4D Contrastive Superflows are Dense 3D Representation Learners

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

Abstract. In the realm of autonomous driving, accurate 3D perception is the foundation. However, developing such models relies on extensive human annotations – a process that is both costly and labor-intensive. To address this challenge from a data representation learning perspective, we introduce **SuperFlow**, a novel framework designed to harness consecutive LiDAR-camera pairs for establishing spatiotemporal pretraining objectives. SuperFlow stands out by integrating two key designs: **1)** a dense-to-sparse consistency regularization, which promotes insensitivity to point cloud density variations during feature learning, and **2)** a flow-based contrastive learning module, carefully crafted to extract meaningful temporal cues from readily available sensor calibrations. To further boost learning efficiency, we incorporate a plug-and-play view consistency module that enhances the alignment of the knowledge distilled from camera views. Extensive comparative and ablation studies across 11 heterogeneous LiDAR datasets validate our effectiveness and superiority. Additionally, we observe several interesting emerging properties by scaling up the 2D and 3D backbones during pretraining, shedding light on the future research of 3D foundation models for LiDAR-based perception. Code is publicly available at `https://github.com/Xiangxu-0103/SuperFlow`.

Keywords: LiDAR Segmentation · 3D Data Pretraining · Autonomous Driving · Image-to-LiDAR Contrastive Learning · Semantic Superpixels

1 Introduction

Driving perception is one of the most crucial components of an autonomous vehicle system. Recent advancements in sensing technologies, such as light detection and ranging (LiDAR) sensors and surrounding-view cameras, open up new possibilities for a holistic, accurate, and 3D-aware scene perception [3,9,79].

Training a 3D perception model that can perform well in real-world scenarios often requires large-scale datasets and sufficient computing power [27,58]. Different from 2D, annotating 3D data is notably more expensive and labor-intensive,

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

which hinders the scalability of existing 3D perception models [28, 69, 98, 112]. Data representation learning serves as a potential solution to mitigate such a problem [6, 76]. By designing suitable pretraining objectives, the models are anticipated to extract useful concepts from raw data, where such concepts can help improve models' performance on downstream tasks with fewer annotations [51].

Recently, Sautier *et al*. [82] proposed SLidR to distill knowledge from surrounding camera views – using a pretrained 2D backbone such as MoCo [14] and DINO [72] – to LiDAR point clouds, exhibiting promising 3D representation learning properties. The key to its success is the superpixel-driven contrastive objectives between cameras and LiDAR sensors. Subsequent works further extended this framework from various aspects, such as class balancing [66], hybrid-view distillation [110], semantic superpixels [11, 12, 61], and so on. While these methods showed improved performance over their baselines, there exist several issues that could undermine the data representation learning.

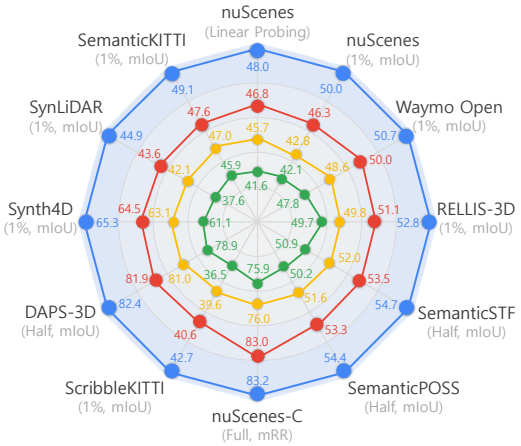


Fig. 1: Performance overview of SuperFlow compared to state-of-the-art image-to-LiDAR pretraining methods, *i.e*., Seal [61], SLidR [82], and PPKT [63], on eleven LiDAR datasets. The scores of prior methods are normalized based on SuperFlow's scores. The larger the area coverage, the better the overall segmentation performance.

The first concern revolves around the inherent temporal dynamics of LiDAR data [4,8]. LiDAR point clouds are acquired sequentially, capturing the essence of motion within the scene. Traditional approaches [61,63,66,82,110] often overlook this temporal aspect, treating each snapshot as an isolated scan. However, this sequential nature holds a wealth of information that can significantly enrich the model's understanding of the 3D environment [71, 96]. Utilizing these temporal cues can lead to more robust and context-aware 3D perception models, which is crucial for dynamic environments encountered in autonomous driving.

Moreover, the varying density of LiDAR point clouds presents a unique challenge [45, 47, 94]. Due to the nature of LiDAR scanning and data acquisition, different areas within the same scene can have significantly different point densities, which can in turn affect the consistency of feature representation across the scene [2,47,108,111]. Therefore, a model that can learn invariant features regardless of point cloud density tends to be effective for recognizing the structural and semantic information in the 3D space.

In lieu of existing challenges, we propose a novel spatiotemporal contrastive learning dubbed **SuperFlow** to encourage effective cross-sensor knowledge dis-

tillation. Our approach features three key components, all centered around the use of the off-the-shelf temporal cues inherent in the LiDAR acquisition process:

- We first introduce a straightforward yet effective view consistency alignment that seamlessly generates semantic superpixels with language guidance, alleviating the "self-conflict" issues in existing works [61, 66, 82]. As opposed to the previous pipeline, our method also aligns the semantics across camera views in consecutive scenes, paving the way for more sophisticated designs.
- To address the varying density of LiDAR point clouds, we present a dense-to-sparse regularization module that encourages consistency between features of dense and sparse point clouds. Dense points are obtained by concatenating multi-sweep LiDAR scans within a suitable time window and propagating the semantic superpixels from sparse to dense points. By leveraging dense point features to regularize sparse point features, the model promotes insensitivity to point cloud density variations.
- To capture useful temporal cues from consecutive scans across different timestamps, we design a flow-based contrastive learning module. This module takes multiple LiDAR-camera pairs as input and excites strong consistency between temporally shifted representations. Analogous to existing image-to-LiDAR representation learning methods [61, 66, 82], we also incorporate useful spatial contrastive objectives into our framework, setting a unified pipeline that emphasizes holistic representation learning from both the structural 3D layouts and the temporal 4D information.

The strong spatiotemporal consistency regularization in SuperFlow effectively forms a semantically rich landscape that enhances data representations. As illustrated in Fig. 1, our approach achieves appealing performance gains over state-of-the-art 3D pretraining methods across a diverse spectrum of downstream tasks. Meanwhile, we also target at scaling the capacity of both 2D and 3D backbones during pretraining, shedding light on the future development of more robust, unified, and ubiquitous 3D perception models.

To summarize, this work incorporates key contributions listed as follows:

- We present **SuperFlow**, a novel framework aimed to harness consecutive LiDAR-camera pairs for establishing spatiotemporal pretraining objectives.
- Our framework incorporates novel designs including view consistency alignment, dense-to-sparse regularization, and flow-based contrastive learning, which better encourages data representation learning effects between camera and LiDAR sensors across consecutive scans.
- Our approach sets a new state-of-the-art performance across 11 LiDAR datasets, exhibiting strong robustness and generalizability. We also reveal intriguing emergent properties as we scale up the 2D and 3D backbones, which could lay the foundation for scalable 3D perception.

2 Related Work

LiDAR-based 3D Perception. The LiDAR sensor has been widely used in today's 3D perception systems, credited to its robust and structural sensing abil-

ities [4, 88, 92]. Due to the sparse and unordered nature of LiDAR point clouds, suitable rasterization strategies are needed to convert them into structural inputs [37, 93]. Popular choices include sparse voxels [18, 19, 33, 34, 90, 118], bird's eye view maps [10,56,111,117], range view images [17,21,44,68,104,107,116], and multi-view fusion [18, 40, 60, 62, 77, 105, 106]. While witnessing record-breaking performances on standard benchmarks, existing approaches rely heavily on human annotations, which hinders scalability [27]. In response to this challenge, we resort to newly appeared 3D representation learning, hoping to leverage the rich collections of unlabeled LiDAR point clouds for more effective learning from LiDAR data. This could further enrich the efficacy of LiDAR-based perception.

Data-Efficient 3D Perception. To better save annotation budgets, previous efforts seek 3D perception in a data-efficient manner [11, 12, 27, 40, 46, 49]. One line of research resorts to weak supervision, *e.g.*, seeding points [36, 53, 86, 115], active prompts [38, 57, 100], and scribbles [94], for weakly-supervised LiDAR semantic segmentation. Another line of research seeks semi-supervised learning approaches [47, 52, 91] to better tackle efficient 3D scene perception and achieve promising results. In this work, different from the prior pursuits, we tackle efficient 3D perception from the data representation learning perspective. We establish several LiDAR-based data representation learning settings that seamlessly combine pretraining with weakly- and semi-supervised learning, further enhancing the scalability of 3D perception systems.

3D Representation Learning. Analog to 2D representation learning strategies [13, 15, 30, 31, 103], prior works designed contrastive [35, 70, 81, 101, 108, 113], masked modeling [32, 50, 95], and reconstruction [7, 67] objectives for 3D pretraining. Most early 3D representation learning approaches use a single modality for pretraining, leaving room for further development. The off-the-shelf calibrations among different types of sensors provide a promising solution for building pretraining objectives [63]. Recently, SLidR [82] has made the first contribution toward multi-modal 3D representation learning between camera and LiDAR sensors. Subsequent works [66,74,110] extended this framework with more advanced designs. Seal [61] leverages powerful vision foundation models [42,109,119,120] to better assist the contrastive learning across sensors. Puy *et al.* [75,76] conducted a comprehensive study on the distillation recipe for better pretraining effects. While these approaches have exhibited better performance than their baselines, they overlooked the rich temporal cues across consecutive scans, which might lead to sub-opt pretraining performance. In this work, we construct dense 3D representation learning objectives using calibrated LiDAR sequences. Our approach encourages the consistency between features from sparse to dense inputs and features across timestamps, yielding superiority over existing endeavors.

4D Representation Learning. Leveraging consecutive scans is promising in extracting temporal relations [2, 23, 33, 85]. For point cloud data pretraining, prior works [16, 64, 83, 84, 114] mainly focused on applying 4D cues on object- and human-centric point clouds, which are often small in scale. For large-scale automotive point clouds, STRL [39] learns spatiotemporal data invariance with different spatial augmentations in the point cloud sequence. TARL [71] and

STSSL [96] encourage similarities of point clusters in two consecutive frames, where such clusters are obtained by ground removal and clustering algorithms, *i.e.*, RANSAC [25], Patchwork [55], and HDBSCAN [24]. BEVContrast [81] shares a similar motivation but utilizes BEV maps for contrastive learning, which yields a more effective implementation. The "one-fits-all" clustering parameters, however, are often difficult to obtain, hindering existing works. Different from existing methods that use a single modality for 4D representation learning, we propose to leverage LiDAR-camera correspondences and semantic-rich superpixels to establish meaningful multi-modality 4D pretraining objectives.

3 SuperFlow

In this section, we first revisit the common setups of the camera-to-LiDAR distillation baseline (*cf*. Sec. 3.1). We then elaborate on the technical details of SuperFlow, encompassing a straightforward yet effective view consistency alignment (*cf*. Sec. 3.2), a dense-to-sparse consistency regularization (*cf*. Sec. 3.3), and a flow-based spatiotemporal contrastive learning (*cf*. Sec. 3.4). The overall pipeline of the proposed SuperFlow framework is depicted in Fig. 4.

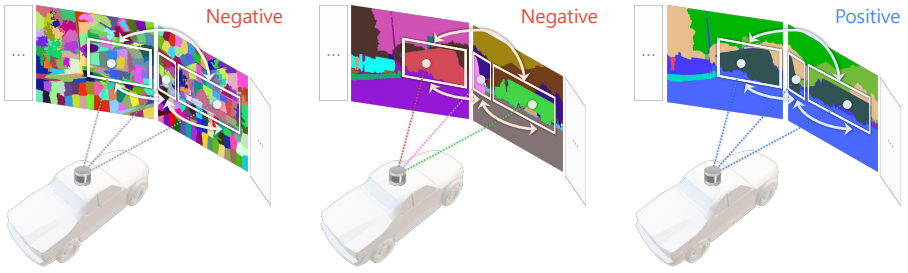
3.1 Preliminaries

Problem Definition. Given a point cloud $\mathcal{P}^t = \{\mathbf{p}_i^t, \mathbf{f}_i^t | i = 1, ..., N\}$ with N points captured by a LiDAR sensor at time t, where $\mathbf{p}_i \in \mathbb{R}^3$ denotes the coordinate of the point and $\mathbf{f}_i \in \mathbb{R}^C$ is the corresponding feature, we aim to transfer knowledge from M surrounding camera images $\mathcal{I}^t = \{\mathbf{I}_i^t | i = 1, ..., M\}$ into the point cloud. Here, $\mathbf{I}_i \in \mathbb{R}^{H \times W \times 3}$ represents an image with height H and width W. Prior works [61, 82] generate a set of class-agnostic superpixels $\mathcal{X}_i = \{\mathbf{X}_i^j | j = 1, ..., V\}$ for each image via the unsupervised SLIC algorithm [1] or the more recent vision foundation models (VFMs) [42, 119, 120], where V denotes the total number of superpixels. Assuming that the point cloud $\mathcal{P}^t$ and images $\mathcal{I}^t$ are calibrated, the point cloud $\mathbf{p}_i = (x_i, y_i, z_i)$ can be then projected to the image plane (u_i, v_i) using the following sensor calibration parameters:

$$[u_i, v_i, 1]^{\mathrm{T}} = \frac{1}{z_i} \times \mathit{\Gamma}_K \times \mathit{\Gamma}_{c \leftarrow l} \times [x_i, y_i, z_i]^{\mathrm{T}} , \tag{1}$$

where $\mathit{\Gamma}_K$ denotes the camera intrinsic matrix and $\mathit{\Gamma}_{c \leftarrow l}$ is the transformation matrix from LiDAR sensors to surrounding-view cameras. We also obtain a set of superpoints $\mathcal{Y} = \{\mathbf{Y}^j | j = 1, ..., V\}$ through this projection.

Network Representations. Let $\mathcal{F}_{\theta_p} : \mathbb{R}^{N \times (3+C)} \rightarrow \mathbb{R}^{N \times D}$ be a 3D backbone with trainable parameters θ_p, which takes LiDAR points as input and outputs D-dimensional point features. Let $\mathcal{G}_{\theta_i} : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{\frac{H}{S} \times \frac{W}{S} \times E}$ be an image backbone with pretrained parameters θ_i that takes images as input and outputs E-dimensional image features with stride S. Let $\mathcal{H}_{\omega_p} : \mathbb{R}^{N \times D} \rightarrow \mathbb{R}^{N \times L}$ and $\mathcal{H}_{\omega_i} : \mathbb{R}^{\frac{H}{S} \times \frac{W}{S} \times E} \rightarrow \mathbb{R}^{H \times W \times L}$ be linear heads with trainable parameters ω_p and ω_i,



(a) Heuristic **(b)** Class Agnostic **(c)** View Consistent

Fig. 2: Comparisons of different superpixels. (a) Class-agnostic superpixels generated by the unsupervised SLIC [1] algorithm. (b) Class-agnostic semantic superpixels generated by vision foundation models (VFMs) [109, 119, 120]. (c) View-consistent semantic superpixels generated by our view consistency alignment module.

which project backbone features to L-dimensional features with ℓ_2-normalization and upsample image features to $H \times W$ with bilinear interpolation.

Pretraining Objective. The overall objective of image-to-LiDAR representation learning [82] is to transfer knowledge from the trained image backbone $\mathcal{G}_{\theta_i}$ to the 3D backbone $\mathcal{F}_{\theta_p}$. The superpixels $\mathcal{X}_i$ generated offline, serve as an intermediate to effectively guide the knowledge transfer process.

3.2 View Consistency Alignment

Motivation. The class-agnostic superpixels $\mathcal{X}_i$ used in prior works [61,66,82] are typically instance-level and do not consider their actual categories. As discussed in [66], instance-level superpixels can lead to "self-conflict" problems, which undermines the effectiveness of pretraining.

Superpixel Comparisons. Fig. 2 compares superpixels generated via the unsupervised SLIC [1] and VFMs. SLIC [1] tends to over-segment objects, causing semantic conflicts. VFMs generate superpixels through a panoptic segmentation head, which can still lead to "self-conflict" in three conditions (see Fig. 2b): ① when the same object appears in different camera views, leading to different parts of the same object being treated as negative samples; ② when objects of the same category within the same camera view are treated as negative samples; ③ when objects across different camera views are treated as negative samples even if they share the same label.

Semantic-Related Superpixels Generation. To address these issues, we propose generating semantic-related superpixels to ensure consistency across camera views. Contrastive Vision-Language Pre-training (CLIP) [78] has shown great generalization in few-shot learning. Building on existing VFMs [42, 119, 120], we employ CLIP's text encoder and fine-tune the last layer of the segmentation head from VFMs with predefined text prompts. This allows the segmentation head to generate language-guided semantic categories for each pixel, which we leverage as superpixels. As shown in Fig. 2c, we unify superpixels across camera

views based on semantic category, alleviating the "self-conflict" problem in prior image-to-LiDAR contrastive learning pipelines.

3.3 D2S: Dense-to-Sparse Consistency Regularization

Motivation. LiDAR points are sparse and often incomplete, significantly restricting the efficacy of the cross-sensor feature representation learning process. In this work, we propose to tackle this challenge by combining multiple LiDAR scans within a suitable time window to create a dense point cloud, which is then used to encourage consistency with the sparse point cloud.

Point Cloud Concatenation. Specifically, given a keyframe point cloud $\mathcal{P}^t$ captured at time t and a set of sweep point clouds $\{\mathcal{P}^s | s = 1, ..., T\}$ captured at previous times s, we first transform the coordinate (x^s, y^s, z^s) of the sweep point cloud $\mathcal{P}^s$ to the coordinate systems of $\mathcal{P}^t$, as they share different systems due to the vehicle's movement:

$$[\tilde{x}^s, \tilde{y}^s, \tilde{z}^s]^{\mathrm{T}} = \Gamma_{t \leftarrow s} \times [x^s, y^s, z^s]^{\mathrm{T}} , \tag{2}$$

where $\Gamma_{t \leftarrow s}$ denotes the transformation matrix from the sweep point cloud at time s to the keyframe point cloud at time t. We then concatenate the transformed sweep points $\{\tilde{\mathcal{P}}^s | s = 1, ...T\}$ with $\mathcal{P}^t$ to obtain a dense point cloud $\mathcal{P}^d$. As shown in Fig. 3, $\mathcal{P}^d$ fuses temporal information from consecutive point clouds, resulting in a dense and semantically rich representation for feature learning.

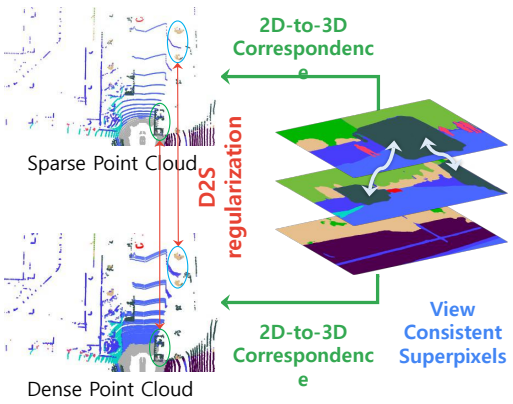


Fig. 3: Dense-to-sparse (D2S) consistency regularization module. Dense point clouds are obtained by combining multiple point clouds captured at different times. A D2S regularization is formulated by encouraging the consistency between dense features and sparse features.

Dense Superpoints. Meanwhile, we generate sets of superpoints $\mathcal{Y}^d$ and $\mathcal{Y}^t$ for $\mathcal{P}^d$ and $\mathcal{P}^t$, respectively, using superpixels $\mathcal{X}^t$. Both $\mathcal{P}^t$ and $\mathcal{P}^d$ are fed into the weight-shared 3D network $\mathcal{F}_{\theta_p}$ and $\mathcal{H}_{\omega_p}$ for feature extraction. The output features are grouped via average pooling based on the superpoint indices to obtain superpoint features $\mathbf{Q}^d$ and $\mathbf{Q}^t$, where $\mathbf{Q}^d \in \mathbb{R}^{V \times L}$ and $\mathbf{Q}^d \in \mathbb{R}^{V \times L}$. We expect $\mathbf{Q}^d$ and $\mathbf{Q}^t$ to share similar features, leading to the following D2S loss:

$$\mathcal{L}_{\mathrm{d2s}} = \frac{1}{V} \sum_{i=1}^{V} (1 - < \mathbf{q}_i^t, \mathbf{q}_i^d >) , \tag{3}$$

where $< \cdot, \cdot >$ denotes the scalar product to measure the similarity of features.

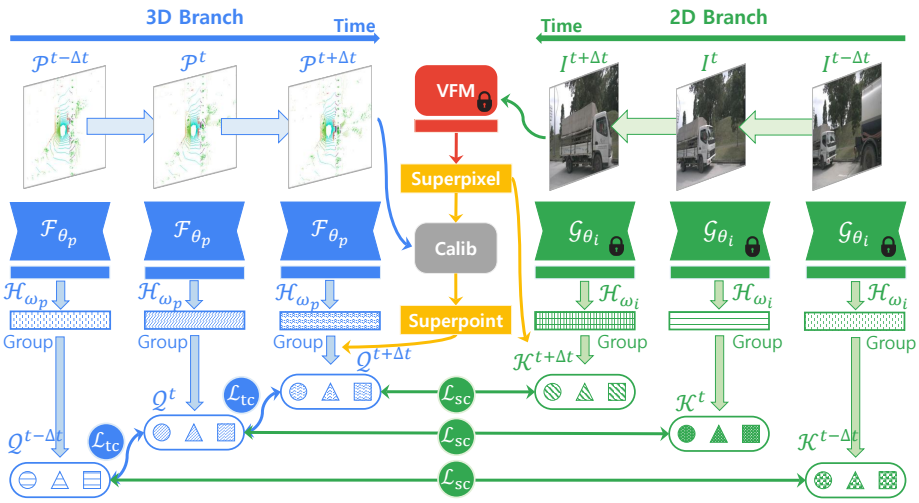


Fig. 4: Flow-based contrastive learning (FCL) pipeline. FCL takes multiple LiDAR-camera pairs from consecutive scans as input. Based on temporally aligned semantic superpixel and superpoints, two contrastive learning objectives are formulated: 1) spatial contrastive learning between each LiDAR-camera pair ($\mathcal{L}_{\text{sc}}$), and 2) temporal contrastive learning among consecutive LiDAR point clouds across scenes ($\mathcal{L}_{\text{tc}}$).

3.4 FCL: Flow-Based Contrastive Learning

Motivation. LiDAR point clouds are acquired sequentially, embedding rich dynamic scene information across consecutive timestamps. Prior works [61, 66, 82] primarily focused on single LiDAR scans, overlooking the consistency of moving objects across scenes. To address these limitations, we propose flow-based contrastive learning (FCL) across sequential LiDAR scenes to encourage spatiotemporal consistency.

Spatial Contrastive Learning. Our framework, depicted in Fig. 4, takes three LiDAR-camera pairs from different timestamps within a suitable time window as input, *i.e.*, $\{(\mathcal{P}^t, \mathcal{I}^t), (\mathcal{P}^{t+\Delta t}, \mathcal{I}^{t+\Delta t}), (\mathcal{P}^{t-\Delta t}, \mathcal{I}^{t-\Delta t})\}$, where timestamp t denotes the current scene and Δt is the timespan. Following previous works [61,82], we first distill knowledge from the 2D network into the 3D network for each scene separately. Taking $(\mathcal{P}^t, \mathcal{I}^t)$ as an example, $\mathcal{P}^t$ and $\mathcal{I}^t$ are fed into the 3D and 2D networks to extract per-point and image features. The output features are then grouped via average pooling based on superpoints $\mathcal{Y}^t$ and superpixels $\mathcal{X}^t$ to obtain superpoint features $\mathbf{Q}^t$ and superpixel features $\mathbf{K}^t$. A spatial contrastive loss is formulated to constrain 3D representation via pretrained 2D prior knowledge. This process is formulated as follows:

$$\mathcal{L}_{\text{sc}} = -\frac{1}{V}\sum_{i=1}^{V}\log\left[\frac{e^{(<\mathbf{q}_i,\mathbf{k}_i>/\tau)}}{\sum_{j\neq i} e^{(<\mathbf{q}_i,\mathbf{k}_j>/\tau)} + e^{(<\mathbf{q}_i,\mathbf{k}_i>/\tau)}}\right], \tag{4}$$

where $\tau > 0$ is a temperature that controls the smoothness of distillation.

Flow-Based Contrastive Learning. The spatial contrastive learning objective between images and point clouds, as depicted in Eq. (4), fails to ensure that moving objects share similar attributes across different scenes. To maintain consistency across scenes, a temporal consistency loss is introduced among superpoint features across different scenes. For the point clouds $\mathcal{P}^t$ and $\mathcal{P}^{t+\Delta t}$, the corresponding superpoint features $\mathbf{Q}^t$ and $\mathbf{Q}^{t+\Delta t}$ are obtained via their superpoints. The temporal contrastive loss operates on $\mathbf{Q}^t$ and $\mathbf{Q}^{t+\Delta t}$:

$$\mathcal{L}_{\text{tc}}^{t \leftarrow t+\Delta t} = -\frac{1}{V} \sum_{i=1}^{V} \log \left[\frac{e^{(<\mathbf{q}_i^t, \mathbf{q}_i^{t+\Delta t}>/\tau)}}{\sum_{j \neq i} e^{(<\mathbf{q}_i^t, \mathbf{q}_j^{t+\Delta t}>/\tau)} + e^{(<\mathbf{q}_i^t, \mathbf{q}_i^{t+\Delta t}>/\tau)}} \right] . \quad (5)$$

The same function is also applied between $\mathbf{Q}^t$ and $\mathbf{Q}^{t-\Delta t}$. This approach enables point features at time t to extract more context-aware information across scenes.

4 Experiments

4.1 Settings

Data. We follow the seminar works SLidR [82] and Seal [61] when preparing the datasets. A total of eleven datasets are used in our experiments, including [1]*nuScenes* [26], [2]*SemanticKITTI* [5], [3]*Waymo Open* [89], [4]*ScribbleKITTI* [94], [5]*RELLIS-3D* [41], [6]*SemanticPOSS* [73], [7]*SemanticSTF* [99], [8]*SynLiDAR* [97], [9]*DAPS-3D* [43], [10]*Synth4D* [80], and [11]*Robo3D* [45]. Due to space limits, kindly refer to the Appendix and [61, 82] for additional details about these datasets.

Implementation Details. SuperFlow is implemented using the MMDetection3D [20] and OpenPCSeg [59] codebases. Consistent with prior works [61, 82], we employ MinkUNet [19] as the 3D backbone and DINOv2 [72] (with ViT backbones [22]) as the 2D backbone, distilling from three variants: small (S), base (B), and large (L). Following Seal [61], OpenSeeD [109] is used to generate semantic superpixels. The framework is pretrained end-to-end on 600 scenes from *nuScenes* [26], then linear probed and fine-tuned on *nuScenes* [26] according to the data splits in SLidR [82]. The domain generalization study adheres to the same configurations as Seal [61] for the other ten datasets. Both the baselines and SuperFlow are pretrained using eight GPUs for 50 epochs, while linear probing and downstream fine-tuning experiments use four GPUs for 100 epochs, all utilizing the AdamW optimizer [65] and OneCycle scheduler [87]. Due to space limits, kindly refer to the Appendix for additional implementation details.

Evaluation Protocols. Following 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 [45] and report the mean Corruption Error (mCE) and mean Resilience Rate (mRR).

4.2 Comparative Study

Linear Probing. We start by investigating the pretraining quality via linear probing. For this setup, we initialize the 3D backbone $\mathcal{F}_{\theta_p}$ with pretrained parameters and fine-tune only the added-on segmentation head. As shown in Tab. 1,

Table 1: Comparisons of state-of-the-art pretraining methods pretrained on *nuScenes* [26] and fine-tuned on *SemanticKITTI* [5] and *Waymo Open* [89] with specified data portions, respectively. All methods use MinkUNet [19] as the 3D semantic segmentation backbone. **LP** denotes linear probing with a frozen backbone. All scores are given in percentage (%). Best scores in each configuration are shaded with colors.

Method	Venue	Distill	nuScenes						KITTI	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
PointContrast [101]	ECCV'20	None ○	21.90	32.50	-	-	-	-	41.10	-
DepthContrast [113]	ICCV'21	None ○	22.10	31.70	-	-	-	-	41.50	-
ALSO [7]	CVPR'23	None ○	-	37.70	-	59.40	-	72.00	-	-
BEVContrast [81]	3DV'24	None ○	-	38.30	-	59.60	-	72.30	-	-
PPKT [63]	arXiv'21	ResNet ○	35.90	37.80	53.74	60.25	67.14	74.52	44.00	47.60
SLidR [82]	CVPR'22	ResNet ○	38.80	38.30	52.49	59.84	66.91	74.79	44.60	47.12
ST-SLidR [66]	CVPR'23	ResNet ○	40.48	40.75	54.69	60.75	67.70	75.14	44.72	44.93
TriCC [74]	CVPR'23	ResNet ○	38.00	41.20	54.10	60.40	67.60	75.60	45.90	-
Seal [61]	NeurIPS'23	ResNet ○	44.95	45.84	55.64	62.97	68.41	75.60	46.63	49.34
HVDistill [110]	IJCV'24	ResNet ○	39.50	42.70	56.60	62.90	69.30	76.60	49.70	-
PPKT [63]	arXiv'21	ViT-S ○	38.60	40.60	52.06	59.99	65.76	73.97	43.25	47.44
SLidR [82]	CVPR'22	ViT-S ○	44.70	41.16	53.65	61.47	66.71	74.20	44.67	47.57
Seal [61]	NeurIPS'23	ViT-S ○	45.16	44.27	55.13	62.46	67.64	75.58	46.51	48.67
SuperFlow	**Ours**	ViT-S ●	46.44	47.81	59.44	64.47	69.20	76.54	47.97	49.94
PPKT [63]	arXiv'21	ViT-B ○	39.95	40.91	53.21	60.87	66.22	74.07	44.09	47.57
SLidR [82]	CVPR'22	ViT-B ○	45.35	41.64	55.83	62.68	67.61	74.98	45.50	48.32
Seal [61]	NeurIPS'23	ViT-B ○	46.59	45.98	57.15	62.79	68.18	75.41	47.24	48.91
SuperFlow	**Ours**	ViT-B ●	47.66	48.09	59.66	64.52	69.79	76.57	48.40	50.20
PPKT [63]	arXiv'21	ViT-L ○	41.57	42.05	55.75	61.26	66.88	74.33	45.87	47.82
SLidR [82]	CVPR'22	ViT-L ○	45.70	42.77	57.45	63.20	68.13	75.51	47.01	48.60
Seal [61]	NeurIPS'23	ViT-L ○	46.81	46.27	58.14	63.27	68.67	75.66	47.55	50.02
SuperFlow	**Ours**	ViT-L ●	48.01	49.95	60.72	65.09	70.01	77.19	49.07	50.67

SuperFlow consistently outperforms state-of-the-art methods under diverse configurations. We attribute this to the use of temporal consistency learning, which captures the structurally rich temporal cues across consecutive scenes and enhances the semantic representation learning of the 3D backbone. We also observe improved performance with larger 2D networks (*i.e.*, from ViT-S to ViT-L), revealing a promising direction of achieving higher quality 3D pretraining.

Downstream Fine-Tuning. It is known that data representation learning can mitigate the need for large-scale human annotations. Our study systematically compares SuperFlow with prior works on three popular datasets, including *nuScenes* [26], *SemanticKITTI* [5], and *Waymo Open* [89], under limited annotations for few-shot fine-tuning. From Tab. 1, we observe that SuperFlow achieves promising performance gains among three datasets across all fine-tuning tasks. We also use the pretrained 3D backbone as initialization for the fully-supervised learning study on *nuScenes* [26]. As can be seen from Tab. 1, models pretrained via representation learning consistently outperform the random initialization counterparts, highlighting the efficacy of conducting data pretraining. We also find that distillations from larger 2D networks show consistent improvements.

Cross-Domain Generalization. To verify the strong generalizability of SuperFlow, we conduct a comprehensive study using seven diverse LiDAR datasets and

Table 2: Domain generalization study of different pretraining methods pretrained on the *nuScenes* [26] dataset and fine-tuned on other *seven* heterogeneous 3D semantic segmentation datasets with specified data portions, respectively. All scores are given in percentage (%). Best scores in each configuration are shaded with colors.

Method	ScriKITTI		Rellis-3D		SemPOSS		SemSTF		SynLiDAR		DAPS-3D		Synth4D	
	1%	10%	1%	10%	Half	Full	Half	Full	1%	10%	Half	Full	1%	10%
Random	23.81	47.60	38.46	53.60	46.26	54.12	48.03	48.15	19.89	44.74	74.32	79.38	20.22	66.87
PPKT [63]	36.50	51.67	49.71	54.33	50.18	56.00	50.92	54.69	37.57	46.48	78.90	84.00	61.10	62.41
SLidR [82]	39.60	50.45	49.75	54.57	51.56	55.36	52.01	54.35	42.05	47.84	81.00	85.40	63.10	62.67
Seal [61]	40.64	52.77	51.09	55.03	53.26	56.89	53.46	55.36	43.58	49.26	81.88	85.90	64.50	66.96
SuperFlow	42.70	54.00	52.83	55.71	54.41	57.33	54.72	56.57	44.85	51.38	82.43	86.21	65.31	69.43

Table 3: Out-of-distribution 3D robustness study of state-of-the-art pretraining methods under corruption and sensor failure scenarios in the *nuScenes-C* dataset from the *Robo3D* benchmark [45]. **Full** denotes fine-tuning with full labels. **LP** denotes linear probing with a frozen backbone. All mCE (↓), mRR (↑), and mIoU (↑) scores are given in percentage (%). Best scores in each configuration are shaded with colors.

#	Initial	Backbone	mCE	mRR	Fog	Rain	Snow	Blur	Beam	Cross	Echo	Sensor	Avg
Full	Random	MinkU-18 ○	115.61	70.85	53.90	71.10	48.22	51.85	62.21	37.73	57.47	38.97	52.68
	SuperFlow	MinkU-18 ●	109.00	75.66	54.95	72.79	49.56	57.68	62.82	42.45	59.61	41.77	55.21
	Random	MinkU-34 ○	112.20	72.57	62.96	70.65	55.48	51.71	62.01	31.56	59.64	39.41	54.18
	PPKT [63]	MinkU-34 ○	105.64	75.87	64.01	72.18	59.08	57.17	63.88	36.34	60.59	39.57	56.60
	SLidR [82]	MinkU-34 ○	106.08	75.99	65.41	72.31	56.01	56.07	62.87	41.94	61.16	38.90	56.83
	Seal [61]	MinkU-34 ○	92.63	83.08	72.66	74.31	66.22	66.14	65.96	57.44	59.87	39.85	62.81
	SuperFlow	MinkU-34 ●	91.67	83.17	70.32	75.77	65.41	61.05	68.09	60.02	58.36	50.41	63.68
	Random	MinkU-50 ○	113.76	72.81	49.95	71.16	45.36	55.55	62.84	36.94	59.12	43.15	53.01
	SuperFlow	MinkU-50 ●	107.35	74.02	54.36	73.08	50.07	56.92	64.05	38.10	62.02	47.02	55.70
	Random	MinkU-101 ○	109.10	74.07	50.45	73.02	48.85	58.48	64.18	43.86	59.82	41.47	55.02
	SuperFlow	MinkU-101 ●	96.44	78.57	56.92	76.29	54.70	59.35	71.89	55.13	60.27	51.60	60.77
LP	PPKT [63]	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 [82]	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 [61]	MinkU-34 ○	166.18	75.38	37.33	42.77	29.93	37.73	40.32	20.31	37.73	24.94	33.88
	SuperFlow	MinkU-34 ●	161.78	75.52	37.59	43.42	37.60	39.57	41.40	23.64	38.03	26.69	35.99

show results in Tab. 2. It is worth noting that these datasets are collected under different acquisition and annotation conditions, including adverse weather, weak annotations, synthetic collection, and dynamic objects. For all fourteen domain generalization fine-tuning tasks, SuperFlow exhibits superior performance over the prior arts [61,63,82]. This study strongly verifies the effectiveness of the proposed flow-based contrastive learning for image-to-LiDAR data representation.

Out-of-Distribution Robustness. The robustness of 3D perception models against unprecedented conditions directly correlates with the model's applicability to real-world applications [29, 48, 54, 102]. We compare our SuperFlow with prior models in the *nuScenes-C* dataset from the *Robo3D* benchmark [45] and show results in Tab. 3. We observe that models pretrained using SuperFlow exhibit improved robustness over the random initialization counterparts. Besides, we find that 3D networks with different capacities often pose diverse robustness.

Quantitative Assessments. We visualize the prediction results fine-tuned on nuScenes [26], SemanticKITTI [5] and Waymo Open [89], compared with random

Table 4: Ablation study of SuperFlow using different # of sweeps. All methods use ViT-B [72] for distillation. All scores are given in percentage (%). Baseline results are shaded with colors.

Backbone	nuScenes LP	nuScenes 1%	KITTI 1%	Waymo 1%
1× Sweeps ○	47.41	47.52	48.14	49.31
2× Sweeps ●	47.66	48.09	48.40	50.20
5× Sweeps ○	47.23	48.00	47.94	49.14
7× Sweeps ○	46.03	47.98	46.83	47.97

Table 5: Ablation study of SuperFlow on network capacity (# params) of 3D backbones. All methods use ViT-B [72] for distillation. All scores are given in percentage (%). Baseline results are shaded with colors.

Backbone	Layer	nuScenes LP	nuScenes 1%	KITTI 1%	Waymo 1%
MinkUNet ○	18	47.20	47.70	48.04	49.24
MinkUNet ●	34	47.66	48.09	48.40	50.20
MinkUNet ○	50	54.11	52.86	49.22	51.20
MinkUNet ○	101	52.56	51.19	48.51	50.01

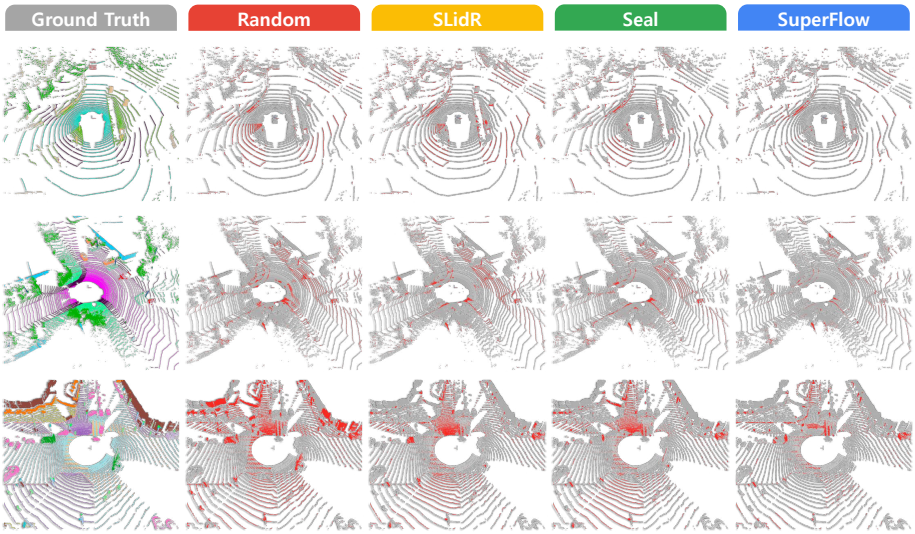


Fig. 5: Qualitative assessments of state-of-the-art pretraining methods pretrained on *nuScenes* [26] and fine-tuned on *nuScenes* [26], *SemanticKITTI* [5], and *Waymo Open* [89], with 1% annotations. The error maps show the correct and incorrect predictions in gray and red, respectively. Best viewed in colors and zoomed-in for details.

initialization, SLiDR [82], and Seal [61]. As shown in Fig. 5, Superflow performs well, especially on backgrounds, *i.e.*, "road" and "sidewalk" in complex scenarios.

4.3 Ablation Study

In this section, we are tailored to understand the efficacy of each design in our SuperFlow framework. Unless otherwise specified, we adopt MinkUNet-34 [19] and ViT-B [72] as the 3D and 2D backbones, respectively, throughout this study.

3D Network Capacity. Existing 3D backbones are relatively small in scale compared to their 2D counterparts. We study the scale of the 3D network and the results are shown in Tab. 5. We observe improved performance as the network capacity scales up, except for MinkUNet-101 [19]. We conjecture that this is due

Table 6: Ablation study of each component in SuperFlow. All variants use a MinkUNet-34 [19] as the 3D backbone and ViT-B [72] for distillation. **VC**: View consistency. **D2S**: Dense-to-sparse regularization. **FCL**: Flow-based contrastive learning. All scores are given in percentage (%).

#	VC	D2S	FCL	nuScenes LP	nuScenes 1%	KITTI 1%	Waymo 1%
-	Random			8.10	30.30	39.50	39.41
(a)	✗	✗	✗	44.65	44.47	46.65	47.77
(b)	✓	✗	✗	45.57	45.21	46.87	48.01
(c)	✓	✓	✗	46.17	46.91	47.26	49.01
(d)	✓	✗	✓	47.24	47.67	48.21	49.80
(e)	✓	✓	✓	47.66	48.09	48.40	50.20

Table 7: Ablation study on spatiotemporal consistency. All variants use a MinkUNet-34 [19] as the 3D backbone and ViT-B [72] for distillation. **0** denotes current timestamp. **0.5s** corresponds to a 20Hz timespan. All scores are given in percentage (%).

Timespan	nuScenes LP	nuScenes 1%	KITTI 1%	Waymo 1%
Single-Frame	46.17	46.91	47.26	49.01
$0, -0.5s$	46.39	47.08	47.99	49.78
$-0.5s, 0, +0.5s$	47.66	48.09	48.40	50.20
$-1.0s, 0, +1.0s$	47.60	47.99	48.43	50.18
$-1.5s, 0, +1.5s$	46.43	48.27	48.34	49.93
$-2.0s, 0, +2.0s$	46.20	48.49	48.18	50.01

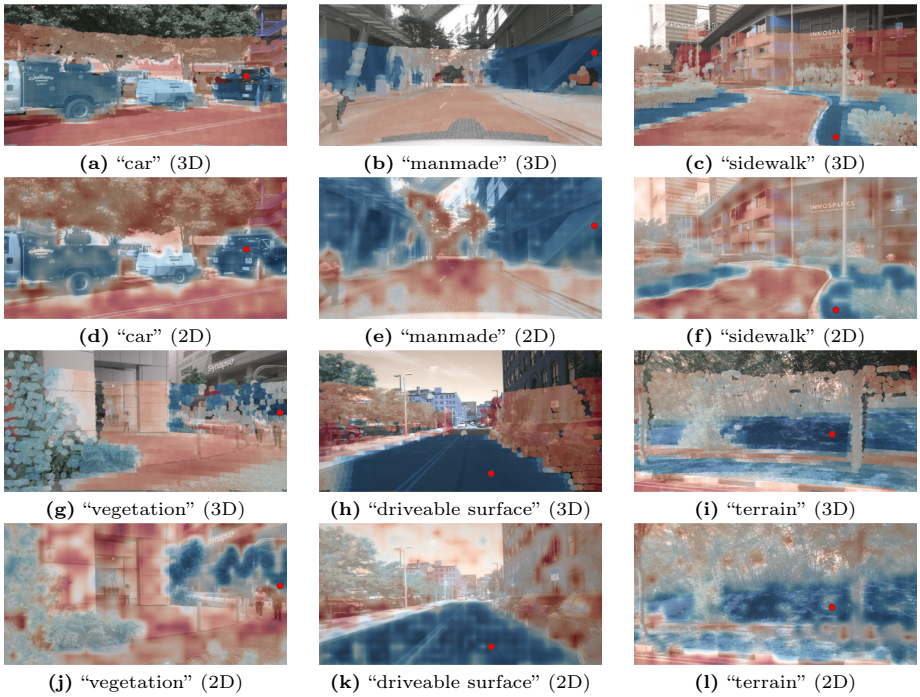
to the fact that models with limited parameters are less effective in capturing patterns during representation learning, and, conversely, models with a large set of trainable parameters tend to be difficult to converge.

Representation Density. The consistency regularization between sparse and dense point clouds encourages useful representation learning. To analyze the degree of regularization, we investigate various point cloud densities and show the results in Tab. 4. We observe that a suitable point cloud density can improve the model's ability to feature representation. When the density of point clouds is too dense, the motion of objects is obvious in the scene. However, we generate superpoints of the dense points based on superpixels captured at the time of sparse points. The displacement difference of dynamic objects makes the projection misalignment. A trade-off selection would be two or three sweeps.

Temporal Consistency. The ability to capture semantically coherent temporal cues is crucial in our SuperFlow framework. In Eq. (5), we operate temporal contrastive learning on superpoints features across scenes. As shown in Tab. 7, we observe that temporal contrastive learning achieves better results compared to single-frame methods. We also compare the impact of frames used to capture temporal cues. When we use 3 frames, it acquires more context-aware information than 2 frames and achieves better results. Finally, we study the impact of the timespan between frames. The performance will drop with a longer timespan. We conjecture that scenes with short timespans have more consistency, while long timespans tend to have more uncertain factors.

Component Analysis. In Tab. 6, we analyze each component in the SuperFlow framework, including view consistency, dense-to-sparse regularization, and flow-based contrastive learning. The baseline is SLiDR [82] with VFMs-based superpixels. View consistency brings slight improvements among the popular datasets with a few annotations. D2S distills dense features into sparse features and it brings about 1% mIoU gains. FCL extracts temporal cues via temporal contrastive learning and it significantly leads to about 2.0% mIoU gains.

Visual Inspections. Similarity maps presented in Fig. 6 denote the segmentation ability of our pretrained model. The query points include "car", "man-



(a) "car" (3D) **(b)** "manmade" (3D) **(c)** "sidewalk" (3D)
(d) "car" (2D) **(e)** "manmade" (2D) **(f)** "sidewalk" (2D)
(g) "vegetation" (3D) **(h)** "driveable surface" (3D) **(i)** "terrain" (3D)
(j) "vegetation" (2D) **(k)** "driveable surface" (2D) **(l)** "terrain" (2D)

Fig. 6: Cosine similarity between features of a query point (red dot) and: 1) features of other points projected in the image (the 1st and 3rd rows); and 2) features of an image with the same scene (the 2nd and 4th rows). The color goes from red to blue denoting low and high similarity scores, respectively. Best viewed in color.

made", "sidewalk", "vegetation", "driveable surface", and "terrain". SuperFlows shows strong semantic discriminative ability without fine-tuning. We conjecture that it comes from three aspects: 1) View consistent superpixels enable the network to learn semantic representation; 2) Dense-to-sparse regularization enhances the network to learn varying density features; 3) Temporal contrastive learning extracts semantic cues across scenes.

5 Conclusion

In this work, we presented **SuperFlow** to tackle the challenging 3D data representation learning. Motivated by the sequential nature of LiDAR acquisitions, we proposed three novel designs to better encourage spatiotemporal consistency, encompassing view consistency alignment, dense-to-sparse regularization, and flow-based contrastive learning. Extensive experiments across 11 diverse LiDAR datasets showed that SuperFlow consistently outperforms prior approaches in linear probing, downstream fine-tuning, and robustness probing. Our study on scaling up 2D and 3D network capacities reveals insightful findings. We hope this work could shed light on future designs of powerful 3D foundation models.

Acknowledgements. This work was supported by the Scientific and Technological Innovation 2030 - "New Generation Artificial Intelligence" Major Project (No. 2021ZD0112200), the Joint Funds of the National Natural Science Foundation of China (No. U21B2044), the Key Research and Development Program of Jiangsu Province (No. BE2023016-3), and the Talent Research Start-up Foundation of Nanjing University of Posts and Telecommunications (No. NY223172). This work was also supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221- 0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012)
2. Aygun, M., Osep, A., Weber, M., Maximov, M., Stachniss, C., Behley, J., Leal-Taixé, L.: 4d panoptic lidar segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5527–5537 (2021)
3. Badue, C., Guidolini, R., Carneiro, R.V., Azevedo, P., Cardoso, V.B., Forechi, A., Jesus, L., Berriel, R., Paixão, T.M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., Souza, A.F.D.: Self-driving cars: A survey. Expert Systems with Applications **165**, 113816 (2021)
4. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Gall, J., Stachniss, C.: Towards 3d lidar-based semantic scene understanding of 3d point cloud sequences: The semantickitti dataset. International Journal of Robotics Research **40**, 959–96 (2021)
5. 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)
6. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013)
7. 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)
8. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
9. Cao, A.Q., Dai, A., de Charette, R.: Pasco: Urban 3d panoptic scene completion with uncertainty awareness. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14554–14564 (2024)
10. Chen, Q., Vora, S., Beijbom, O.: Polarstream: Streaming lidar object detection and segmentation with polar pillars. In: Advances in Neural Information Processing Systems. vol. 34 (2021)

11. Chen, R., Liu, Y., Kong, L., Chen, N., Zhu, X., Ma, Y., Liu, T., Wang, W.: Towards label-free scene understanding by vision foundation models. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
12. Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., Wang, W.: Clip2scene: Towards label-efficient 3d scene understanding by clip. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7020–7030 (2023)
13. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607 (2020)
14. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
15. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
16. Chen, Y., Nießner, M., Dai, A.: 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In: European Conference on Computer Vision. pp. 543–560 (2022)
17. Cheng, H., Han, X., Xiao, G.: Cenet: Toward concise and efficient lidar semantic segmentation for autonomous driving. In: IEEE International Conference on Multimedia and Expo. pp. 1–6 (2022)
18. Cheng, R., Razani, R., Taghavi, E., Li, E., Liu, B.: Af2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12547–12556 (2021)
19. 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)
20. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d` (2020)
21. Cortinhal, T., Tzelepis, G., Aksoy, E.E.: Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In: International Symposium on Visual Computing. pp. 207–222 (2020)
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
23. Duerr, F., Pfaller, M., Weigel, H., Beyerer, J.: Lidar-based recurrent 3d semantic segmentation with temporal memory alignment. In: International Conference on 3D Vision. pp. 781–790 (2020)
24. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 226–231 (1996)
25. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
26. 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)

27. Gao, B., Pan, Y., Li, C., Geng, S., Zhao, H.: Are we hungry for 3d lidar data for semantic segmentation? a survey of datasets and methods. IEEE Transactions on Intelligent Transportation Systems **23**(7), 6063–6081 (2021)
28. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012)
29. Hao, X., Wei, M., Yang, Y., Zhao, H., Zhang, H., Zhou, Y., Wang, Q., Li, W., Kong, L., Zhang, J.: Is your hd map constructor reliable under sensor corruptions? arXiv preprint arXiv:2406.12214 (2024)
30. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
31. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
32. Hess, G., Jaxing, J., Svensson, E., Hagerman, D., Petersson, C., Svensson, L.: Masked autoencoders for self-supervised learning on automotive point clouds. arXiv preprint arXiv:2207.00531 (2022)
33. Hong, F., Kong, L., Zhou, H., Zhu, X., Li, H., Liu, Z.: Unified 3d and 4d panoptic segmentation via dynamic shifting networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(5), 3480–3495 (2024)
34. Hong, F., Zhou, H., Zhu, X., Li, H., Liu, Z.: Lidar-based panoptic segmentation via dynamic shifting network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13090–13099 (2021)
35. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring data-efficient 3d scene understanding with contrastive scene contexts. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15587–15597 (2021)
36. Hu, Q., Yang, B., Fang, G., Guo, Y., Leonardis, A., Trigoni, N., Markham, A.: Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In: European Conference on Computer Vision. pp. 600–619 (2022)
37. Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A.: Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4977–4987 (2021)
38. Hu, Z., Bai, X., Zhang, R., Wang, X., Sun, G., Fu, H., Tai, C.L.: Lidal: Interframe uncertainty based active learning for 3d lidar semantic segmentation. In: European Conference on Computer Vision. pp. 248–265 (2022)
39. Huang, S., Xie, Y., Zhu, S.C., Zhu, Y.: Spatio-temporal self-supervised representation learning for 3d point clouds. In: IEEE/CVF International Conference on Computer Vision. pp. 6535–6545 (2021)
40. Jaritz, M., Vu, T.H., de Charette, R., Wirbel, E., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12605–12614 (2020)
41. 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)
42. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)

43. 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)
44. Kong, L., Liu, Y., Chen, R., Ma, Y., Zhu, X., Li, Y., Hou, Y., Qiao, Y., Liu, Z.: Rethinking range view representation for lidar segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 228–240 (2023)
45. 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)
46. Kong, L., Quader, N., Liong, V.E.: Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In: IEEE International Conference on Robotics and Automation. pp. 9338–9345 (2023)
47. Kong, L., Ren, J., Pan, L., Liu, Z.: Lasermix for semi-supervised lidar semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21705–21715 (2023)
48. Kong, L., Xie, S., Hu, H., Ng, L.X., Cottereau, B.R., Ooi, W.T.: Robodepth: Robust out-of-distribution depth estimation under corruptions. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
49. Kong, L., Xu, X., Ren, J., Zhang, W., Pan, L., Chen, K., Ooi, W.T., Liu, Z.: Multi-modal data-efficient 3d scene understanding for autonomous driving. arXiv preprint arXiv:2405.05258 (2024)
50. Krispel, G., Schinagl, D., Fruhwirth-Reisinger, C., Possegger, H., Bischof, H.: Maeli: Masked autoencoder for large-scale lidar point clouds. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3383–3392 (2024)
51. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. IEEE Transactions on Pattern Analysis and Machine Intelligence **8**, 193907–193934 (2020)
52. Li, L., Shum, H.P., Breckon, T.P.: Less is more: Reducing task and model complexity for 3d point cloud semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9361–9371 (2023)
53. Li, R., de Charette, R., Cao, A.Q.: Coarse3d: Class-prototypes for contrastive learning in weakly-supervised 3d point cloud segmentation. In: British Machine Vision Conference (2022)
54. Li, Y., Kong, L., Hu, H., Xu, X., Huang, X.: Optimizing lidar placements for robust driving perception in adverse conditions. arXiv preprint arXiv:2403.17009 (2024)
55. Lim, H., Oh, M., Myung, H.: Patchwork: Concentric zone-based region-wise ground segmentation with ground likelihood estimation using a 3d lidar sensor. IEEE Robotics and Automation Letters **6**(4), 6458–6465 (2021)
56. Liong, V.E., Nguyen, T.N.T., Widjaja, S., Sharma, D., Chong, Z.J.: Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. arXiv preprint arXiv:2012.04934 (2020)
57. Liu, M., Zhou, Y., Qi, C.R., Gong, B., Su, H., Anguelov, D.: Less: Label-efficient semantic segmentation for lidar point clouds. In: European Conference on Computer Vision. pp. 70–89 (2022)
58. Liu, M., Yurtsever, E., Zhou, X., Fossaert, J., Cui, Y., Zagar, B.L., Knoll., A.C.: A survey on autonomous driving datasets: Data statistic, annotation, and outlook. arXiv preprint arXiv:2401.01454 (2024)

59. Liu, Y., Bai, Y., Kong, L., Chen, R., Hou, Y., Shi, B., Li, Y.: Pcseg: An open source point cloud segmentation codebase. `https://github.com/PJLab-ADG/PCSeg` (2023)
60. Liu, Y., Chen, R., Li, X., Kong, L., Yang, Y., Xia, Z., Bai, Y., Zhu, X., Ma, Y., Li, Y., Qiao, Y., Hou, Y.: Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In: IEEE/CVF International Conference on Computer Vision. pp. 21662–21673 (2023)
61. 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)
62. Liu, Y., Kong, L., Wu, X., Chen, R., Li, X., Pan, L., Liu, Z., Ma, Y.: Multi-space alignments towards universal lidar segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14648–14661 (2024)
63. 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)
64. Liu, Y., Chen, J., Zhang, Z., Huang, J., Yi, L.: Leaf: Learning frames for 4d point cloud sequence understanding. In: IEEE/CVF International Conference on Computer Vision. pp. 604–613 (2023)
65. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
66. 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)
67. Michele, B., Boulch, A., Puy, G., Vu, T.H., Marlet, R., Courty, N.: Saluda: Surface-based automotive lidar unsupervised domain adaptation. arXiv preprint arXiv:2304.03251 (2023)
68. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4213–4220 (2019)
69. Muhammad, K., Ullah, A., Lloret, J., Ser, J.D., de Albuquerque, V.H.C.: Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Transactions on Intelligent Transportation Systems **22**(7), 4316–4336 (2020)
70. Nunes, L., Marcuzzi, R., Chen, X., Behley, J., Stachniss, C.: Segcontrast: 3d point cloud feature representation learning through self-supervised segment discrimination. IEEE Robotics and Automation Letters **7**(2), 2116–2123 (2022)
71. Nunes, L., Wiesmann, L., Marcuzzi, R., Chen, X., Behley, J., Stachniss, C.: Temporal consistent 3d lidar representation learning for semantic perception in autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5217–5228 (2023)
72. 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)
73. 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)

74. 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)
75. Puy, G., Gidaris, S., Boulch, A., Siméoni, O., Sautier, C., Pérez, P., Bursuc, A., Marlet, R.: Revisiting the distillation of image representations into point clouds for autonomous driving. arXiv preprint arXiv:2310.17504 (2023)
76. Puy, G., Gidaris, S., Boulch, A., Siméoni, O., Sautier, C., Pérez, P., Bursuc, A., Marlet, R.: Three pillars improving vision foundation model distillation for lidar. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21519–21529 (2024)
77. Qiu, H., Yu, B., Tao, D.: Gfnet: Geometric flow network for 3d point cloud semantic segmentation. Transactions on Machine Learning Research (2022)
78. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
79. Rizzoli, G., Barbato, F., Zanuttigh, P.: Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives. Technologies **10**(4) (2022)
80. 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)
81. 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)
82. 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)
83. Shen, Z., Sheng, X., Fan, H., Wang, L., Guo, Y., Liu, Q., Wen, H., Zhou, X.: Masked spatio-temporal structure prediction for self-supervised learning on point cloud videos. In: IEEE/CVF International Conference on Computer Vision. pp. 16580–16589 (2023)
84. Sheng, X., Shen, Z., Xiao, G., Wang, L., Guo, Y., Fan, H.: Point contrastive prediction with semantic clustering for self-supervised learning on point cloud videos. In: IEEE/CVF International Conference on Computer Vision. pp. 16515–16524 (2023)
85. Shi, H., Lin, G., Wang, H., Hung, T.Y., Wang, Z.: Spsequencenet: Semantic segmentation network on 4d point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4574–4583 (2020)
86. Shi, H., Wei, J., Li, R., Liu, F., Lin, G.: Weakly supervised segmentation on outdoor 4d point clouds with temporal matching and spatial graph propagation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11840–11849 (2022)
87. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. arXiv preprint arXiv:1708.07120 (2017)
88. Sun, J., Xu, X., Kong, L., Liu, Y., Li, L., Zhu, C., Zhang, J., Xiao, Z., Chen, R., Wang, T., Zhang, W., Chen, K., Qing, C.: An empirical study of training state-of-the-art lidar segmentation models. arXiv preprint arXiv:2405.14870 (2024)

89. 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)
90. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: European Conference on Computer Vision. pp. 685–702 (2020)
91. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
92. Triess, L.T., Dreissig, M., Rist, C.B., Zöllner, J.M.: A survey on deep domain adaptation for lidar perception. In: IEEE Intelligent Vehicles Symposium Workshops. pp. 350–357 (2021)
93. Uecker, M., Fleck, T., Pflugfelder, M., Zöllner, J.M.: Analyzing deep learning representations of point clouds for real-time in-vehicle lidar perception. arXiv preprint arXiv:2210.14612 (2022)
94. 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)
95. Wei, W., Nejadasl, F.K., Gevers, T., Oswald, M.R.: T-mae: Temporal masked autoencoders for point cloud representation learning. arXiv preprint arXiv:2312.10217 (2023)
96. Wu, Y., Zhang, T., Ke, W., Süsstrunk, S., Salzmann, M.: Spatiotemporal self-supervised learning for point clouds in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5251–5260 (2023)
97. 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)
98. Xiao, A., Huang, J., Guan, D., Zhang, X., Lu, S., Shao, L.: Unsupervised point cloud representation learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 11321–11339 (2023)
99. 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)
100. Xie, B., Li, S., Guo, Q., Liu, C.H., Cheng, X.: Annotator: A generic active learning baseline for lidar semantic segmentation. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
101. 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)
102. Xie, S., Kong, L., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Benchmarking and improving bird's eye view perception robustness in autonomous driving. arXiv preprint arXiv:2405.17426 (2024)
103. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)

104. Xu, C., Wu, B., Wang, Z., Zhan, W., Vajda, P., Keutzer, K., Tomizuka, M.: Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In: European Conference on Computer Vision. pp. 1–19 (2020)
105. Xu, J., Zhang, R., Dou, J., Zhu, Y., Sun, J., Pu, S.: Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 16024–16033 (2021)
106. Xu, W., Li, X., Ni, P., Guang, X., Luo, H., Zhao, X.: Multi-view fusion driven 3d point cloud semantic segmentation based on hierarchical transformer. IEEE Sensors Journal **23**(24), 31461–31470 (2023)
107. Xu, X., Kong, L., Shuai, H., Liu, Q.: Frnet: Frustum-range networks for scalable lidar segmentation. arXiv preprint arXiv:2312.04484 (2023)
108. Yin, J., Zhou, D., Zhang, L., Fang, J., Xu, C.Z., Shen, J., Wang, W.: Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In: European Conference on Computer Vision. pp. 17–33 (2022)
109. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Gao, J., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: IEEE/CVF International Conference on Computer Vision. pp. 1020–1031 (2023)
110. 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)
111. Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., Foroosh, H.: Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9601–9610 (2020)
112. Zhang, Y., Hou, J., Yuan, Y.: A comprehensive study of the robustness for lidar-based 3d object detectors against adversarial attacks. International Journal of Computer Vision pp. 1–33 (2023)
113. 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)
114. Zhang, Z., Dong, Y., Liu, Y., Yi, L.: Complete-to-partial 4d distillation for self-supervised point cloud sequence representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17661–17670 (2023)
115. Zhang, Z., Yang, B., Wang, B., Li, B.: Growsp: Unsupervised semantic segmentation of 3d point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17619–17629 (2023)
116. Zhao, Y., Bai, L., Huang, X.: Fidnet: Lidar point cloud semantic segmentation with fully interpolation decoding. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4453–4458 (2021)
117. Zhou, Z., Zhang, Y., Foroosh, H.: Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13194–13203 (2021)
118. Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9939–9948 (2021)
119. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., Peng, N., Wang, L., Lee, Y.J., Gao, J.: Generalized decoding for pixel, image, and language. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)

120. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. In: Advances in Neural Information Processing Systems. vol. 36 (2023)