Implicit Surface Contrastive Clustering for LiDAR Point Clouds

– Zaiwei Zhang, Min Bai, Erran Li

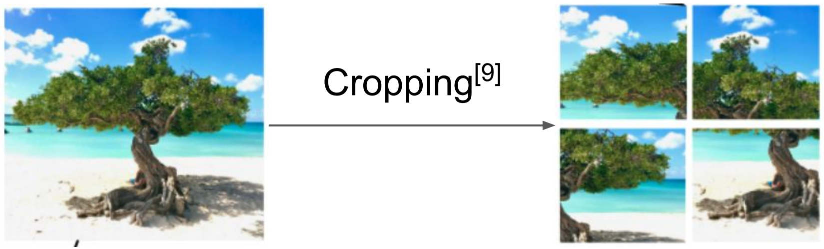
aws machine learning

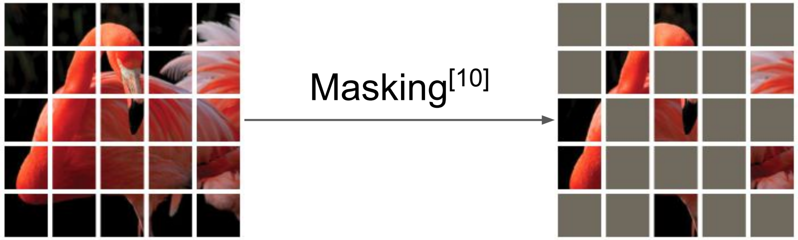
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Session THU-PM-106

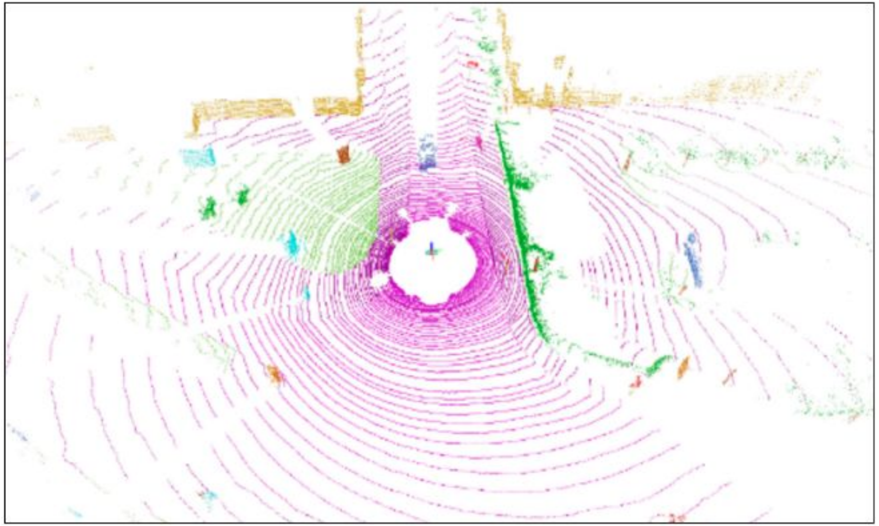
Self Supervised Pretraining on ImageNet

Sample images from ImageNet[5]





Large-scale Point Clouds in 3D



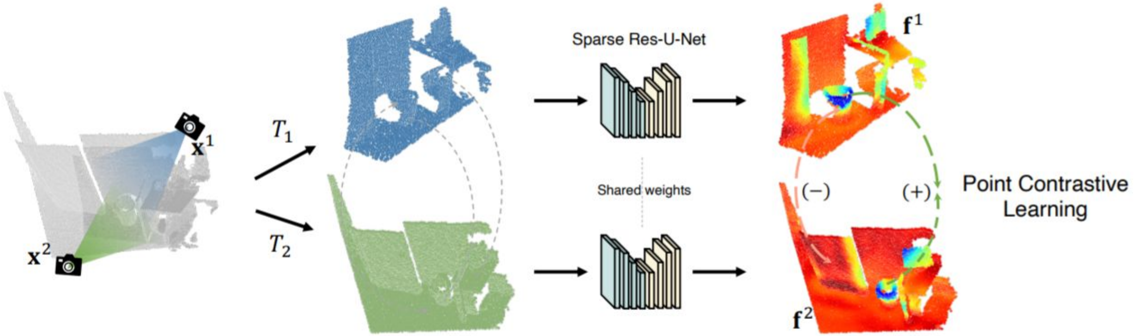
Background points are the majority!

Prior Approaches on Self-supervised Learning in 3D

Most methods focus on single 3D object

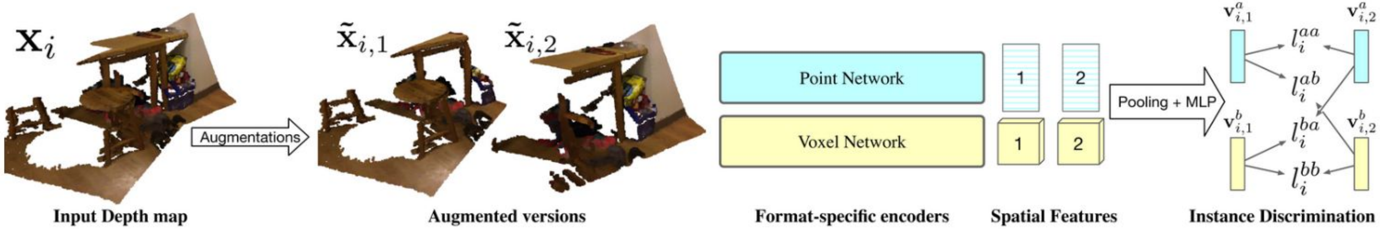
Point level reasoning

- Contrastive learning

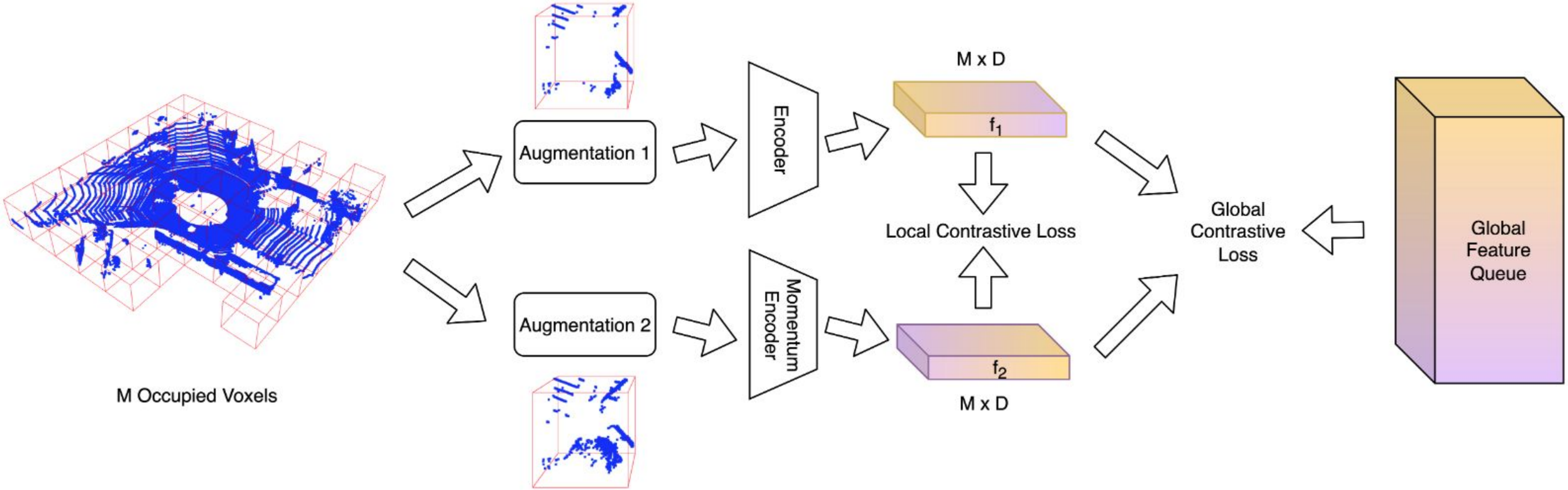


Scene level reasoning:

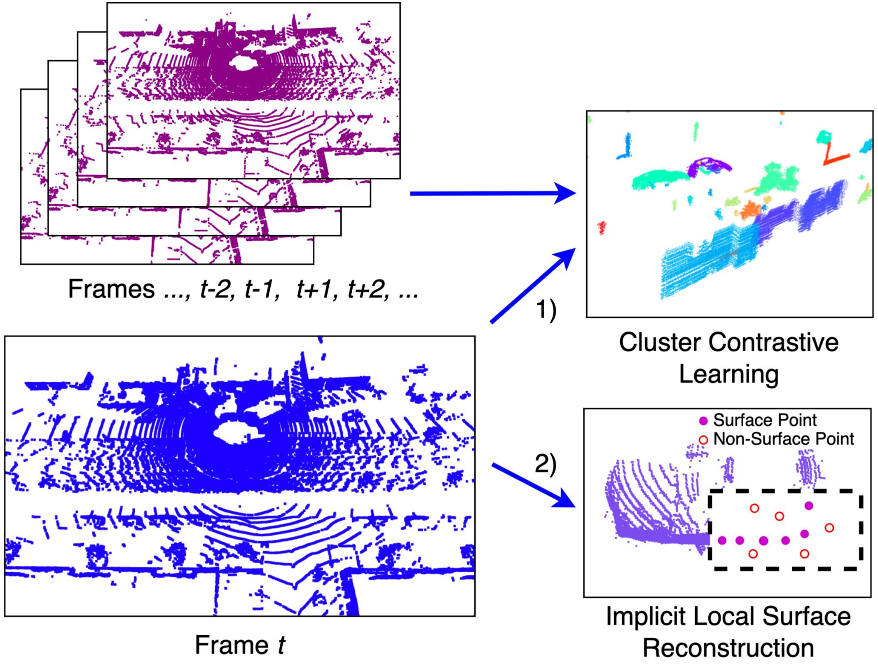
- Contrastive Learning
- Feature regression



Prior Approaches on Self-supervised Learning in 3D



Approach Overview



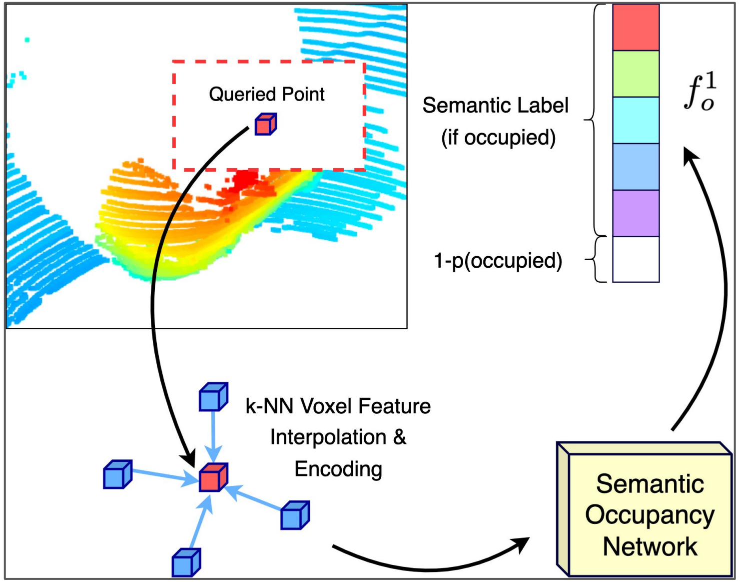
*Global semantic clustering (**what is it?**) and surface reasoning (**what is its shape?**) are complementary.*

Algorithm Outline

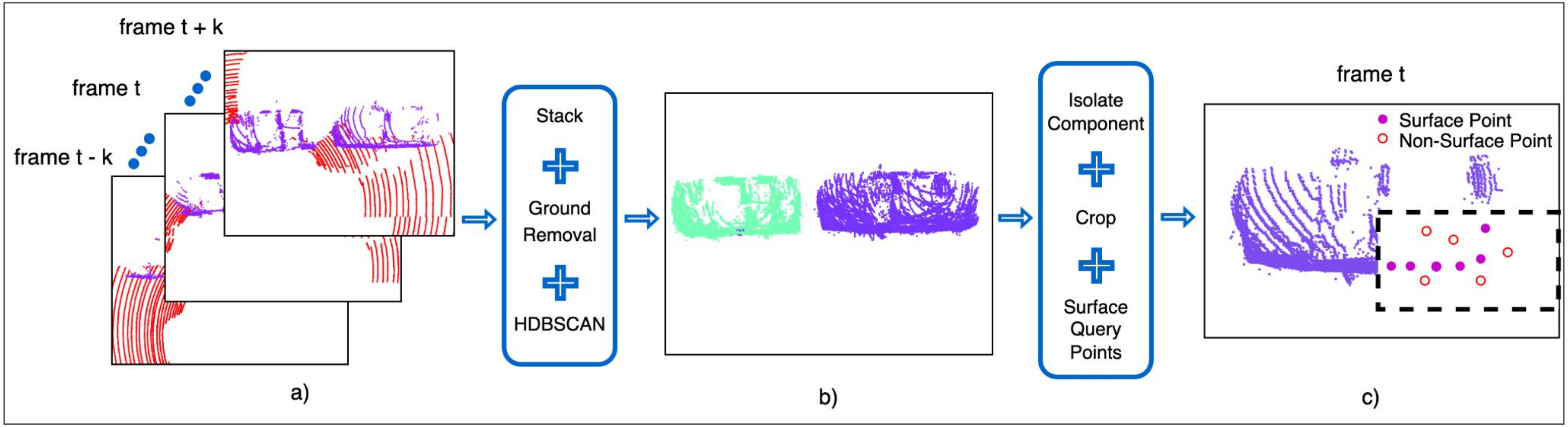
- We adopt teacher-student framework and the teacher network is a momentum encoder.

Input Network encoder F_θ and momentum encoder F_m; Point cloud frames $\mathcal{D} = \{X\}_{i=1}^{N}$; Pre-computed point group labels $\{V\}_{i=1}^{N}$; Global feature queues $F_g \in \mathrm{R}^{C\times d}$;
Output Pre-trained weight for the network encoder F_θ
for x_i in X **do**
- Sample a different frame x_j from D
- Generate two augmented versions $\hat{x}_i$ and $\hat{x}_j$
- For shared point groups V between $\hat{x}_i$ and $\hat{x}_j$, apply random cropping and create query points Q^1 for $\hat{x}_i$
- Feature embeddings: $h^1 = F_\theta(\hat{x}_i), h^2 = F_m(\hat{x}_j)$
- Point group features: $f^1 = \text{avg\_pool}(h^1, V), f^2 = \text{avg\_pool}(h^2, V)$
- Global contrastive clustering loss: $L_c(f^1, f^2, F_g)$
- Local occupancy prediction loss: $L_o(h^1, Q^1)$
- Update F_g with f^2 and update F_θ

end for



Training Supervision Creation



Leverage temporal information and regularities of autonomous driving scenes:

- LiDAR frame stacking + ground points removal
- HDBSCAN clustering → temporal linked group IDs for contrastive clustering task
- Random masking on local object point clusters
- Generate point queries near the target crops for implicit surface reconstruction task

Loss formulation

Contrastive Loss

$$-\sum_{m=1}^{M} y_m \log(p_m^c) + -\sum_{i \in K} \log \frac{\exp(f_i^1 \cdot f_i^2/\tau)}{\sum_{j \in K} \exp(f_i^1 \cdot f_j^2/\tau)}$$

Global Contrastive Loss: consistent semantic clustering across samples

Local Contrastive Loss: form distinct semantic clusters within scene

+

Local Occupancy Prediction Loss

$$L_o(P^o, Z) = -\sum_{m=1}^{K} z_m \log(p_m^o)$$

Experimental Results

Semantic Segmentation Fine-tuning Performance (mIoU) on SemanticKITTI (SK) and Waymo Open Dataset (WOD) with Subset of Labels

	% of SK Used for Fine-Tuning				% of WOD Used for Fine-Tuning			
Self-Supervision Method	1%	2%	5%	10%	1%	2%	5%	10%
No Pre-training	38.9	44.0	51.7	53.4	42.5	45.8	50.4	52.8
PointContrast [44]	41.1(+2.2)	45.0(+1.0)	51.0(-0.7)	52.3(-1.1)	43.8(+1.3)	46.7(+0.9)	49.0(-1.4)	53.4(+0.6)
DepthContrast [49]	39.2(+0.3)	44.7(+0.7)	49.9(-1.8)	52.3(-1.1)	42.7(+0.2)	45.8(+0.0)	50.7(+0.3)	53.0(+0.2)
SegContrast [30]	42.2(+3.3)	45.7(+1.7)	51.0(-0.7)	53.9(+0.5)	43.4(+0.9)	46.2(+0.4)	50.9(+0.5)	53.8(+1.0)
SSPL [48]	42.5(+3.6)	46.4(+2.4)	51.0(-0.7)	53.6(+0.2)	44.8(+2.3)	47.3(+1.5)	51.3(+0.9)	53.5(+0.7)
Ours	**45.1**(+6.2)	**49.0**(+5.0)	**53.0**(+1.3)	**55.2**(+1.8)	**46.0**(+3.5)	**47.9**(+2.1)	**51.7**(+1.3)	**54.1**(+1.3)

Experimental Results

3D Object Detection Fine-tuning Performance on sub-sampled KITTI Dataset (mAP_R11) with subset of labels

	Car (Moderate)				Pedestrian (Moderate)				Cyclist (Moderate)			
Self-Supervision Method	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
No Pre-training	60.2	69.1	74.3	77.8	48.2	**58.8**	59.7	59.2	44.9	57.6	63.3	70.5
PointContrast [44]	62.2	70.6	66.9	77.4	48.6	58.2	58.6	59.1	46.8	58.4	64.6	70.9
DepthContrast [49]	65.0	72.5	77.1	77.7	48.5	55.1	57.1	57.7	51.9	59.6	65.3	71.8
SegContrast [30]	65.4	73.0	77.0	77.9	48.0	57.2	57.6	58.1	50.6	59.3	65.8	72.0
SSPL [48]	63.3	71.1	76.8	76.8	48.1	55.3	57.0	58.2	48.0	58.8	64.2	71.3
Ours	**68.9**	**74.3**	**77.3**	**78.4**	**48.9**	56.5	**59.9**	**59.8**	**53.2**	**60.7**	**69.5**	**73.8**

Ablation Study

Semantic Segmentation Fine-tuning Performance (mIoU) on SemanticKITTI

	1%	2%	5%	10%
None	38.9	44.0	51.7	53.4
occ-only(8)	40.5(+1.6)	45.7(+1.7)	51.4(-0.3)	52.0(-1.4)
occ-only(16)	43.0(+4.1)	**46.1**(+2.1)	**51.9**(+0.2)	54.0(+0.6)
occ-only(32)	**43.6**(+4.7)	46.0(+2.0)	51.7(+0.0)	**54.1**(+0.7)
local-only	41.3(+2.4)	45.9(+1.9)	51.7(+0.0)	53.0(-0.4)
global-only	41.7(+2.8)	45.5(+1.5)	51.8(+0.1)	52.5(-0.9)
full(100)	**43.6**(+4.7)	**47.0**(+3.0)	51.8(+0.1)	53.1(-0.3)
full(200)	43.5(+4.6)	46.1(+2.1)	**51.9**(+0.2)	**53.6**(+0.2)
Ours	**45.1**(+6.2)	**49.0**(+5.0)	**53.0**(+1.3)	**55.2**(+1.8)

Thank you!