CleverDistiller: Simple and Spatially Consistent Cross-modal Distillation

Hariprasath Govindarajan[†2,4]
hargov@qti.qualcomm.com
Maciej K. Wozniak[†1]
maciejw@kth.se
Marvin Klingner[3]
mklingne@qti.qualcomm.com
Camille Maurice[3]
cmaurice@qti.qualcomm.com
B Ravi Kiran[5]
ravkira@qti.qualcomm.com
Senthil Yogamani[6]
syogaman@qti.qualcomm.com

† equal contribution

[1] KTH Royal Institute of Technology, Sweden
[2] Linköping University, Sweden
[3] Qualcomm Arriver Software GmbH
[4] Qualcomm Auto Ltd Sweden Filial
[5] Qualcomm France, S.A.R.L.
[6] Qualcomm Technologies, Inc.

Abstract

Vision foundation models have revolutionized 2D camera-based perception by extracting generalized features for downstream tasks. Recent work applies self-supervised cross-modal knowledge distillation (KD) to transfer these capabilities to 3D LiDAR models, but often relies on complex losses or pseudo-semantic maps. We introduce CleverDistiller, a self-supervised, cross-modal 2D-to-3D KD framework, introducing simple yet effective design choices. Our method uses a direct feature similarity loss and an MLP projection head to capture complex semantic dependencies without relying on pseudo-semantic maps or explicit semantic supervision. In addition, we improve the 3D spatial reasoning capabilities of learned representations through a self-supervised occupancy prediction task. Experiments on autonomous driving benchmarks show that CleverDistiller achieves state-of-the-art performance in both 3D semantic segmentation and 3D object detection, with up to 10% mIoU improvement, particularly when fine-tuning with limited data, demonstrating the effectiveness of our approach.

1 Introduction

For many vision applications such as autonomous driving, improving the generalized understanding of the environment is important to increase safety and autonomy. However, constant updates of sensor types and setups from different manufacturers makes it difficult to train one perception system that excels on all use cases. Most perception models for

object detection or semantic segmentation are known for their lack of robustness towards weather conditions, lighting, sensor setup, or driving environments that usually deviate between training and inference [60]. Especially, LiDAR-based 3D models are known for their adaptation challenges [57, 61, 72].

Due to these inherent challenges, recent focus has shifted towards using Vision Foundation Models (VFMs) such as DINO [11, 33, 46]. These models provide features that are robust and generalize well under domain shift, while also supporting multiple downstream tasks. Although many works address and extend VFMs in camera-based setups [40, 53, 62], our work aims to transfer VFM knowledge to more challenging LiDAR-based 3D networks. Cross-modal knowledge distillation has been used to transfer learned feature representations between the backbones of different sensor modalities. Initial methods focused mainly on distilling LiDAR-based features into camera-based models [17, 34]. However, with the emergence of VFMs the reverse direction, to distill generalized semantic 2D features into 3D LiDAR-based networks, has become more relevant [42, 48, 70, 71].

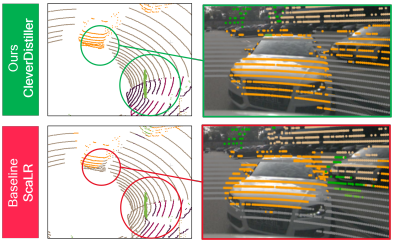

Figure 1: We observe how **CleverDistiller** (top) improves over the baseline ScaLR (bottom) by producing spatially consistent semantic outputs.

Many previous approaches for cross-modal KD have tried to explicitly model the semantic dependencies by using semantic priors, that depend on extra models or preprocessing steps to extract semantically coherent image regions (SAM or OpenSeeD [53, 70, 71, 75, 77]). They show improvements in performance and robustness, but tend to complicate the cross-modal KD process. On the other hand, ScaLR [48] shows the effectiveness of a simple and scalable distillation solution, but does not achieve SOTA performance. As demonstrated in Fig. 1 ScaLR yields semantic outputs that are sometimes inconsistent with the scene's spatial structure. Overall, previous methods are either complicated and involve semantic priors to beat SOTA or are simple and promising but failing to improve over SOTA performance.

To address these shortcomings, we propose **CleverDistiller**, a fully self-supervised KD framework that transfers features from a camera-based VFM to a LiDAR-based network. Our approach focuses on obtaining multi-task features by integrating cross-modal KD from a VFM and an auxiliary spatial task, i.e., occupancy prediction [7], to capture both semantic and spatial information in a unified framework, thereby improving spatial and semantic feature consistency. Our contributions can be summarized as follows:

- We identify linear projection of 3D LiDAR features into 2D space of camera-based VFM features to be a key limitation of cross-modal KD frameworks (overlooked by *all* previous methods). We show that an MLP projection head on top of the 3D LiDAR backbone enables the backbone to learn more informative features that excel across multiple tasks, while also retaining the simplicity of the distillation (**Secs. 3.2 and 3.3**).

- VFM features are semantically rich but lack spatial and geometric information. We incorporate a self-supervised auxiliary task of occupancy prediction to encourage the LiDAR model to additionally encode spatial information. Thus, the general semantic features distilled from the VFM are further enriched to be spatially consistent across the 3D scene (**Sec. 3.4**).

- We demonstrate that CleverDistiller outperforms previous approaches in standard 2D-to-3D KD benchmarks for both 3D semantic segmentation and 3DOD. Our model is particularly successful when fine-tuned on limited data, highlighting the generalization capabilities of our learned 3D features (**Sec. 4.3**).

2 Related work

Pre-training Vision Models: Supervised pre-training on large labeled datasets [20, 29, 39, 55, 56] has being gradually replaced by self-supervised pre-training [24]. It involves training a model using a pretext task on large unlabeled datasets. Currently, self-supervised learning (SSL) methods predominantly use a contrastive [13] or reconstruction formulation [46, 80]. Initially, a contrastive objective was considered over data instances in a batch [13, 66] or a memory bank [14, 15, 30]. Recent works such as DINO [11] contrast over prototypical clusters [2, 9, 10, 27, 41] and learn representations that are highly data efficient [8, 49]. Reconstruction based methods [5, 31] typically lack data efficiency [9]. Then, DINOv2 [46] consolidated a range of strategies such as masked image modeling [80] and produced a model that excels on a wide range of downstream tasks, adopted in many recent works [23, 28, 58]. While pre-training models on the target domain is a promising direction [40, 62], there is a lack of off-the-shelf models for autonomous driving, thus in this work we chose DINOv2 as a teacher model to train a 3D student backbone.

Pre-training 3D Backbones Self-supervised pre-training of 3D backbones started with single object scans (e.g. ShapeNet [12, 65]) similar to ImageNet in the case of images [16, 47, 50, 74, 79]. Several methods have been developed to pre-train on complex 3D indoor and outdoor scenes [45, 63, 69, 73]. Inspired by their vision counterpart, these methods also mainly use contrastive and reconstruction objectives. Some works leverage LiDAR specific augmentations [38] or use specialized tasks such as occupancy prediction consistent under sparse point sampling [1, 7] and different mixing strategies for pointclouds [37, 44, 67]. However, these methods are limited by the volume of diverse 3D pre-training data, owing to the difficulty in collecting such a large-scale dataset.

Cross-modal 2D-to-3D Knowledge Distillation: Recently, distilling knowledge from a pre-trained 2D image backbone into a 3D LiDAR backbone has been shown to be an effective pre-training strategy. Initial works focused on single object pointclouds [21, 76] but recent works have demonstrated that this strategy is also suitable for complex scenes [42, 43]. SLidR [61] groups points and pixels into semantically coherent superpixels/superpoints based on the teacher vision model features. Other works use additional vision models such as SAM [33] or OpenSeeD [75] to extract pseudo-semantic regions for the same purpose [42, 70]. Recently, some works have also used spatial and temporal consistency to regularize learned representations [42, 70, 71, 77, 78]. Others demonstrated the potential of scaling up different distillation components such as the dataset, the 3D backbone and the 2D backbone while using a simple distillation loss that matches the corresponding point-pixel features directly [48]. However, direct feature comparison impairs the 3D network's ability to learn geometrically rich features. In this work, we propose an MLP projection head and the auxiliary spatial task to guide the 3D network towards learning more spatially aware features.

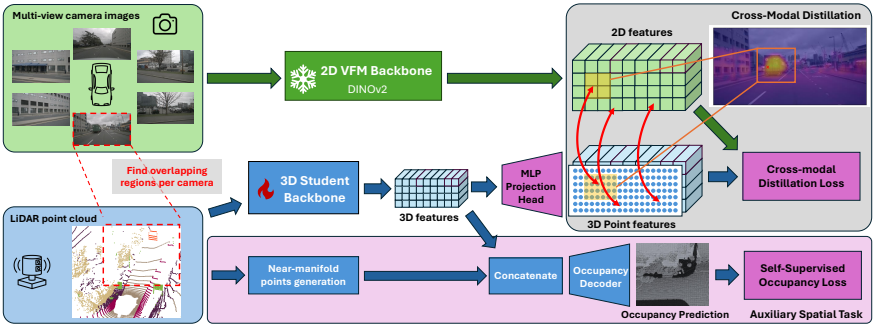

Figure 2: Overview of the **CleverDistiller** framework. Sensor calibration is used to associate 3D points with image regions. Features are extracted using modality-specific backbones. A cross-modal KD loss distills camera features into the 3D backbone via an MLP projection head, while an occupancy loss enforces spatial consistency. The pre-trained 3D backbone is then used for downstream tasks.

3 Method Description

3.1 Problem Definition and Motivation

Given a point cloud $\mathcal{P} = \{\mathbf{p}_i \mid i = 1,\ldots,N\}$ with N points captured by a LiDAR sensor, where $\mathbf{p}_i \in \mathbb{R}^4$ denotes the coordinates of the point and the laser intensity. Let $\mathcal{I}_t = \{\mathbf{I}_k \mid k = 1,\ldots,M\}$ denote M surround camera images where $\mathbf{I}_k \in \mathbb{R}^{H\times W\times 3}$ denotes an image with height H and width W. The objective of cross-modal distillation is to take advantage of a pre-trained vision foundation model $T_\phi(\cdot)$ to learn a LiDAR foundation model, $S_{\theta_p}(\cdot)$ by distilling the features from the image modality to the LiDAR modality.

Many recent works generate a set of class-agnostic or pseudo-semantic superpixels (*e.g.* using text prompts in vision language models) for each image using additional supervised or unsupervised segmentation methods such as SAM [33] or OpenSeeD [75]. These superpixels are not ground truths and require careful hyperparameter tuning to obtain suitable segments for a particular domain. Moreover with larger dataset, generating superpixels requires incrementally more time and computational resources. This complicates the overall distillation process and *we question the need for semantic priors* in the presence of good features from a VFM. Further, most prior works use a contrastive loss [42, 51, 70, 71, 77, 78] which creates an unnecessary *self-conflict* issue: contrasting image segments belonging to similar objects.

In our approach, we find corresponding point and image features, without relying on any additional processing steps or semantic priors. Assuming that the point cloud $\mathcal{P}_t$ and images $\mathcal{I}_t$ are calibrated, we find corresponding points and pixels by projecting the point cloud $\mathbf{p}_i = (x_i, y_i, z_i)$ onto the image plane (u_i, v_i) using the following sensor calibration parameters, $\begin{bmatrix} u_i & v_i & 1 \end{bmatrix}^T = \rho(i) = \frac{1}{z_i} \times \Gamma_K \times \Gamma_{c\leftarrow l} \times \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^T$, where Γ_K denotes the camera-intrinsic matrix and $\Gamma_{c\leftarrow l}$ is the transformation matrix from LiDAR sensors to surround-view cameras. This enables a direct mapping between points and pixels in the image.

3.2 Cross-modal Learning Objectives

Let the student model, $S_{\theta_p} : \mathbb{R}^{N\times 4} \rightarrow \mathbb{R}^{N\times D_p}$ be a 3D backbone with trainable parameters θ_p, which takes LiDAR points as input and outputs D_p-dimensional point features. Let

$T_\phi : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{\frac{H}{S}\times\frac{W}{S}\times D_v}$ be an image backbone with pre-trained parameters ϕ that takes images as input and outputs D_v-dimensional image features.

The primary objective function of self-supervised knowledge distillation is to transfer knowledge from the trained image backbone T_ϕ to the 3D backbone S_{θ_p}. The point features are obtained as $q_p = \mathcal{H}_p(S_{\theta_p}(\mathbf{p}))$ and the image features as $q_v = T_\phi(\mathbf{I})$, where $\mathcal{H}_p : \mathbb{R}^{D_p} \to \mathbb{R}^{D_v}$ is a projection head that transforms the point features to the same dimensions as the VFM teacher. Both features, q_p and q_v are ℓ_2-normalized. Inspired by ScaLR [48], we use a simple pairwise cosine similarity loss given by,

$$\mathcal{L}_{\text{distillation}}(q_p, q_v) = \frac{1}{N}\sum_{i\in 1,\ldots,N} \|q_{p,i} - q_{v,\rho(i)}\|_2. \tag{1}$$

Such a loss is preferable to a contrastive loss as it avoids the complexities of careful negative sample selection, training sensitivity to batch composition and self-conflict issues while maintaining robust feature alignment. In a multi-camera setup, the above loss is applied to each camera view using only those points that are visible in that camera view.

3.3 Improving the Projection Head

We investigate the role of the projection head by evaluating the informativeness of the 3D backbone features using the RankMe metric [25]. RankMe is computed *after* self-supervised distillation and *before* any fine-tuning with labels, enabling us to assess the information content of the learned representations in a label-free setting. Originally proposed for evaluating self-supervised models without labels, RankMe serves as a proxy for representational quality and potential downstream performance. We find that this relationship holds in our context as well: higher RankMe scores consistently correlate with better downstream performance after fine-tuning (see Fig. 3 and Tab. 1). This shows that RankMe can help identify design choices in distillation setting, such as importance of projection head or auxiliary tasks (Tab. 3) that yield more informative features for initialization and task-specific training.

Every other method we compare with [42, 48, 51, 70, 71] uses a linear projection head and we find that this results in less informative 3D backbone features (see Tab. 1). Using an MLP projection head improves the informativeness of these features as evidenced by the improved RankMe metric. Consequently, this results in significant downstream performance improvements (see Tab. 1). A deeper MLP with larger hidden dimensions results in the most informative backbone features and best downstream performance. A complex MLP projection head adds only a minimal increase in computational cost ($\approx 9\%$ more peak GPU memory and compute cost). Based on this analysis, we use a 3-layer MLP projection head with a similar formulation as the MLP head used while pre-training the DINOv2 model, which yields the best results (see Tab. 1 and Fig. 3). While the MLP outputs features that align well with the image featureenables the backbone (before the MLP) to retain additional 3D modality-specific information. This is evident in Tab. 3 comparing results with and without occupancy the MLP: with the MLP, the features retain significantly richer information content, as reflected by a higher RankMe metric.

3.4 Auxiliary Spatial Occupancy Task

Cross-modal distillation alone does not encourage the model to learn spatially rich information, unique to LiDAR that can be useful for downstream tasks. To encourage the learning

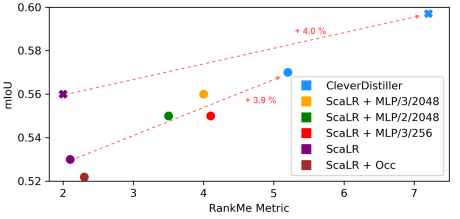

Figure 3: MinkUNet performance (1% finetuning) on nuScenes dataset vs RankMe metric of the 3D backbone features. The colors refer to the distillation method, the image teachers are denoted as • for ViT-S/14 and × for ViT-B/14.

MLP Layers.	Hidden dim.	RankMe	nuScenes		KITTI	Waymo
			LP	1%	1%	1%
1 (Linear)	-	1.89	40.76	53.01	44.42	47.98
2	2048	2.94	45.46	54.38	50.09	48.19
3	256	4.18	44.79	54.80	50.10	48.93
3	2048	**4.24**	**46.36**	**55.01**	**50.15**	**49.08**

Table 1: Comparison of different MLP architectures with the PT-v3 [64] 3D backbone model. The MLP projection improves performance as well as the RankMe metric. All scores in mIoU.

of spatial features, we incorporate an auxiliary spatial task. Combined with an MLP projection head, this controls the spatial information retained in the backbone features. While many works prioritize segmentation tasks where semantic knowledge is sufficient, tasks such as 3D object detection require a deeper spatial reasoning. We experimented with auxiliary tasks like scene flow estimation and occupancy prediction, ultimately selecting occupancy as the most effective (see supplementary Sec. A8).

Occupancy prediction injects spatial and geometric information of the objects into the learned features, which we expect to be useful for tasks such as 3D object detection. The occupancy prediction loss is formulated based on ALSO [7]. We observe that incorporating the occupancy task improves the informativeness of the backbone features by enabling it to learn additional information complementary to those distilled from the VFM features (see improvement on RankMe metric when adding the occupancy task in Tab. 3). This further boosts downstream performance in various tasks. The semantic information from the 2D backbone combined with the spatial knowledge encouraged by occupancy prediction task enable the model to understand and interpret complex 3D environments better. This also makes the model more robust to domain shifts and LiDAR corruption (see Tab. 7).

Following ALSO [7], for each 3D point p sampled on the surface, we create two query points: $q_{\text{front}} = p - \delta p$ and $q_{\text{behind}} = p + \delta p$, where δp is a small distance. We consider q_{front} to be empty and q_{behind} to be filled. Additionally, a third empty query point q_{sight} is randomly picked in the segment between the sensor and p. Importantly, the occupancy prediction is *not* performed on the original surface points themselves but on these generated query points An occupancy decoder head takes a concatenation of the 3D backbone features of the surface point and the 3D coordinates of the query points to predict the occupancy and intensity of the query points. The occupancy loss consists of a binary cross-entropy loss $\mathcal{L}_{\text{O}}$ based on the occupancy classification and a reconstruction MSE loss $\mathcal{L}_{\text{Int}}$ based on the intensity prediction and we use $\lambda = 1$ similar to ALSO as $\mathcal{L}_{\text{occ}} = \mathcal{L}_{\text{O}} + \lambda \mathcal{L}_{\text{Int}}$.

Overall, we combine two simple but significant design choices. First, we use an MLP projection head and cosine similarity loss. Second, we add a complementary auxiliary spatial task. Combined, these improve the informativeness and spatial consistency of the 3D backbone features as well as downstream performance. This combination of design choices leads to a more robust and efficient cross-modal distillation framework.

4 Results

4.1 Experiments setup

Pre-training: *CleverDistiller* is implemented using the Pointcept framework [19]. Consistent with prior works we employ MinkUNet-34 [18] as the 3D backbone and DINOv2 [46] (with ViT backbones [22]) as the 2D backbone, distilling from three variants: small (S), base (B) and large (L). Additionally, we experiment with SOTA 3D backbone, PointTransformer-V3 (PT-v3) [64], to investigate if our method scales to larger and newer architectures, showing that CleverDistiller can work with any point-based 3D backbone. For PT-v3, we use the default configuration and only increase the decoder dimensions to 256. The framework is pre-trained on nuScenes using 4 NVIDIA A100-80GB GPUs for 50 and 100 epochs. We follow the same procedure to pre-train both ScaLR and CleverDistiller, while the results for other methods are based on results reported in their works [42, 43, 51, 70, 71]. Further training setup is provided in Sec. A2. Though most cross-modal image-to-LiDAR distillation works claim to pre-train for only 50 epochs, some recent works such as SuperFlow [70] utilize 3 LiDAR sweeps per sample, increasing the effective length and compute cost of the training by a factor of three (effectively 150 epochs). We characterize this using effective epochs in Tab. 4. For fairer comparison, we show results of our method for both 50 and 100 epochs of pre-training. CleverDistiller^{50} provides a direct comparison to previous methods that used only one LiDAR sweep per sample while CleverDistiller^{100} serves as a more efficient comparison to SuperFlow, ScaLR, and LiMoE+SuperFlow. We experiment on a total of nine datasets in our experiments, namely: nuScenes [8], SemanticKITTI [6] (SKITTI), KITTI [26], Waymo [54], ScribbleKITTI [59], RELLIS-3D [32], SemanticSTF [68], and DAPS-3D [35] and nuScenes-C from Robo3D [36].

Fine tuning: Pre-trained LiDAR backbone is evaluated on a range of downstream tasks on various datasets. Firstly, we conduct segmentation experiments on nuScenes using linear probing (LP) and limited data finetuning scenarios using 4 GPUs for 100 epochs. Further fine-tuning details are provided in Sec. A4. For domain generalization study using Robo3D, we follow the same protocol as previous works. We also test our features on the 3DOD task using the KITTI and Waymo datasets. For 3DOD, we use weights obtained during pre-training on nuScenes. We substitute PointRCNN [52] backbone with MinkUNet (following SLiDR approach). Finally, we train for 80 epochs on partial data for Waymo and KITTI.

Evaluation Protocols: Following standard conventions, we report the Intersection-over-Union (IoU) on each semantic class and mean IoU (mIoU) over all classes for downstream tasks. For 3D robustness evaluations, we follow Robo3D [36] and report the mean Corruption Error (mCE) and mean Resilience Rate (mRR).

4.2 Ablation studies

We conduct ablation studies for the two proposed components, the MLP projection head and the auxiliary occupancy task. We pre-train a MinkUNet-34 and PT-V3 model for 100 epochs using the ViT-S/14 variant of DINOv2. We use the 3-layer MLP head selected in Sec. 3.3 and experimented with different weights for the occupancy loss. From Tab. 2, we see that a weight of 0.05 is sufficient to incorporate additional spatial information. Then, we ablate the best versions of the two components Tab. 3. We observe that the addition of the MLP

Occupancy loss weight	nuScenes		SKITTI	Waymo
	LP	1%	1%	1%
0.0	55.72	61.79	56.30	60.78
0.01	56.49	61.40	57.01	62.14
0.05	**58.49**	**62.54**	**62.19**	61.95
0.2	57.39	61.65	58.96	**62.39**
1.0	53.89	58.99	56.78	62.30

Table 2: Comparison of different occupancy loss weights with PTV3 [64], results in mIoU. MinkUNet-34 [18], results in mIoU.

#	MLP	Occ	Matrix Rank	RankMe	nuScenes		SKITTI	Waymo
					LP	1%	1%	1%
(a)	✗	✗	22	1.89	40.76	53.01	44.42	47.98
(b)	✗	✓	25	2.33	41.67	52.16	48.19	48.11
(c)	✓	✗	33	4.24	46.36	55.01	50.15	49.08
(d)	✓	✓	41	5.27	49.81	56.90	50.59	50.99

Table 3: Ablation study of different components with MinkUNet-34 [18], results in mIoU.

(c) improves the informativeness of the backbone features (RankMe metric) and provides a significant performance boost compared to the baseline (a) in Tab. 3. The occupancy task (d) provides an orthogonal improvement that further improves the features informativeness. Note that the MLP projection head enhances the effectiveness of the occupancy task, as it enables the backbone features to learn additional LiDAR-specific spatial information which may not align with the purely semantic image features.

4.3 Performance Comparison

As shown in Tab. 4 and Tab. 5, CleverDistiller achieves state-of-the-art performance across all fine-tuning tasks, as well as in linear probing, on both MinkUnet [18] and PTV3 [64]. In particular, when fine-tuning with limited data on nuScenes, we observe substantial improvements of up to 10% mIoU in comparison to baselines. CleverDistiller also consistently outperforms competing methods in domain transfer scenarios, such as SKITTI and Waymo.

Method	Venue	Distill	S.P.	Effective Epochs	nuScenes						SKITTI	Waymo
					LP	1%	5%	10%	25%	Full	1%	1%
Random	-	-	✗	-	8.10	30.30	47.84	56.15	65.48	74.66	39.50	39.41
PPKT [43]	arXiv'21	ViT-S	✗	50	38.60	40.60	52.06	59.99	65.76	73.97	43.25	47.44
SLidR [61]	CVPR'22	ViT-S	✓	50	44.70	41.16	53.65	61.47	66.71	74.20	44.67	47.57
Seal [45]	NIPS'23	ViT-S	✓	50	45.16	44.27	55.13	62.46	67.64	75.58	46.51	48.67
CleverDistiller^{50}	-	ViT-S	✗	50	**46.02**	**53.67**	**62.34**	**64.04**	**68.06**	**77.44**	**49.58**	**50.45**
SuperFlow [44]	ECCV'24	ViT-S	✓	150	46.44	47.81	59.44	64.47	69.20	76.54	47.97	49.94
ScaLR [18]	CVPR'24	ViT-S	✗	100	40.76	53.01	61.58	63.47	67.56	74.51	44.42	47.98
LiMoE [71]	CVPR'25	ViT-S	✓	150	48.20	49.60	60.54	65.65	**71.39**	77.27	49.53	**51.42**
CleverDistiller^{100}	-	ViT-S	✗	100	**49.81**	**56.90**	**64.55**	**65.92**	70.11	**77.61**	**50.59**	50.99
PPKT [43]	arXiv'21	ViT-B	✗	50	39.95	40.91	53.21	60.87	66.22	74.07	44.09	47.57
SLidR [61]	CVPR'22	ViT-B	✓	50	45.35	41.64	55.83	62.68	67.61	74.98	45.50	48.32
Seal [45]	NIPS'23	ViT-B	✓	50	46.59	45.98	57.15	62.79	68.18	75.41	47.24	48.91
CleverDistiller^{50}	-	ViT-B	✗	50	**48.40**	**55.80**	**63.49**	**65.56**	**69.06**	**78.17**	**49.83**	**52.80**
SuperFlow [44]	ECCV'24	ViT-B	✓	150	47.66	48.09	59.66	64.52	69.79	76.57	48.40	50.20
ScaLR [18]	CVPR'24	ViT-B	✗	100	41.80	55.83	63.46	65.24	68.70	74.76	45.59	49.60
LiMoE [71]	CVPR'25	ViT-B	✓	150	49.07	50.23	61.51	66.17	**71.56**	77.81	50.30	51.77
CleverDistiller^{100}	-	ViT-B	✗	100	**51.89**	**59.80**	**66.44**	**67.65**	69.53	**78.49**	**51.48**	**53.56**
PPKT [43]	arXiv'21	ViT-L	✗	50	41.57	42.05	55.75	61.26	66.88	74.33	45.87	47.82
SLidR [61]	CVPR'22	ViT-L	✓	50	45.70	42.77	57.45	63.20	68.13	75.51	47.01	48.60
Seal [45]	NIPS'23	ViT-L	✓	50	46.81	46.27	58.14	63.27	68.67	75.66	47.55	50.02
CleverDistiller^{50}	-	ViT-L	✗	50	**48.32**	**56.70**	**64.51**	**66.25**	**69.33**	**78.68**	**49.86**	**53.32**
SuperFlow [44]	ECCV'24	ViT-L	✓	150	48.01	49.95	60.72	65.09	70.01	77.19	49.07	50.67
ScaLR [18]	CVPR'24	ViT-L	✗	100	40.12	55.78	63.28	64.76	68.19	75.09	44.85	50.34
LiMoE [71]	CVPR'25	ViT-L	✓	150	49.35	51.41	62.07	66.64	**71.59**	77.85	50.69	51.93
CleverDistiller^{100}	-	ViT-L	✗	100	**52.45**	**60.64**	**67.03**	**67.29**	70.45	**78.29**	**52.28**	**54.83**

Table 4: Comparison of SOTA pretraining methods pretrained in self-supervised manner on nuScenes and fine-tuned on nuScenes, SemanticKITTI, and Waymo. All methods use MinkUNet-34 [18] as the backbone for the 3D semantic segmentation task. We also show which 2D VFM backbone was used as a teacher and if the approach uses pseudo semantic priors from off-the shelf segmentation models. LP denotes linear probing with a frozen backbone. All scores in mIoU. For LiMoE [71], we use LiMoE+SuperFlow. "S.P." refers to whether the method uses Semantic Priors.

To the best of our knowledge, we include all recently published and accepted methods (including [71]), and demonstrate that high performance can be achieved through simple design choices, without relying on semantic priors or contrastive losses.

The performance of CleverDistiller highlights strong generalization of the learned representations, making it well-suited for settings with limited annotated data. The results also confirm that models pretrained via representation learning significantly outperform those initialized randomly, underscoring the importance of data-driven pretraining. Furthermore, distillation from larger 2D networks consistently improves downstream performance, aligning with findings from prior work. We further provide additional qualitative results, feature visualizations, and class-specific analyses in the supplementary (Secs. A6 and A9).

4.4 Domain generalization

In domain generalization experiments, we evaluate all methods pre-trained on nuScenes across four diverse datasets as shown in Tab. 6. We fine-tune the models on different subsets of data from the new dataset using the fine-tuning protocol described in Sec. 4.1. Each dataset poses distinct challenges, with some originating outside the autonomous driving domain. Notably, ScriKITTI features extremely sparse scribble-level annotations, yet our method still achieves the best performance among all approaches, highlighting its ability to learn effectively from limited supervision. Rellis-3D, in contrast, is an off-road mobile robotics dataset with a significant domain shift from nuScenes in terms of terrain, sensor configuration, and scene layout. Despite this, our method adapts well and continues to outperform others, indicating strong generalization across both structured and unstructured environments. The results on SemSTF and DAPS-3D further support the robustness of our approach in diverse settings.

4.5 Robustness

We compare our method against SOTA approaches on the Robo3D benchmark [36] presented in Tab. 7 based on the linear probing models. CleverDistiller achieves superior performance, particularly in corrupted settings where other methods degrade significantly for example, when LiDAR data is sparse. It consistently shows improved robustness over random initialization and competing methods, obtaining the lowest mean Corruption Error (mCE) in linear probing (robustness results based on full data finetuned models are provided in the

Method	nuScenes					SKITTI	Waymo
	LP	1%	5%	10%	25%	1%	1%
ViT-S							
ScaLR	44.99	62.07	69.10	72.91	75.69	56.92	60.09
Ours	**58.49**	**62.54**	**71.09**	**73.89**	**76.64**	**61.95**	**62.19**
ViT-B							
ScaLR	41.62	63.54	71.74	73.84	76.56	60.74	57.17
Ours	**60.54**	**63.93**	**72.27**	**74.11**	**76.93**	**63.31**	**62.20**
ViT-L							
ScaLR	45.54	66.17	72.34	74.96	77.56	58.48	61.47
Ours	**61.92**	**66.18**	**73.13**	**75.18**	**77.57**	**62.23**	**65.94**

Table 5: Comparisons of SOTA methods pre-trained on nuScenes and fine-tuned on datasets subset, using PTV3 [54] for 3D semantic segmentation. All scores in mIoU.

Method	ScriKITTI		Rellis-3D		SemSTF		DAPS-3D	
	1%	10%	1%	10%	50%	100%	50%	100%
Random	23.81	47.60	38.46	53.60	48.03	48.15	74.32	79.38
PPKT^{50}	36.50	51.67	49.71	54.33	50.92	54.69	78.90	84.00
SLidR^{50}	39.60	50.45	49.75	54.57	52.01	54.35	81.00	85.40
Seal^{50}	40.64	52.77	51.09	55.03	53.46	**55.36**	81.88	85.90
Ours^{50}	**40.77**	**53.31**	**56.23**	**56.60**	**53.77**	54.72	**82.58**	**86.15**
ScaLR^{100}	36.45	49.16	47.91	48.86	52.10	54.40	-	-
SF^{150}	42.70	54.00	52.83	55.71	54.72	56.57	82.43	86.21
LiMoE^{150}	43.95	55.96	53.74	56.67	**55.60**	**57.31**	83.24	86.68
Ours^{100}	**44.03**	**56.70**	**58.35**	**60.92**	53.99	55.66	**83.06**	**87.95**

Table 6: Domain generalization benchmark. Models are pre-trained on nuScenes and finetuned on different datasets. (all scores in mIoU)

Method	Backbone	mCE ↓	mRR ↑	mIoU ↑								
				Fog	Rain	Snow	Blur	Beam	Cross	Echo	Sensor	Avg Full
PPKT[50] [43]	MinkU-34	183.44	78.15	30.65	35.42	28.12	29.21	32.82	19.52	28.01	20.71	28.06
SLidR[50] [51]	MinkU-34	179.38	77.18	34.88	38.09	32.64	26.44	33.73	20.81	31.54	21.44	29.95
Seal[50] [42]	MinkU-34	166.18	75.38	37.33	42.77	29.93	37.73	40.32	20.31	37.73	24.94	33.88
CleverDistiller[50]	MinkU-34	**157.50**	**84.84**	**42.45**	**43.45**	**38.77**	**37.85**	**41.01**	**43.11**	**39.17**	**26.55**	**39.05**
SuperFlow[150] [40]	MinkU-34	161.78	75.52	37.59	43.42	37.60	39.57	41.40	23.64	38.03	26.69	35.99
Scalr[100] [38]	MinkU-34	173.18	78.91	37.55	37.96	36.29	33.64	33.06	23.01	33.62	23.70	33.59
LiMoE[150]	MinkU-34	155.77	78.23	40.35	45.28	39.14	**42.10**	**44.21**	27.33	39.20	**29.49**	38.39
CleverDistiller[100]	MinkU-34	**151.21**	**89.99**	**43.96**	**46.91**	**41.20**	41.05	42.15	**45.67**	**41.30**	28.85	**41.39**

Table 7: 3D robustness study under corruption and sensor failure scenarios in the nuScenes-C dataset from the Robo3D benchmark [36]. This evaluation is based on the linear probing model trained on frozen MinkUNet-34 backbones obtained using different knowledge distillation methods. All mCE (↓), mRR (↑), and mIoU (↑) scores are given in percentage (%).

supplementary, Sec. A7). Our method also achieves the highest mean Relative Robustness (mRR) across all trials, indicating strong resilience to various corruptions. In the Fog scenario, for instance, CleverDistiller reaches an mIoU of 43.96, outperforming the previous best of 40.35. The most pronounced improvement is observed in the cross-sensor corruption scenario, characterized by extremely sparse LiDAR input. Here, CleverDistiller performs up to 20% points better in linear probing accuracy compared to other methods.

Although LiMoE [71] (very recent unpublished work) slightly outperforms our method in some cases, CleverDistiller remains more robust on average, achieving lower mCE and higher mRR. Notably, our method is simpler and requires significantly lesser training compute than the mixture-of-experts model, LiMoE, highlighting its efficiency and effectiveness.

5 Conclusions

We introduce CleverDistiller, a simple yet effective framework for cross-modal distillation from 2D vision foundation models (VFMs) to 3D LiDAR models. Unlike prior work relying on complex losses or pseudo-labels, we use a direct feature similarity loss and an MLP projection head to improve semantic information transfer. We further enhance spatial reasoning by incorporating an auxiliary task of occupancy prediction. CleverDistiller achieves SOTA performance with up to 10% mIoU gains in low-data regimes.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, and Raquel Urtasun. UnO: Unsupervised occupancy fields for perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[2] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Proceedings of the International Conference on Learning Representations*, 2020.

[3] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Proceedings of the European Conference on Computer Vision*, 2022.

[4] Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *International Conference on Machine Learning*, 2024.

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *Proceedings of the International Conference on Learning Representations*, 2022.

[6] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *Proceedings of the International Conference on Computer Vision*, 2019.

[7] Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, and Renaud Marlet. ALSO: Automotive LiDAR self-supervision by occupancy estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision*, 2018.

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision*, 2021.

[12] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

[14] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision*, 2021.

[16] Ye Chen, Jinxian Liu, Bingbing Ni, Hang Wang, Jiancheng Yang, Ning Liu, Teng Li, and Qi Tian. Shape self-correction for unsupervised point cloud understanding. In *Proceedings of the International Conference on Computer Vision*, 2021.

[17] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. BEVDistill: Cross-modal BEV distillation for multi-view 3D object detection. In *Proceedings of the International Conference on Learning Representations*, 2023.

[18] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[19] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. https://github.com/Pointcept/Pointcept, 2023.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[21] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2D image transformers help 3D representation learning? In *Proceedings of the International Conference on Learning Representations*, 2023.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

[23] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[24] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[25] Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, 2023.

[26] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 2013.

[27] Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. DINO as a von Mises-Fisher mixture model. In *Proceedings of the International Conference on Learning Representations*, 2023.

[28] Guangxing Han and Ser-Nam Lim. Few-shot object detection with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[32] Peng Jiang, Philip Osteen, Maggie Wigness, and Srikanth Saripalli. RELLIS-3D dataset: Data, benchmarks and analysis. In *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021.

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the International Conference on Computer Vision*, 2023.

[34] Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X3KD: Knowledge distillation across modalities, tasks and stages for multi-camera 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[35] Alexey A Klokov, Di Un Pak, Aleksandr Khorin, Dmitry A Yudin, Leon Kochiev, Vladimir D Luchinskiy, and Vitaly D Bezuglyj. DAPS3D: Domain adaptive projective segmentation of 3D LiDAR point clouds. *IEEE Access*, 11:79341–79356, 2023.

[36] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3D: Towards robust and reliable 3D perception against corruptions. In *Proceedings of the International Conference on Computer Vision*, 2023.

[37] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. LaserMix for semi-supervised LiDAR semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[38] Lingdong Kong, Xiang Xu, Jiawei Ren, Wenwei Zhang, Liang Pan, Kai Chen, Wei Tsang Ooi, and Ziwei Liu. Multi-modal data-efficient 3D scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2012.

[40] Tim Lebailly, Thomas Stegmüller, Behzad Bozorgtabar, Jean-Philippe Thiran, and Tinne Tuytelaars. CrIBo: Self-supervised learning via cross-image object-level bootstrapping. In *Proceedings of the International Conference on Learning Representations*, 2024.

[41] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *Proceedings of the International Conference on Learning Representations*, 2020.

[42] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36, 2024.

[43] Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H Hsu. Learning from 2D: Contrastive pixel-to-point knowledge transfer for 3D pretraining. *arXiv preprint arXiv:2104.04687*, 2021.

[44] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3D: Out-of-context data augmentation for 3D scenes. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2021.

[45] Lucas Nunes, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. SegContrast: 3D point cloud feature representation learning through self-supervised segment discrimination. *IEEE Robotics and Automation Letters*, 2022.

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[47] Omid Poursaeed, Tianxing Jiang, Han Qiao, Nayun Xu, and Vladimir G Kim. Self-supervised learning of point clouds via orientation estimation. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2020.

[48] Gilles Puy, Spyros Gidaris, Alexandre Boulch, Oriane Siméoni, Corentin Sautier, Patrick Pérez, Andrei Bursuc, and Renaud Marlet. Three pillars improving vision foundation model distillation for LiDAR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[49] Yangjun Ruan, Saurabh Singh, Warren Richard Morningstar, Alexander A Alemi, Sergey Ioffe, Ian Fischer, and Joshua V Dillon. Weighted ensemble self-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2023.

[50] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 2019.

[51] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-LiDAR self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[52] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[53] Thomas Stegmüller, Tim Lebailly, Behzad Bozorgtabar, Tinne Tuytelaars, and Jean-Philippe Thiran. CrOC: Cross-view online clustering for dense visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[54] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[55] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021.

[56] Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *Proceedings of the European Conference on Computer Vision*, 2022.

[57] Darren Tsai, Julie Stephany Berrio, Mao Shan, Eduardo Nebot, and Stewart Worrall. MS3D: Leveraging multiple detectors for unsupervised domain adaptation in 3D object detection. In *International Conference on Intelligent Transportation Systems*. IEEE, 2023.

[58] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. DINO-tracker: Taming DINO for self-supervised point tracking in a single video. In *Proceedings of the European Conference on Computer Vision*, 2024.

[59] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised LiDAR semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[60] Maciej K Wozniak, Viktor Kårefjärd, Marko Thiel, and Patric Jensfelt. Toward a robust sensor fusion step for 3D object detection on corrupted data. *IEEE Robotics and automation letters*, 8 (11):7018–7025, 2023.

[61] Maciej K Wozniak, Mattias Hansson, Marko Thiel, and Patric Jensfelt. UADA3D: Unsupervised adversarial domain adaptation for 3d object detection with sparse lidar and large domain gaps. *IEEE Robotics and Automation Letters*, 9(12):11210–11217, 2024.

[62] Maciej K Wozniak, Hariprasath Govindarajan, Marvin Klingner, Camille Maurice, B Ravi Kiran, and Senthil Yogamani. S3PT: Scene semantics and structure guided clustering to boost self-supervised pre-training for autonomous driving. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2025.

[63] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3D representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[64] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point Transformer V3: Simpler, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[65] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[66] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[67] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. PolarMix: A general data augmentation technique for LiDAR point clouds. *Advances in Neural Information Processing Systems*, 2022.

[68] Aoran Xiao, Jiaxing Huang, Weihao Xuan, Ruijie Ren, Kangcheng Liu, Dayan Guan, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. 3D semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[69] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. PointContrast: Unsupervised pre-training for 3D point cloud understanding. In *Proceedings of the European Conference on Computer Vision*, 2020.

[70] Xiang Xu, Lingdong Kong, Hui Shuai, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, and Qingshan Liu. 4D contrastive superflows are dense 3D representation learners. In *Proceedings of the European Conference on Computer Vision*, 2024.

[71] Xiang Xu, Lingdong Kong, Hui Shuai, Liang Pan, Ziwei Liu, and Qingshan Liu. LiMoE: Mixture of LiDAR representation learners from automotive scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[72] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. ST3D++: Denoised self-training for unsupervised domain adaptation on 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[73] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. ProposalContrast: Unsupervised pre-training for LiDAR-based 3D object detection. In *Proceedings of the European Conference on Computer Vision*, 2022.

[74] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[75] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the International Conference on Computer Vision*, 2023.

[76] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3D representations from 2D pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[77] Yifan Zhang and Junhui Hou. Fine-grained image-to-LiDAR contrastive distillation with visual foundation models. *Advances in Neural Information Processing Systems*, 2024.

[78] Yifan Zhang and Junhui Hou. Is contrastive distillation enough for learning comprehensive 3D representations? *arXiv preprint arXiv:2412.08973*, 2024.

[79] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3D features on any point-cloud. In *Proceedings of the International Conference on Computer Vision*, 2021.

[80] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *Proceedings of the International Conference on Learning Representations*, 2022.