4D Contrastive Superflows are Dense 3D Representation Learners

Supplementary Material

Xiang Xu[1,⋆], Lingdong Kong[2,3,*], Hui Shuai[4], Wenwei Zhang[2], Liang Pan[2], Kai Chen[2], Ziwei Liu[5], and Qingshan Liu[4,✉]

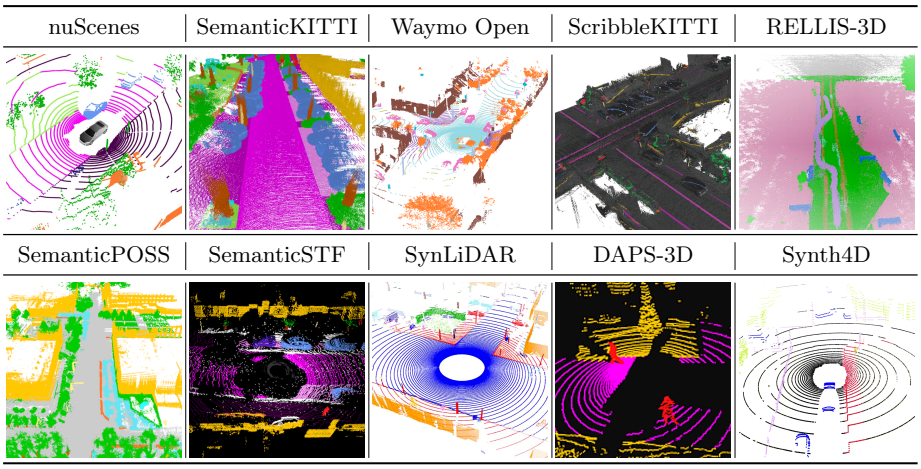
[1] Nanjing University of Aeronautics and Astronautics
[2] Shanghai AI Laboratory
[3] National University of Singapore
[4] Nanjing University of Posts and Telecommunications
[5] S-Lab, Nanyang Technological University

Table of Contents

A Additional Implementation Detail

In this section, we elaborate on additional details regarding the datasets, hyper-parameters, and training/evaluation configuration.

Table A: Summary of datasets used in this work. Our study encompasses a total of 10 datasets in the linear probing and downstream generalization experiments, including [1] *nuScenes* [5], [2] *SemanticKITTI* [1], [3] *Waymo Open* [20], [4] *ScribbleKITTI* [22], [5] *RELLIS-3D* [6], [6] *SemanticPOSS* [14], [7] *SemanticSTF* [24], [8] *SynLiDAR* [23], [9] *DAPS-3D* [7], [10] *Synth4D* [16], and [11] *nuScenes-C* [8]. Images adopted from the original papers.



A.1 Datasets

Pretraining. In this work, we pretrain the model on the *nuScenes* [5] dataset following the data split in SLidR [18]. Specifically, 600 scenes are used as the training set for model pretraining, which is a mini-train split of the whole 700 training scenes. It includes both LiDAR point clouds and six camera image data, from labeled keyframe data to multiple unlabeled sweeps. We conduct spatiotemporal contrastive learning with keyframe data and dense-to-sparse regularization by combining multiple LiDAR sweeps to form dense points.

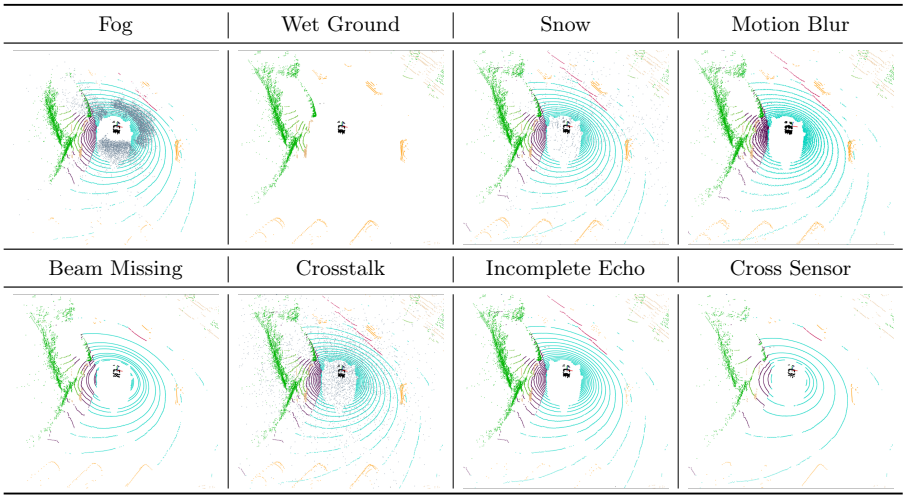
Linear Probing. We train the 3D backbone network with the fixed pretrained backbone on the training set of *nuScenes* [5], and evaluate the performance on the validation set. It consists of 700 training scenes (for 29,130 samples) and 150 validation scenes (for 6,019 samples). Following the conventional setup, the evaluation results are calculated among 16 merged semantic categories.

Downstream Fine-Tuning. To validate the pretraining quality of each self-supervised learning approach, we conduct a comprehensive downstream fine-tuning experiment on the nuScenes [5] dataset, with various configurations. Specifically, we train the 3D backbone network with the pretrained backbone using 1%, 5%, 10%, 25%, and 100% annotated data, respectively, and evaluate the model's performance on the official validation set.

Cross-Domain Fine-Tuning. In this work, we conduct a comprehensive cross-domain fine-tuning experiment on a total of 9 datasets. Tab. A provides a summary of these datasets. Specifically, *SemanticKITTI* [1] and *Waymo Open* [20]

⋆ X. Xu and L. Kong contributed equally to this work. ✉ Corresponding author.

Table B: Examples of the out-of-distribution (OoD) scenarios. Our study encompasses a total of 8 common OoD scenarios in the 3D robustness evaluation experiments, including [1]*fog*, [2]*wet ground*, [3]*snow*, [4]*motion blur*, [5]*beam missing*, [6]*crosstalk*, [7]*incomplete echo*, and [8]*cross sensor*. Images adopted from the Robo3D [8] paper.



contain large-scale LiDAR scans collected from real-world driving scenes, which are acquired by 64-beam LiDAR sensors. We construct the 1% training sample set by sampling every 100 frame from the whole training set. *ScribbleKITTI* [22] shares the same scene with *SemanticKITTI* [1] but are weakly annotated with line scribbles. The total percentage of valid annotated labels is 8.06% compared to fully-supervised methods, while saving about 90% annotation times. *RELLIS-3D* [6] is a multimodal dataset collected in an off-road environment. It contains 13,556 annotated LiDAR scans, which present challenges to class imbalance and environmental topography. *SemanticPOSS* [14] is a small-scale point cloud dataset with rich dynamic instances captured in Peking University. It consists of 6 LiDAR sequences, where sequence 2 is the validation set and the remaining data forms the training set. *SemanticSTF* [24] consists of 2,076 LiDAR scans from various adverse weather conditions, including "snowy", "dense-foggy", "light-foggy", and "rainy" scans. The dataset is split into three sets: 1,326 scans for training, 250 scans for validation, and 500 scans for testing. *SynLiDAR* [23], *Synth4D* [16], and *DAPS-3D* [7] are synthetic datasets captured from various simulators. *SynLiDAR* [23] contains 13 LiDAR sequences with totally 198,396 samples. *Synth4D* [16] includes two subsets and we use Synth4D-nuScenes in this work. It comprises of 20,000 point clouds captured in different scenarios, including town, highway, rural area, and city. *DAPS-3D* includes two subsets and we use DAPS-1, which is semi-synthetic with larger scale in this work. It contains 11 sequences with about 23,000 LiDAR scans.

Out-of-Distribution Robustness Evaluation. In this work, we conduct a comprehensive out-of-distribution (OoD) robustness evaluation experiment on the *nuScenes-C* dataset from the Robo3D [8] benchmark. As shown in Tab. B,

there are a total of 8 OoD scenarios in the *nuScenes-C* dataset, including "fog", "wet ground", "snow", "motion blur", "beam missing", "crosstalk", "incomplete echo", and "cross sensor". Each scenario is further split into three levels ("light", "moderate", "heavy") based on its severity. We test each model on all three levels and report the average results.

A.2 Training Configurations

In this work, we implement the MinkUNet [4] network with the TorchSparse [21] backend as our 3D backbone. The point clouds are partitioned under cylindrical voxels of size 0.10 meter. For the 3D network, point clouds are randomly rotated around the z-axis, flipped along x-axis and y-axis with a 50% probability, and scaled with a factor between 0.95 and 1.05 during pretraining and downstream fine-tuning. For the 2D network, we choose ViT pretrained from DINOV2 [13] with three variants: ViT-S, ViT-B, and ViT-L. The image data are resized to 224×448, and flipped horizontally with a 50% probability during pretraining. For pretraining, we randomly choose 3 camera images as inputs of the 2D network. To enable view consistency alignment, we use the class names as the prompts when generating the semantic superpixels. We train the network with eight GPUs for 50 epochs and the batch size is set to 4 for each GPU. For downstream fine-tuning, we use the same data split as [9] for all datasets. The loss function of segmentation is a combination of cross-entropy loss and Lovász-Softmax loss [2]. We train the segmentation network with four GPUs for 100 epochs and the batch size is set to 2 for each GPU. All the models are trained with the AdamW optimizer [11] and OneCycle scheduler [19]. The learning rate is set as 0.01 and 0.001 for pretraining and fine-tuning, respectively.

A.3 Evaluation Configurations

Following conventions, we report Intersection-over-Union (IoU) for each category i and mean IoU (mIoU) across all categories. IoU can be formulated as follows:

$$\mathrm{IoU}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i} \ , \tag{1}$$

where $\mathrm{TP}_i, \mathrm{FP}_i, \mathrm{FN}_i$ are true positives, false positives, and false negatives for category i, respectively. For robust protocol, we utilize the Corruption Error (CE) and Resilience Rate (RR) metrics, following Robo3D [8], which are defined as follows:

$$\mathrm{CE}_i = \frac{\sum_{j=1}^{3}(1-\mathrm{IoU}_i^j)}{\sum_{j=1}^{3}(1-\mathrm{IoU}_{i_{\mathrm{base}}}^j)} \ , \mathrm{RR}_i = \frac{\sum_{j=1}^{3}(1-\mathrm{IoU}_i^j)}{3\times\mathrm{IoU}_{\mathrm{clean}}} \ , \tag{2}$$

where IoU_i^j is the mIoU calculated at the i-th scene for the j-th level; $\mathrm{IoU}_{i_{\mathrm{base}}}^j$ and $\mathrm{IoU}_{\mathrm{clean}}$ are scores of the baseline model and scores on the "clean" validation set. For a fair comparison with priors, all models are tested without test time augmentation or model ensemble for both linear probing and downstream tasks.

B Additional Quantitative Result

In this section, we supplement the complete results (*i.e.*, the class-wise LiDAR semantic segmentation results) to better support the findings and conclusions drawn in the main body of this paper.

B.1 Class-Wise Linear Probing Results

We present the class-wise IoU scores for the linear probing experiments in Tab. C. We also implement PPKT [10], SLidR [18], and Seal [9] with the distillation of ViT-S, ViT-B, and ViT-L. The results show that SuperFlow outperforms state-of-the-art pretraining methods significantly for most semantic classes. Some notably improved classes are: "barrier", "bus", "traffic cone", and "terrain". Additionally, we observe a consistent trend of performance improvements using larger models for the cross-sensor distillation.

B.2 Class-Wise Fine-Tuning Results

We present the class-wise IoU scores for the 1% fine-tuning experiments in Tab. D. We observe that a holistic improvement brought by SuperFlow compared to state-of-the-art pretraining methods.

C Additional Qualitative Result

In this section, we provide additional qualitative examples to help visually compare different approaches presented in the main body of this paper.

C.1 LiDAR Segmentation Results

We provide additional qualitative assessments in Fig. A, Fig. B, and Fig. C. The results verify again the superiority of SuperFlow over prior pretraining methods.

C.2 Cosine Similarity Results

We provide additional cosine similarity maps in Fig. D and Fig. E. The results consistently verify the efficacy of SuperFlow in learning meaningful representations during flow-based spatiotemporal contrastive learning.

D Limitation and Discussion

In this section, we elaborate on the limitations and potential negative societal impact of this work.

D.1 Potential Limitations

Although SuperFlow holistically improves the image-to-LiDAR self-supervised learning efficacy, there are still rooms for further explorations.

Dynamic Objects. As shown in Fig. F, dynamic objects across frames may share different superpixels due to variant scales in the images. The objects across frames will be regarded as negative samples, which will cause "temporal conflict" under temporal contrastive learning.

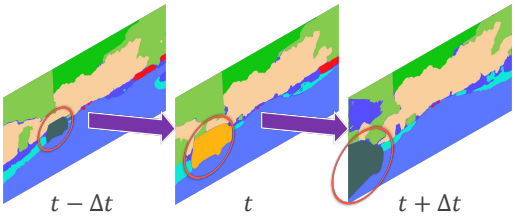


Fig. F: Possible temporal conflicts.

Mis-Align Between LiDAR and Cameras. The calibration parameters between LiDAR and cameras are not perfect as they work at different frequencies. This will cause possible misalignment between superpoints and superpixels, especially when using dense point clouds to distill sparse point clouds. This also restricts the scalability to form much denser points from sweep points.

D.2 Potential Societal Impact

LiDAR systems can capture detailed 3D images of environments, potentially including private spaces or sensitive information. If not properly managed, this could lead to privacy intrusions, as individuals might be identifiable from the data collected, especially when combined with other data sources. Additionally, dependence on automated systems that use LiDAR semantic segmentation could lead to overreliance and trust in technology, potentially causing safety issues if the systems fail or make incorrect decisions. This is particularly critical in applications involving human safety.

E Public Resources Used

In this section, we acknowledge the use of the following public resources, during the course of this work.

E.1 Public Codebase Used

We acknowledge the use of the following public codebase during this work:

- MMCV[6] ..... Apache License 2.0
- MMDetection[7] ..... Apache License 2.0
- MMDetection3D[8] ..... Apache License 2.0
- MMEngine[9] ..... Apache License 2.0
- MMPreTrain[10] ..... Apache License 2.0
- OpenPCSeg[11] ..... Apache License 2.0

[6] `https://github.com/open-mmlab/mmcv`.
[7] `https://github.com/open-mmlab/mmdetection`.
[8] `https://github.com/open-mmlab/mmdetection3d`.
[9] `https://github.com/open-mmlab/mmengine`.
[10] `https://github.com/open-mmlab/mmpretrain`.
[11] `https://github.com/PJLab-ADG/OpenPCSeg`.

E.2 Public Datasets Used

We acknowledge the use of the following public datasets during this work:

- nuScenes[12] ..... CC BY-NC-SA 4.0
- nuScenes-devkit[13] ..... Apache License 2.0
- SemanticKITTI[14] ..... CC BY-NC-SA 4.0
- SemanticKITTI-API[15] ..... MIT License
- WaymoOpenDataset[16] ..... Waymo Dataset License
- Synth4D[17] ..... GPL-3.0 License
- ScribbleKITTI[18] ..... Unknown
- RELLIS-3D[19] ..... CC BY-NC-SA 3.0
- SemanticPOSS[20] ..... CC BY-NC-SA 3.0
- SemanticSTF[21] ..... CC BY-NC-SA 4.0
- SynthLiDAR[22] ..... MIT License
- DAPS-3D[23] ..... MIT License
- Robo3D[24] ..... CC BY-NC-SA 4.0

E.3 Public Implementations Used

We acknowledge the use of the following implementations during this work:

- SLidR[25] ..... Apache License 2.0
- DINOv2[26] ..... Apache License 2.0
- Segment-Any-Point-Cloud[27] ..... CC BY-NC-SA 4.0
- OpenSeeD[28] ..... Apache License 2.0
- torchsparse[29] ..... MIT License

[12] https://www.nuscenes.org/nuscenes.
[13] https://github.com/nutonomy/nuscenes-devkit.
[14] http://semantic-kitti.org.
[15] https://github.com/PRBonn/semantic-kitti-api.
[16] https://waymo.com/open.
[17] https://github.com/saltoricristiano/gipso-sfouda.
[18] https://github.com/ouenal/scribblekitti.
[19] https://github.com/unmannedlab/RELLIS-3D.
[20] http://www.poss.pku.edu.cn/semanticposs.html.
[21] https://github.com/xiaoaoran/SemanticSTF.
[22] https://github.com/xiaoaoran/SynLiDAR.
[23] https://github.com/subake/DAPS3D.
[24] https://github.com/ldkong1205/Robo3D.
[25] https://github.com/valeoai/SLidR.
[26] https://github.com/facebookresearch/dinov2.
[27] https://github.com/youquanl/Segment-Any-Point-Cloud.
[28] https://github.com/IDEA-Research/OpenSeeD.
[29] https://github.com/mit-han-lab/torchsparse.

Table C: The **per-class IoU scores** of state-of-the-art pretraining methods pre-trained and linear-probed on the *nuScenes* [5] dataset. All IoU scores are given in percentage (%). The **best** IoU scores in each configuration are shaded with colors.

Method	mIoU	barrier	bicycle	bus	car	construction vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	driveable surface	other flat	sidewalk	terrain	manmade	vegetation
Random	8.1	0.5	0.0	0.0	3.9	0.0	0.0	0.0	6.4	0.0	3.9	59.6	0.0	0.1	16.2	30.6	12.0
• Distill: None																	
PointContrast [25]	21.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DepthContrast [27]	22.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ALSO [3]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BEVContrast [17]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
• Distill: ResNet-50																	
PPKT [10]	35.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SLidR [18]	39.2	44.2	0.0	30.8	60.2	15.1	22.4	47.2	27.7	16.3	34.3	80.6	21.8	35.2	48.1	71.0	71.9
ST-SLidR [12]	40.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TriCC [15]	38.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Seal [9]	45.0	54.7	5.9	30.6	61.7	18.9	28.8	48.1	31.0	22.1	39.5	83.8	35.4	46.7	56.9	74.7	74.7
HVDistill [26]	39.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
• Distill: ViT-S																	
PPKT [10]	38.6	43.8	0.0	31.2	53.1	15.2	0.0	42.2	16.5	18.3	33.7	79.1	37.2	45.2	52.7	75.6	74.3
SLidR [18]	44.7	45.0	8.2	34.8	58.6	23.4	40.2	43.8	19.0	22.9	40.9	82.7	38.3	47.6	53.9	77.8	77.9
Seal [9]	45.2	48.9	8.4	30.7	68.1	17.5	37.7	57.7	17.9	20.9	40.4	83.8	36.6	44.2	54.5	76.2	79.3
SuperFlow	46.4	49.8	6.8	45.9	63.4	18.5	31.0	60.3	28.1	25.4	47.4	86.2	38.4	47.4	56.7	74.9	77.8
• Distill: ViT-B																	
PPKT [10]	40.0	29.6	0.0	30.7	55.8	6.3	22.4	56.7	18.1	24.3	42.7	82.3	33.2	45.1	53.4	71.3	75.7
SLidR [18]	45.4	46.7	7.8	46.5	58.7	23.9	34.0	47.8	17.1	23.7	41.7	83.4	39.4	47.0	54.6	76.6	77.8
Seal [9]	46.6	49.3	8.2	35.1	70.8	22.1	41.7	57.4	15.2	21.6	42.6	84.5	38.1	46.8	55.4	77.2	79.5
SuperFlow	47.7	45.8	12.4	52.6	67.9	17.2	40.8	59.5	25.4	21.0	47.6	85.8	37.2	48.4	56.6	76.2	78.2
• Distill: ViT-L																	
PPKT [10]	41.6	30.5	0.0	32.0	57.3	8.7	24.0	58.1	19.5	24.9	44.1	83.1	34.5	45.9	55.4	72.5	76.4
SLidR [18]	45.7	46.9	6.9	44.9	60.8	22.7	40.6	44.7	17.4	23.0	40.4	83.6	39.9	47.8	55.2	78.1	78.3
Seal [9]	46.8	53.1	6.9	35.0	65.0	22.0	46.1	59.2	16.2	23.0	41.8	84.7	35.8	46.6	55.5	78.4	79.8
SuperFlow	48.0	54.1	14.9	47.6	65.9	23.4	46.5	56.9	27.5	20.7	44.4	84.8	39.2	47.4	58.0	76.0	79.2

Table D: The **per-class IoU scores** of state-of-the-art pretraining methods pre-trained and fine-tuned on *nuScenes* [5] with 1% annotations. All IoU scores are given in percentage (%). The **best** IoU scores in each configuration are shaded with colors.

Method	mIoU	barrier	bicycle	bus	car	construction vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	driveable surface	other flat	sidewalk	terrain	manmade	vegetation
Random	30.3	0.0	0.0	8.1	65.0	0.1	6.6	21.0	9.0	9.3	25.8	89.5	14.8	41.7	48.7	72.4	73.3
• Distill: None																	
PointContrast [25]	32.5	0.0	1.0	5.6	67.4	0.0	3.3	31.6	5.6	12.1	30.8	91.7	21.9	48.4	50.8	75.0	74.6
DepthContrast [27]	31.7	0.0	0.6	6.5	64.7	0.2	5.1	29.0	9.5	12.1	29.9	90.3	17.8	44.4	49.5	73.5	74.0
ALSO [3]	37.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BEVContrast [17]	37.9	0.0	1.3	32.6	74.3	1.1	0.9	41.3	8.1	24.1	40.9	89.8	36.2	44.0	52.1	79.9	79.7
• Distill: ResNet-50																	
PPKT [10]	37.8	0.0	2.2	20.7	75.4	1.2	13.2	45.6	8.5	17.5	38.4	92.5	19.2	52.3	56.8	80.1	80.9
SLidR [18]	38.8	0.0	1.8	15.4	73.1	1.9	19.9	47.2	17.1	14.5	34.5	92.0	27.1	53.6	61.0	79.8	82.3
ST-SLidR [12]	40.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TriCC [15]	41.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Seal [9]	45.8	0.0	9.4	32.6	77.5	10.4	28.0	53.0	25.0	30.9	49.7	94.0	33.7	60.1	59.6	83.9	83.4
HVDistill [26]	42.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
• Distill: ViT-S																	
PPKT [10]	40.6	0.0	0.0	25.2	73.5	9.1	6.9	51.4	8.6	11.3	31.1	93.2	41.7	58.3	64.0	82.0	82.6
SLidR [18]	41.2	0.0	0.0	26.6	72.0	12.4	15.8	51.4	22.9	11.7	35.3	92.9	36.3	58.7	63.6	81.2	82.3
Seal [9]	44.3	20.0	0.0	19.4	74.7	10.6	45.7	60.3	29.2	17.4	38.1	93.2	26.0	58.8	64.5	81.9	81.9
SuperFlow	47.8	38.2	1.8	25.8	79.0	15.3	43.6	60.3	0.0	28.4	55.4	93.7	28.8	59.1	59.9	83.5	83.1
• Distill: ViT-B																	
PPKT [10]	40.9	0.0	0.0	24.5	73.5	12.2	7.0	51.0	13.5	15.4	36.3	93.1	40.4	59.2	63.5	81.7	82.2
SLidR [18]	41.6	0.0	0.0	26.7	73.4	10.3	16.9	51.3	23.3	12.7	38.1	93.0	37.7	58.8	63.4	81.6	82.7
Seal [9]	46.0	43.0	0.0	26.7	81.3	9.9	41.3	56.2	0.0	21.7	51.6	93.6	42.3	62.8	64.7	82.6	82.7
SuperFlow	48.1	39.1	0.9	30.0	80.7	10.3	47.1	59.5	5.1	27.6	55.4	93.7	29.1	61.1	63.5	82.7	83.6
• Distill: ViT-L																	
PPKT [10]	42.1	0.0	0.0	24.4	78.8	15.1	9.2	54.2	14.3	12.9	39.1	92.9	37.8	59.8	64.9	82.3	83.6
SLidR [18]	42.8	0.0	0.0	23.9	78.8	15.2	20.9	55.0	28.0	17.4	41.4	92.2	41.2	58.0	64.0	81.8	82.7
Seal [9]	46.3	41.8	0.0	23.8	81.4	17.7	46.3	58.6	0.0	23.4	54.7	93.8	41.4	62.5	65.0	83.8	83.8
SuperFlow	50.0	44.5	0.9	22.4	80.8	17.1	50.2	60.9	21.0	25.1	55.1	93.9	35.8	61.5	62.6	83.7	83.7

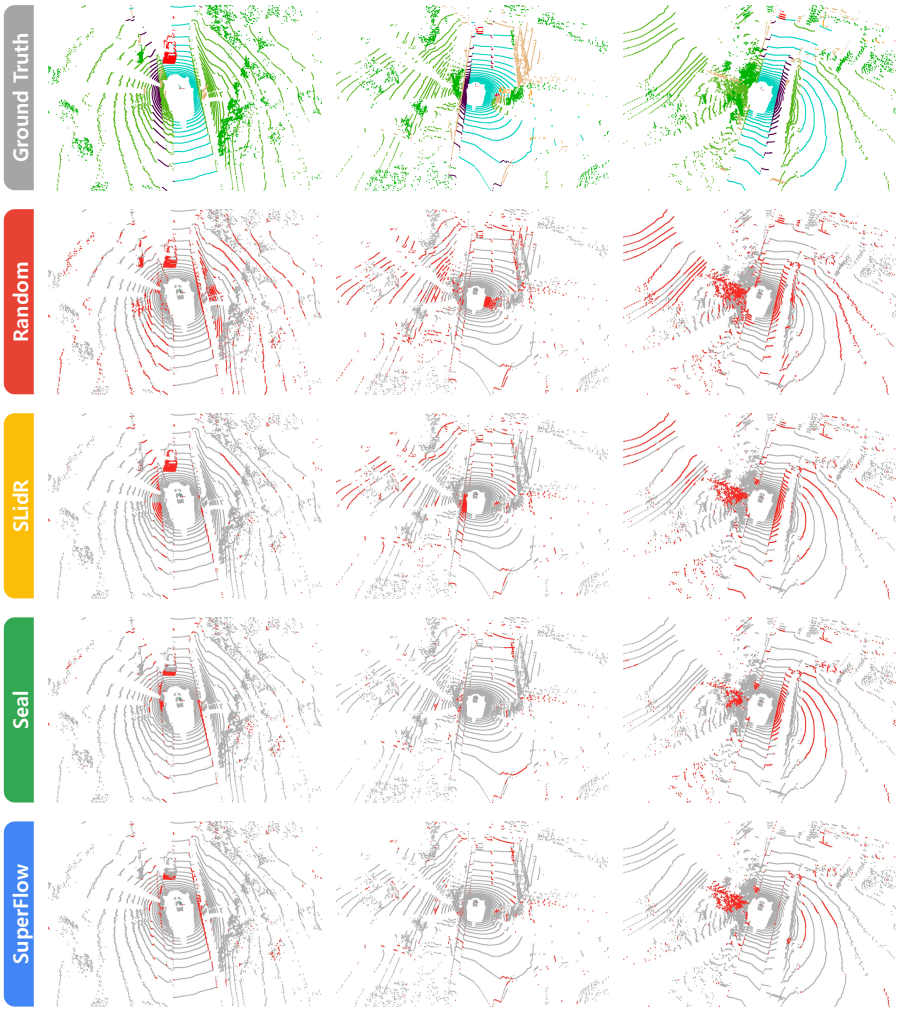


Fig. A: Qualitative assessments of state-of-the-art pretraining methods pretrained on *nuScenes* [5] and fine-tuned on *nuScenes* [5] with 1% annotations. The error maps show the correct and incorrect predictions in gray and red, respectively. Best viewed in colors and zoomed-in for details.

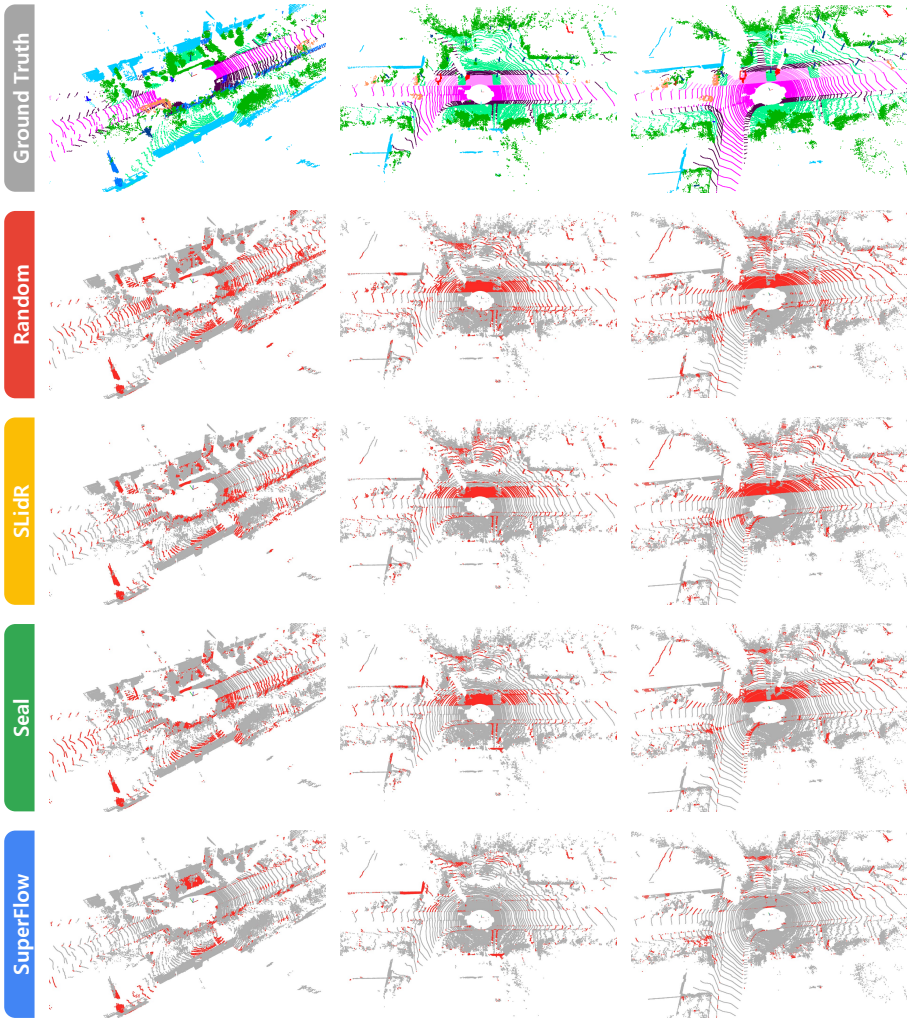


Fig. B: Qualitative assessments of state-of-the-art pretraining methods pretrained on *nuScenes* [5] and fine-tuned on *SemanticKITTI* [1] with 1% annotations. The error maps show the correct and incorrect predictions in gray and red, respectively. Best viewed in colors and zoomed-in for details.

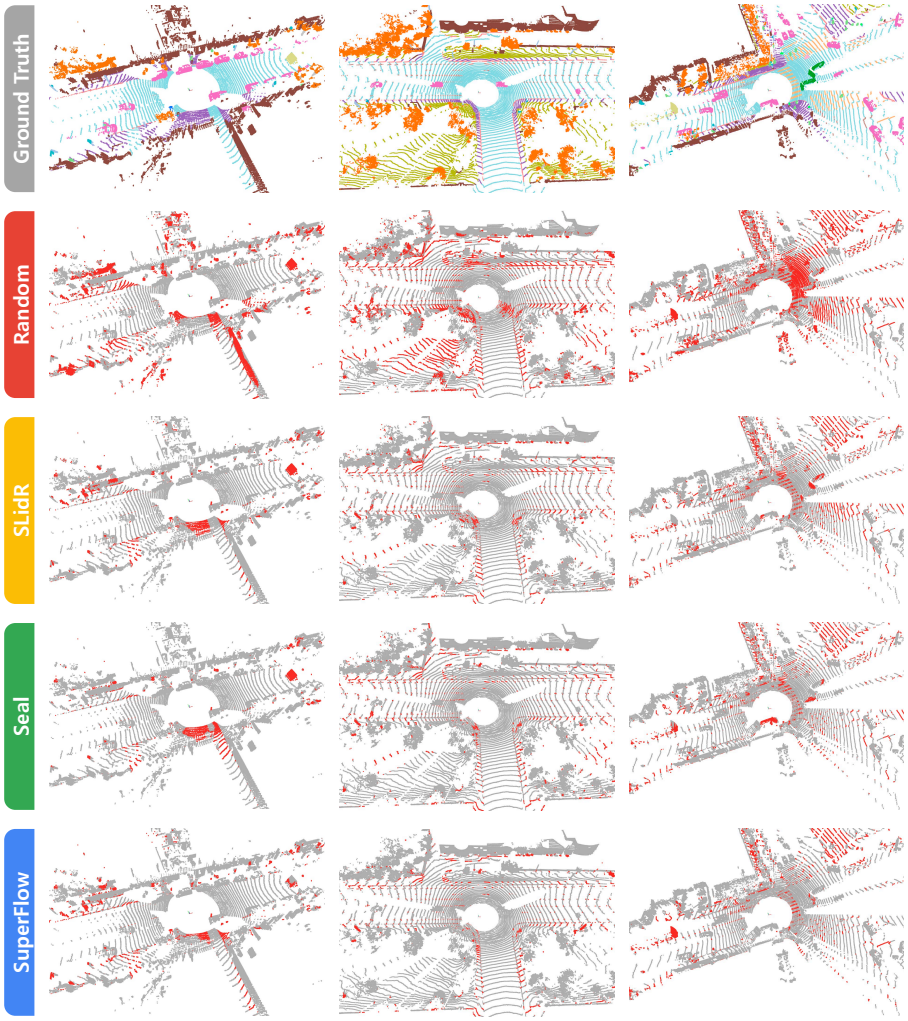
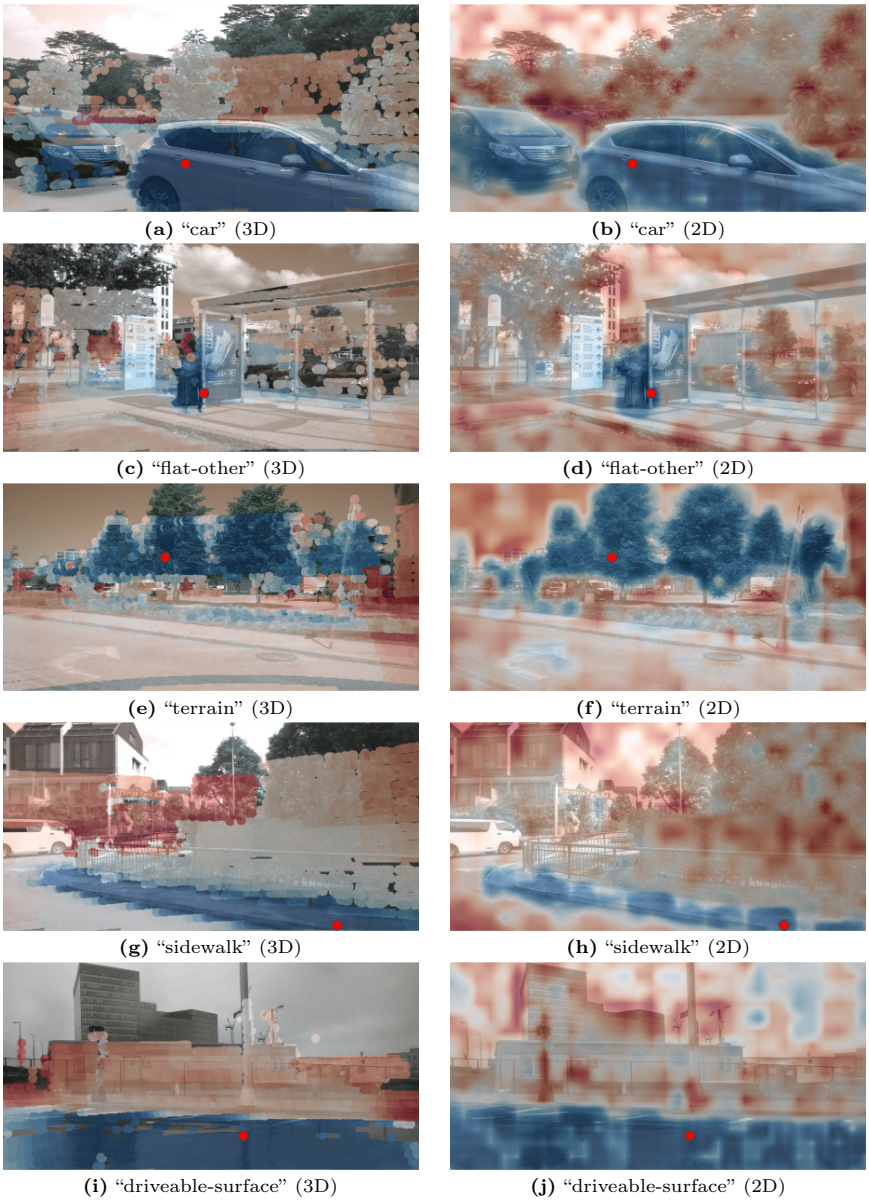


Fig. C: Qualitative assessments of state-of-the-art pretraining methods pretrained on *nuScenes* [5] and fine-tuned on *Waymo Open* [20] with 1% annotations. The error maps show the correct and incorrect predictions in gray and red, respectively. Best viewed in colors and zoomed-in for details.



(a) "car" (3D) **(b)** "car" (2D)

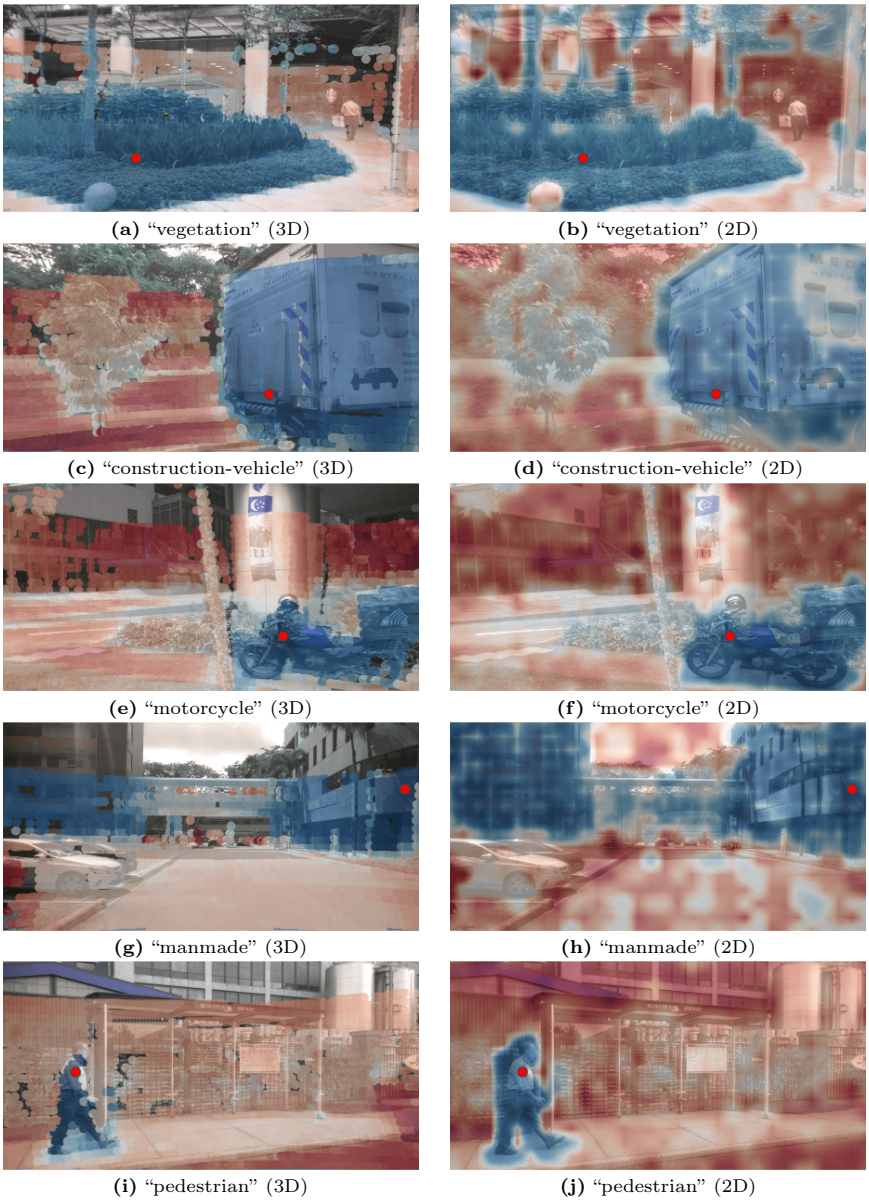
(c) "flat-other" (3D) **(d)** "flat-other" (2D)

(e) "terrain" (3D) **(f)** "terrain" (2D)

(g) "sidewalk" (3D) **(h)** "sidewalk" (2D)

(i) "driveable-surface" (3D) **(j)** "driveable-surface" (2D)

Fig. D: Cosine similarity between the features of a query point (denoted as a red dot) and the features of other points projected in the image (the left column), and the features of an image with the same scene (the right column). The color goes from red to blue denoting low and high similarity scores, respectively.



(a) "vegetation" (3D) **(b)** "vegetation" (2D)

(c) "construction-vehicle" (3D) **(d)** "construction-vehicle" (2D)

(e) "motorcycle" (3D) **(f)** "motorcycle" (2D)

(g) "manmade" (3D) **(h)** "manmade" (2D)

(i) "pedestrian" (3D) **(j)** "pedestrian" (2D)

Fig. E: Cosine similarity between the features of a query point (denoted as a red dot) and the features of other points projected in the image (the left column), and the features of an image with the same scene (the right column). The color goes from red to blue denoting low and high similarity scores, respectively.

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: IEEE/CVF International Conference on Computer Vision. pp. 9297–9307 (2019)
2. Berman, M., Triki, A.R., Blaschko, M.B.: The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4413–4421 (2018)
3. Boulch, A., Sautier, C., Michele, B., Puy, G., Marlet, R.: Also: Automotive lidar self-supervision by occupancy estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13455–13465 (2023)
4. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019)
5. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. IEEE Robotics and Automation Letters **7**, 3795–3802 (2022)
6. Jiang, P., Osteen, P., Wigness, M., Saripallig, S.: Rellis-3d dataset: Data, benchmarks and analysis. In: IEEE International Conference on Robotics and Automation. pp. 1110–1116 (2021)
7. Klokov, A., Pak, D.U., Khorin, A., Yudin, D., Kochiev, L., Luchinskiy, V., Bezuglyj, V.: Daps3d: Domain adaptive projective segmentation of 3d lidar point clouds. IEEE Access **11**, 79341–79356 (2023)
8. Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robo3d: Towards robust and reliable 3d perception against corruptions. In: IEEE/CVF International Conference on Computer Vision. pp. 19994–20006 (2023)
9. Liu, Y., Kong, L., Cen, J., Chen, R., Zhang, W., Pan, L., Chen, K., Liu, Z.: Segment any point cloud sequences by distilling vision foundation models. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
10. Liu, Y.C., Huang, Y.K., Chiang, H.Y., Su, H.T., Liu, Z.Y., Chen, C.T., Tseng, C.Y., Hsu, W.H.: Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. arXiv preprint arXiv:2104.04687 (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
12. Mahmoud, A., Hu, J.S., Kuai, T., Harakeh, A., Paull, L., Waslander, S.L.: Self-supervised image-to-point distillation via semantically tolerant contrastive loss. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7102–7110 (2023)
13. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
14. Pan, Y., Gao, B., Mei, J., Geng, S., Li, C., Zhao, H.: Semanticposs: A point cloud dataset with large quantity of dynamic instances. In: IEEE Intelligent Vehicles Symposium. pp. 687–693 (2020)
15. Pang, B., Xia, H., Lu, C.: Unsupervised 3d point cloud representation learning by triangle constrained contrast for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5229–5239 (2023)

16. Saltori, C., Krivosheev, E., Lathuiliére, S., Sebe, N., Galasso, F., Fiameni, G., Ricci, E., Poiesi, F.: Gipso: Geometrically informed propagation for online adaptation in 3d lidar segmentation. In: European Conference on Computer Vision. pp. 567–585 (2022)
17. Sautier, C., Puy, G., Boulch, A., Marlet, R., Lepetit, V.: Bevcontrast: Self-supervision in bev space for automotive lidar point clouds. arXiv preprint arXiv:2310.17281 (2023)
18. Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R.: Image-to-lidar self-supervised distillation for autonomous driving data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9891–9901 (2022)
19. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. arXiv preprint arXiv:1708.07120 (2017)
20. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2446–2454 (2020)
21. Tang, H., Liu, Z., Li, X., Lin, Y., Han, S.: Torchsparse: Efficient point cloud inference engine. Proceedings of Machine Learning and Systems **4**, 302–315 (2022)
22. Unal, O., Dai, D., Gool, L.V.: Scribble-supervised lidar semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2697–2707 (2022)
23. Xiao, A., Huang, J., Guan, D., Zhan, F., Lu, S.: Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In: AAAI Conference on Artificial Intelligence. pp. 2795–2803 (2022)
24. Xiao, A., Huang, J., Xuan, W., Ren, R., Liu, K., Guan, D., Saddik, A.E., Lu, S., Xing, E.: 3d semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9382–9392 (2023)
25. 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 (2020)
26. Zhang, S., Deng, J., Bai, L., Li, H., Ouyang, W., Zhang, Y.: Hvdistill: Transferring knowledge from images to point clouds via unsupervised hybrid-view distillation. International Journal of Computer Vision pp. 1–15 (2024)
27. Zhang, Z., Girdhar, R., Joulin, A., Misra, I.: Self-supervised pretraining of 3d features on any point-cloud. In: IEEE/CVF International Conference on Computer Vision. pp. 10252–10263 (2021)