32.46	39.50	26.01
armchair	47.67	52.60	**57.08**	54.08	50.58	53.05	52.31
coffee table	66.09	66.52	**70.13**	63.88	63.61	64.70	65.04
box	22.25	**24.98**	21.38	22.50	24.65	20.77	23.88
refrigerator	67.48	64.92	56.49	68.95	58.80	**69.89**	65.83
lamp	65.61	60.48	60.96	**69.75**	67.51	69.10	69.02
kitchen cabinet	**67.25**	66.36	63.67	60.97	61.87	62.13	61.39
towel	34.45	40.68	39.90	38.95	40.33	**49.51**	43.77
clothes	13.33	13.55	13.10	15.17	14.78	15.98	**17.20**
tv	80.26	78.17	81.30	85.57	81.30	84.45	**87.15**
nightstand	65.93	**70.83**	64.54	63.64	68.46	68.27	69.17
counter	23.56	24.29	23.95	21.12	21.70	25.78	**26.59**
dresser	37.64	33.10	35.06	32.65	38.64	37.31	**39.46**
stool	36.03	30.45	34.84	43.21	46.42	49.54	**56.44**
cushion	0.00	0.00	0.00	0.00	0.00	0.00	0.00
plant	68.54	**73.90**	68.58	65.83	66.67	71.15	70.80
ceiling	91.50	91.00	90.95	91.53	91.25	**91.70**	91.41
bathtub	81.30	79.60	80.90	78.29	81.58	**81.69**	80.35
end table	21.13	15.36	19.06	15.38	24.01	**24.01**	16.48
dining table	3.91	2.29	5.35	4.43	3.10	**5.65**	0.00
keyboard	25.37	27.32	32.18	33.81	34.60	32.22	**36.64**
bag	**16.12**	12.25	12.89	15.25	9.24	12.69	12.77
backpack	53.09	50.68	50.62	56.73	53.58	57.40	**57.91**
toilet paper	30.46	35.51	33.94	36.52	36.28	39.81	**40.65**
printer	**37.01**	35.76	35.69	29.67	29.06	29.57	29.48
tv stand	62.72	58.32	66.17	69.72	**70.78**	61.92	67.01
whiteboard	42.60	42.39	43.29	45.04	47.14	46.55	**48.66**
blanket	0.46	1.14	0.41	0.61	0.53	0.10	**1.22**
shower curtain	60.80	58.01	61.96	63.04	58.72	**67.85**	65.64
trash can	57.87	56.73	58.29	57.74	54.44	59.26	**61.34**
closet	10.25	7.91	9.19	**15.48**	9.84	7.27	11.18
stairs	55.36	52.20	64.29	49.02	58.42	49.42	**65.48**
microwave	33.50	40.70	37.09	42.85	**45.46**	35.98	40.17
stove	59.17	64.41	59.30	61.99	59.04	**65.18**	65.11
shoe	46.34	38.29	38.85	43.23	38.72	44.52	**47.49**
computer tower	46.52	49.50	50.28	47.87	48.51	**52.96**	52.33
bottle	2.34	6.04	6.87	6.97	8.16	**15.35**	7.14
bin	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ottoman	40.04	52.73	**55.30**	44.97	37.31	44.62	50.21
bench	39.98	41.57	33.65	36.34	38.44	49.73	**52.76**
board	7.83	10.11	7.58	11.91	**16.13**	13.33	11.03
washing machine	79.11	85.10	85.86	81.02	80.65	**86.72**	80.11
mirror	28.55	30.91	32.10	29.13	33.40	35.18	**35.78**
copier	75.21	79.77	83.45	**88.28**	87.01	86.83	88.15
basket	0.00	0.00	0.00	0.00	0.00	0.00	**2.94**
sofa chair	21.73	21.48	27.32	21.82	24.09	**37.55**	36.05
file cabinet	**37.87**	31.78	34.34	29.64	35.37	33.58	32.88
fan	24.39	**31.70**	21.03	27.47	17.05	21.23	31.02
laptop	33.13	34.29	26.23	36.11	41.85	**44.69**	43.09
shower	16.71	20.08	20.80	17.42	20.67	18.93	**32.82**
paper	4.30	3.76	**14.32**	3.79	12.09	2.45	5.84
person	9.98	15.85	18.03	12.09	13.91	17.42	**24.42**
paper towel dispenser	41.18	37.18	32.59	39.04	37.99	37.70	**41.58**
oven	1.48	0.03	0.00	2.86	1.01	**4.91**	0.21
blinds	0.44	1.07	1.42	0.62	3.01	0.53	**3.09**
rack	0.00	0.00	0.00	0.00	0.00	0.00	0.00
plate	0.00	0.00	0.00	0.00	0.00	0.00	0.00
blackboard	48.93	**53.86**	49.19	53.71	48.87	50.76	53.02
piano	38.68	54.40	**58.86**	47.77	48.06	10.02	24.04
suitcase	46.43	46.78	41.31	**61.56**	51.55	57.09	57.84
rail	0.30	**2.58**	1.46	2.57	1.17	1.93	1.15
radiator	50.56	50.60	47.42	52.29	50.40	**55.15**	54.24
recycling bin	32.10	28.50	27.01	34.03	30.09	33.41	**39.05**
container	0.00	9.20	9.53	0.00	**15.11**	0.00	0.00
wardrobe	12.12	9.95	12.63	8.56	11.76	15.69	**24.51**
soap dispenser	61.76	62.38	61.20	56.99	65.05	64.95	**70.48**
telephone	18.38	17.62	21.03	16.54	18.00	20.55	**22.54**
bucket	20.12	20.67	**25.95**	19.55	21.12	24.02	22.04
clock	4.88	9.96	13.88	11.52	18.17	17.51	**26.48**
stand	0.04	**0.90**	0.00	0.00	0.13	0.00	0.00
light	1.91	0.49	3.16	**19.11**	16.25	12.21	7.67
laundry basket	1.42	1.47	0.00	**5.46**	4.35	1.27	0.20
pipe	**3.52**	1.99	1.92	1.17	0.01	1.09	1.98
clothes dryer	14.32	7.40	0.67	41.58	29.33	4.77	**52.12**
guitar	0.00	13.04	0.00	**29.25**	26.05	8.02	1.78
toilet paper holder	0.00	0.00	0.00	0.00	0.00	0.00	**4.69**
seat	0.00	0.00	0.00	0.00	0.00	**1.62**	0.00
speaker	0.00	0.00	0.00	0.00	0.00	0.00	0.00
column	25.80	17.91	12.24	0.00	0.64	**25.89**	0.68

bicycle	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ladder	40.67	33.61	45.94	36.88	47.83	55.96	46.14
bathroom stall	43.90	44.55	45.57	41.21	43.77	48.14	43.56
shower wall	52.94	53.76	50.95	46.90	48.12	43.78	47.90
cup	10.28	12.48	14.42	14.44	16.53	21.00	14.53
jacket	11.18	6.58	10.10	6.16	7.98	9.72	14.45
storage bin	3.46	0.00	0.00	0.88	2.35	4.35	1.04
coffee maker	52.84	34.53	36.58	49.47	57.62	75.29	78.46
dishwasher	5.05	7.17	19.18	11.77	22.50	21.62	28.03
paper towel roll	24.18	19.22	28.64	27.99	31.98	24.73	23.16
machine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mat	6.42	0.00	8.07	0.88	10.44	7.22	10.29
windowsill	14.68	23.71	17.96	16.27	19.98	15.15	21.92
bar	10.21	5.71	12.57	7.74	11.44	13.73	13.26
toaster	1.91	6.50	8.23	0.00	0.14	0.00	0.00
bulletin board	7.95	2.80	10.59	13.68	17.39	19.25	27.59
ironing board	1.33	1.36	12.23	6.51	10.55	0.27	2.02
fireplace	40.93	49.23	60.24	42.96	50.68	47.52	68.43
soap dish	18.11	18.64	23.41	26.71	28.37	28.15	29.71
kitchen counter	50.11	39.74	42.74	32.74	32.14	38.22	42.21
doorframe	34.18	29.46	33.16	36.43	33.90	40.06	40.01
toilet paper dispenser	40.43	31.76	44.93	42.89	37.10	36.93	44.52
mini fridge	18.85	12.48	11.32	8.25	11.57	24.15	17.30
fire extinguisher	3.45	9.75	0.12	8.79	0.37	34.88	41.51
ball	18.74	74.81	35.19	58.87	35.49	2.20	49.23
hat	0.00	0.00	0.00	0.00	0.00	0.00	0.00
shower curtain rod	33.53	29.08	27.35	33.41	27.59	32.87	34.80
water cooler	4.73	11.66	21.04	0.36	0.00	0.00	0.00
paper cutter	18.60	26.97	19.93	23.51	22.60	23.45	21.58
tray	0.00	0.00	0.02	0.00	1.39	0.68	0.00
shower door	13.86	18.79	13.46	14.71	24.77	20.34	23.26
pillar	0.00	0.00	0.10	0.66	0.02	0.29	0.00
ledge	1.83	4.00	6.22	5.15	7.63	9.95	6.86
toaster oven	3.13	15.28	1.39	9.79	18.67	9.96	7.25
mouse	0.00	0.00	0.00	0.00	0.00	0.00	3.18
toilet seat cover dispenser	42.78	45.41	61.57	43.21	62.56	64.32	58.95
furniture	0.00	0.09	0.00	0.00	0.00	0.00	0.00
cart	21.51	29.49	14.34	23.07	33.53	35.27	23.22
storage container	0.00	0.00	0.00	0.00	0.00	0.00	0.00
scale	20.62	10.30	10.56	17.90	17.23	27.05	24.43
tissue box	5.13	3.41	6.26	3.85	10.62	8.94	9.26
light switch	0.00	0.00	0.00	0.00	0.00	0.00	0.00
crate	0.00	0.00	0.00	0.00	0.00	0.00	0.00
power outlet	0.00	0.00	0.00	0.00	0.00	0.00	0.00
decoration	31.19	1.07	1.63	3.03	9.60	18.18	8.42
sign	7.07	6.81	6.73	8.14	5.76	5.09	0.05
projector	2.13	5.28	6.09	0.00	12.94	0.00	18.45
closet door	13.78	6.13	6.91	7.68	15.33	9.57	13.27
vacuum cleaner	27.05	57.37	43.79	55.23	65.48	79.14	51.90
candle	0.00	0.00	0.00	0.00	0.00	0.00	0.00
plunger	24.08	32.55	39.71	30.40	28.56	38.44	26.17
stuffed animal	0.00	0.00	0.00	0.00	0.00	0.00	0.00
headphones	0.00	14.79	0.16	0.00	0.00	0.00	5.50
dish rack	52.57	51.98	41.78	44.19	59.29	72.73	68.41
broom	1.07	3.24	4.84	1.47	0.38	1.64	2.76
guitar case	0.00	0.00	0.00	0.00	0.00	0.00	0.00
range hood	74.72	73.60	74.91	74.28	69.66	75.41	66.40
dustpan	0.54	0.00	1.46	3.90	9.26	0.18	1.78
hair dryer	0.00	0.00	0.67	0.00	2.16	0.00	18.91
water bottle	0.00	0.00	0.00	0.00	0.00	0.00	0.00
handicap bar	4.91	5.24	2.52	4.78	3.54	1.23	4.11
purse	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vent	0.00	0.00	1.45	2.00	27.51	0.00	15.94
shower floor	55.41	18.08	61.75	75.50	45.89	61.17	73.29
water pitcher	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mailbox	32.32	34.76	30.79	37.72	37.64	32.82	42.72
bowl	0.00	1.43	0.00	0.60	1.45	0.00	0.00
paper bag	0.11	5.37	3.51	2.05	1.43	2.30	1.91
alarm clock	0.00	0.00	0.00	0.00	0.00	0.00	0.00
music stand	0.00	0.00	0.00	0.00	0.00	0.00	0.00
projector screen	2.81	1.60	2.01	5.81	0.00	0.25	0.00
divider	40.84	35.55	49.48	43.32	47.47	46.25	53.51
laundry detergent	23.42	8.32	0.00	13.09	0.00	0.61	1.13
bathroom counter	10.11	16.16	17.86	6.25	20.45	22.25	33.43
object	3.51	3.32	4.08	4.75	5.67	5.36	5.11
bathroom vanity	35.99	40.17	40.87	40.70	41.05	50.54	41.94
closet wall	11.00	10.64	12.36	14.75	12.48	8.33	10.74
laundry hamper	10.45	0.85	16.47	8.97	6.63	19.14	23.61
bathroom stall door	59.08	42.73	58.54	55.43	46.60	63.83	58.87
ceiling light	42.51	39.06	38.76	41.53	47.09	49.92	36.99
trash bin	33.77	41.67	33.78	35.63	37.82	39.60	36.98
dumbbell	21.47	33.23	27.16	47.20	57.58	74.11	6.59
stair rail	18.04	23.58	16.66	18.39	15.34	15.69	31.16
tube	3.23	0.07	6.99	0.00	0.19	0.00	4.54
bathroom cabinet	21.93	24.86	23.99	14.25	26.37	18.87	33.12
cd case	0.00	0.00	0.00	0.00	0.00	0.00	0.00
closet rod	25.83	8.50	17.37	19.48	29.02	32.16	19.81
coffee kettle	0.33	0.15	4.56	0.17	14.36	0.00	35.60
structure	0.00	0.00	0.00	0.00	0.00	0.00	0.00
shower head	0.00	20.40	11.80	0.00	14.06	0.11	27.80
keyboard piano	0.00	0.00	0.00	0.15	0.00	0.00	0.01
case of water bottles	0.00	0.00	0.00	0.44	0.00	0.00	0.00
coat rack	0.00	0.00	0.00	0.00	0.00	3.62	0.00
storage organizer	0.00	0.00	0.00	0.00	0.00	0.00	0.00
folded chair	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fire alarm	0.00	0.00	0.00	0.00	0.00	0.00	0.00
power strip	0.00	0.00	0.00	0.00	0.00	0.00	0.00
calendar	0.00	0.00	0.00	0.00	0.00	0.00	0.00
poster	0.00	0.00	0.00	0.00	0.00	0.00	0.00
potted plant	34.20	21.10	85.25	15.78	28.75	46.78	75.46
luggage	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mattress	0.02	0.00	0.00	0.06	0.00	0.00	0.00

Table 6: Class IoU scores on the ScanNet200 benchmark of our proposed method, and compared with other state-of-the-art approaches.