

Beyond One Shot, Beyond One Perspective: Cross-View and Long-Horizon Distillation for Better LiDAR Representations

Xiang Xu¹ Lingdong Kong^{2,*} Song Wang³ Chuanwei Zhou⁴ Qingshan Liu^{4,5,✉}

¹Nanjing University of Aeronautics and Astronautics ²National University of Singapore ³Zhejiang University

⁴Nanjing University of Posts and Telecommunications ⁵SKL-TI

Code & Project: <https://github.com/Xiangxu-0103/LiMA>

Abstract

LiDAR representation learning aims to extract rich structural and semantic information from large-scale, readily available datasets, reducing reliance on costly human annotations. However, existing LiDAR representation strategies often overlook the inherent spatiotemporal cues in LiDAR sequences, limiting their effectiveness. In this work, we propose **LiMA**, a novel long-term image-to-LiDAR **Memory Aggregation** framework that explicitly captures longer range temporal correlations to enhance LiDAR representation learning. LiMA comprises three key components: 1) a **Cross-View Aggregation** module that aligns and fuses overlapping regions across neighboring camera views, constructing a more unified and redundancy-free memory bank; 2) a **Long-Term Feature Propagation** mechanism that efficiently aligns and integrates multi-frame image features, reinforcing temporal coherence during LiDAR representation learning; and 3) a **Cross-Sequence Memory Alignment** strategy that enforces consistency across driving sequences, improving generalization to unseen environments. LiMA maintains **high pretraining efficiency** and incurs no additional computational overhead during downstream tasks. Extensive experiments on mainstream LiDAR-based perception benchmarks demonstrate that LiMA significantly improves both LiDAR semantic segmentation and 3D object detection. We hope this work inspires more effective pretraining paradigms for autonomous driving. The code has been made publicly accessible for future research.

1. Introduction

LiDAR sensors provide high-resolution spatial information and are essential for precise environmental perception and safe navigation [3, 34, 40, 71]. However, achieving accurate perception relies on large-scale, densely labeled datasets, which are costly and labor-intensive to acquire, thereby lim-

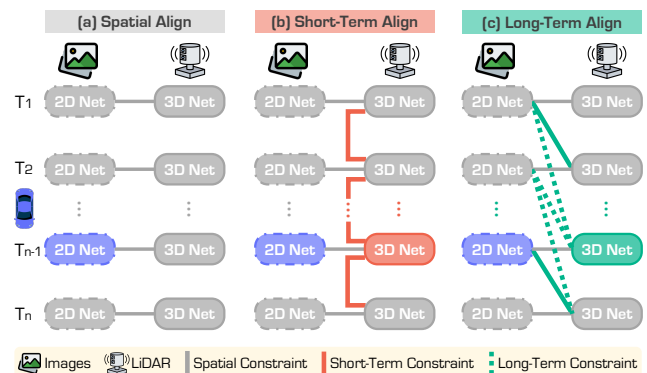


Figure 1. **Illustrative examples of image-to-LiDAR pretraining paradigms.** (a) Spatial Alignment aligns LiDAR features with corresponding image features in the spatial domain without considering temporal consistency. (b) **Short-Term** methods propagate LiDAR features frame by frame, ensuring feature consistency across neighboring frames, but fail to capture long-term dependencies. (c) Our approach leverages **Long-Term** image sequences to enrich LiDAR representations, enabling a more comprehensive understanding of long-range dependencies and motion patterns.

iting scalability in real-world applications [14, 18, 19, 78].

To mitigate this challenge, recent advances in data representation learning leverage large-scale, easily accessible datasets to explore inherent semantic structures [14, 49, 84, 86]. Among these approaches, image-to-LiDAR pretraining methods [52, 61, 66, 102] utilize rich visual priors from RGB images to improve LiDAR feature learning. However, as illustrated in Fig. 1(a), these methods primarily focus on spatial alignment while overlooking the temporal dynamics in LiDAR sequences. As a result, they struggle to capture motion patterns and scene evolution, which are crucial for robust perception in dynamic environments [20, 25, 43, 88].

Recent works [47, 58, 94] incorporate short-term temporal modeling to mitigate this limitation. As shown in Fig. 1(b), these approaches enforce feature consistency across adjacent frames, capturing local motion cues and ensuring smooth feature transitions. However, they rely primarily on frame-to-frame propagation, which maintains local consistency but lacks the capacity to model long-range

(*) Project lead. (✉) Corresponding author.

dependencies and global scene transformations over extended durations. This limitation hinders the ability to construct stable, high-level representations crucial for understanding long-term motion trends and structural variations.

Since LiDAR-based perception tasks inherently involve sequential data, long-term modeling is essential for capturing stable, high-level features that encode extended motion trajectories and scene transitions [42, 48, 77]. Unlike short-term methods that focus on immediate feature alignment, long-term modeling provides a more comprehensive understanding of dynamic environments [89]. This is particularly beneficial for tasks such as motion forecasting, behavior prediction, and autonomous decision-making [15, 64, 74].

To bridge this gap, we propose long-term image-to-LiDAR Memory Aggregation (**LiMA**), a simple yet effective pretraining framework that explicitly models temporal structures in driving sequences, as illustrated in Fig. 1(c). LiMA introduces three key components to enhance LiDAR representation learning from long-term image sequences.

- **Cross-View Aggregation.** To mitigate redundancy in overlapping camera views, this module identifies and unifies shared regions across multiple viewpoints, reinforcing spatial consistency in LiDAR representations. By aligning and fusing features, it constructs a high-quality memory bank for long-term feature propagation.
- **Long-Term Feature Propagation.** To capture long-range dependencies, stored features in the memory bank are transformed into the ego-vehicle coordinate frame, ensuring feature alignment across frames for seamless temporal fusion. This process establishes a temporally coherent image representation, which is distilled into the LiDAR model to encode motion patterns from extended image sequences. The memory bank is dynamically updated in a first-in, first-out (FIFO) manner, enabling continuous temporal propagation across LiDAR frames.
- **Cross-Sequence Memory Alignment.** To improve robustness across diverse driving conditions, this strategy synthesizes mixed scenes from distinct sequences, facilitating adaptation to varying environments. By maintaining structural coherence with sequence-specific memory banks, it ensures robust feature alignment and improves generalization across heterogeneous driving scenarios.

By integrating these three components, LiMA effectively captures both spatial and temporal correlations within and across driving sequences. Extensive experiments on multiple 3D perception benchmarks validate its effectiveness, demonstrating substantial **performance gains** in both LiDAR semantic segmentation and object detection tasks. Notably, our framework ensures **high pretraining efficiency** with an increased number of frames, requiring no more than 20 hours for pretraining. Furthermore, LiMA introduces no additional computational overhead during downstream tasks, ensuring efficient deployment.

To summarize, this work makes contributions as follows:

- We propose **LiMA**, a long-term image-to-LiDAR memory aggregation framework that captures extended temporal dependencies to improve LiDAR representations.
- We introduce three key components: a cross-view aggregation module, a long-term feature propagation module, and a cross-sequence memory alignment strategy, which collectively enhance spatial consistency and temporal robustness across both intra- and inter-sequence domains.
- Extensive experiments across multiple 3D perception benchmarks demonstrate the effectiveness of our method, achieving significant performance improvements in LiDAR semantic segmentation and object detection tasks.

2. Related Work

LiDAR Scene Understanding. LiDAR sensors provide high-precision 3D environmental representations essential for autonomous driving [7, 8, 33, 39]. However, the sparsity and irregularity of LiDAR point clouds pose challenges for accurate perception. To address this, existing methods transform point clouds into various representations, including raw points [24, 62, 63, 68, 69, 75], bird’s-eye view (BEV) [6, 37, 103, 106], range images [1, 30, 54, 90, 95], sparse voxels [11, 22, 23, 73, 107], and multi-view formats [10, 46, 60, 91, 92, 98, 108]. While these methods achieve strong performance, they rely heavily on large-scale labeled datasets, which are expensive and time-consuming to obtain. To mitigate this challenge, recent studies have explored semi-supervised [32, 36, 38, 80] and weakly-supervised [44, 53, 76, 79, 93] learning strategies to reduce annotation costs while maintaining high performance in LiDAR-based perception tasks.

Image-to-LiDAR Data Pretraining. Pretraining facilitates effective feature learning by leveraging tailored objectives such as mask modeling [59, 101], contrastive learning [67, 87], and reconstruction [26, 104]. However, early methods are limited to single-modal point clouds, limiting their scalability and adaptability in large-scale driving environments. To address this constraint, SLiDR [66] introduces a pioneering cross-sensor contrastive learning, aligning pre-trained image and corresponding LiDAR features. Building on this foundation, subsequent works have incorporated more advanced techniques, including class balance strategies [52], semantic-coherent superpixels [35, 47, 94, 97], hybrid representations [96, 102], and knowledge distillation [61]. Despite these advancements, existing approaches largely overlook the temporal dynamics within LiDAR sequences, which are crucial for capturing motion patterns.

Temporal Modeling for LiDAR Representation. LiDAR data inherently capture spatiotemporal dynamics, yet early research predominantly focused on object- or human-centric point clouds [9, 27, 45, 55, 81], limiting scalability in the highly dynamic and unstructured environments

of autonomous driving. To enhance temporal coherence across consecutive scans, recent works have explored various strategies. TriCC [58] enforces triangle consistency to learn temporal relationships, preserving structural integrity over time. Seal [47] utilizes RANSAC-based object segmentation [16] across LiDAR frames to facilitate contrastive learning. SuperFlow [94] enhances temporal modeling by estimating semantic flow across scans, effectively capturing motion cues. However, these methods primarily operate in a frame-by-frame manner, which limits their ability to model long-term dependencies and global scene evolution. In this work, we propose a novel framework that implicitly incorporates long-range temporal information from image sequences into LiDAR representation learning, enabling a more comprehensive understanding of spatiotemporal relationships beyond local pairwise constraints.

3. Revisit Image-to-LiDAR Data Pretraining

In this section, we revisit common strategies for image-to-LiDAR pretraining, including contrastive learning, knowledge distillation, and temporal modeling.

3.1. Preliminaries

Image-to-LiDAR Calibration. Autonomous driving systems integrate LiDAR and multiple cameras, requiring precise spatial and temporal alignment for effective multimodal perception. Temporal alignment ensures both sensors capture the scene simultaneously, mitigating motion-induced artifacts. Spatial alignment involves estimating extrinsic parameters to transform LiDAR point clouds into the camera frame, enabling accurate feature fusion.

Formally, let $\mathcal{P}^t \in \mathbb{R}^{N \times 4}$ denote a point cloud with N points at timestamp t , where each point \mathbf{p}_i^t consists of spatial coordinates (x_i^t, y_i^t, z_i^t) and intensity r_i^t . The corresponding camera image $\mathcal{I}^t \in \mathbb{R}^{H \times W \times 3}$ has a resolution of $H \times W$. Each LiDAR point \mathbf{p}_i is projected onto the camera plane (u_i^t, v_i^t) via the following formulation:

$$\begin{bmatrix} u_i^t & v_i^t \end{bmatrix}^T = \frac{1}{z_i^t} \times \Gamma_K^t \times \Gamma_C^t \times \begin{bmatrix} x_i^t & y_i^t & z_i^t \end{bmatrix}^T, \quad (1)$$

where Γ_K^t is the camera intrinsic matrix and Γ_C^t is the extrinsic transformation from LiDAR to the camera frame. This process establishes a set of point-pixel correspondences $\{\mathbf{p}_i, \mathbf{c}_i\}_{i=1}^M$, where M denotes the number of valid projections within the image bounds.

Pretraining Objective. The image network $\mathcal{G}_{\theta_i} : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H \times W \times C}$, with pretrained parameters θ_i , extracts a C -dimensional feature map \mathcal{F}_i^t from \mathcal{I}^t . Similarly, the LiDAR network $\mathcal{G}_{\theta_p} : \mathbb{R}^{N \times 4} \rightarrow \mathbb{R}^{N \times C}$, parameterized by θ_p , extracts point features \mathcal{F}_p^t from \mathcal{P}^t . The pretraining objective optimizes θ_p such that \mathcal{F}_p^t aligns with \mathcal{F}_i^t , allowing the LiDAR model to inherit semantic priors from the image domain while reducing reliance on labeled datasets.

3.2. Pretraining Methods

Contrastive Learning. Contrastive learning improves cross-modal alignment by constructing positive and negative feature pairs. In image-to-LiDAR learning, contrastive losses can be formulated at the point-pixel [50] or superpoint-superpixel [47, 52, 66, 94] levels:

$$\mathcal{L}_{\text{cont}}(\mathcal{F}_i^t, \mathcal{F}_p^t) = \frac{1}{M} \sum_{j=1}^M \log \frac{e^{\langle \mathbf{f}_{(i,j)}^t, \mathbf{f}_{(p,j)}^t \rangle / \tau}}{\sum_{k=1}^M e^{\langle \mathbf{f}_{(i,k)}^t, \mathbf{f}_{(p,j)}^t \rangle / \tau}}, \quad (2)$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product of feature embeddings, and $\tau > 0$ is a temperature scaling factor. This loss encourages positive feature pairs to be closer while pushing apart negatives, fostering modality-invariant representations.

Knowledge Distillation. Knowledge distillation transfers knowledge from a pretrained image model to a LiDAR model, enforcing feature consistency [61]. A common objective is the ℓ_2 loss, directly operating in the feature space:

$$\mathcal{L}_{\text{dist}}(\mathcal{F}_i^t, \mathcal{F}_p^t) = \frac{1}{M} \sum_{j=1}^M \|\mathbf{f}_{(i,j)}^t - \mathbf{f}_{(p,j)}^t\|_2. \quad (3)$$

By minimizing feature discrepancies, this loss aligns image and LiDAR features, allowing the LiDAR model to leverage image priors while maintaining computational efficiency.

Temporal Modeling. Temporal modeling captures dependencies across consecutive LiDAR scans. To enforce feature consistency across frames, objects are first segmented, and temporal contrastive learning is applied [47, 94]:

$$\mathcal{L}_{\text{cont}}(\mathcal{F}_p^t, \mathcal{F}_p^{t+1}) = \frac{1}{S} \sum_{j=1}^S \log \frac{e^{\langle \mathbf{f}_{(p,j)}^t, \mathbf{f}_{(p,j)}^{t+1} \rangle / \tau}}{\sum_{k=1}^S e^{\langle \mathbf{f}_{(p,k)}^t, \mathbf{f}_{(p,j)}^{t+1} \rangle / \tau}}, \quad (4)$$

where S is the number of segmented objects. By maximizing feature similarity between corresponding objects across frames while contrasting them with non-corresponding ones, this loss enforces temporal coherence, enhancing motion-awareness in LiDAR representations.

4. Methodology

As revisited in Sec. 3, existing approaches [47, 61, 66, 94] focus on spatial alignment or short-term feature propagation, neglecting long-term dependencies, which are crucial for modeling stable representations and extended scene transitions. To address these limitations, we introduce long-term image-to-LiDAR Memory Aggregation (**LiMA**), a simple yet effective framework that distills rich motion patterns from longer sequences into LiDAR representations, enhancing the robustness and generalizability in LiDAR data pretraining. The overall architecture of LiMA is illustrated in Fig. 2. It comprises three key components: **cross-view aggregation**, which unifies multi-view image features to construct a high-quality memory bank (*cf.* Sec. 4.1), a

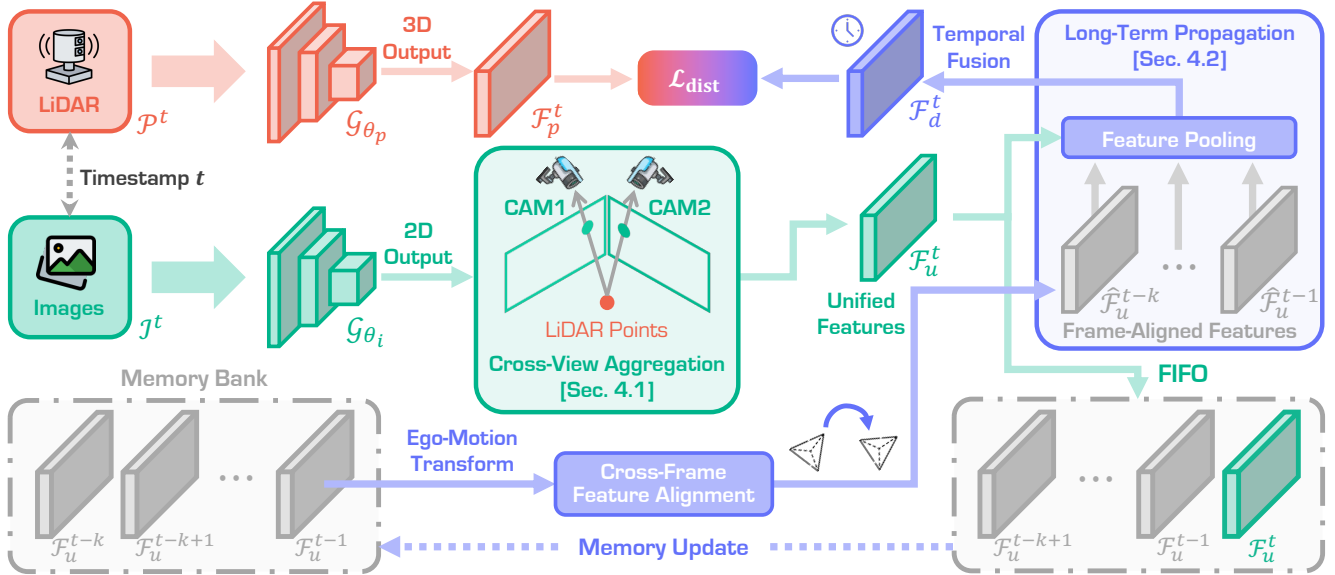


Figure 2. **Overview of the LiMA framework.** At each timestamp t , multi-view image features are first extracted and unified through the **Cross-View Aggregation** module, producing \mathcal{F}_u^t (Sec. 4.1). A **Memory Bank** is introduced to maintain unified image features from the past k frames, enabling temporal feature propagation and fusion with \mathcal{F}_u^t to capture **Long-Term Motion Patterns** \mathcal{F}_d^t (Sec. 4.2). The enriched temporal features are distilled into LiDAR representation \mathcal{F}_p^t , enabling **Cross-Modal Learning**. The memory bank is continuously updated in a first-in, first-out (FIFO) manner, ensuring effective feature propagation and refinement for future frames.

long-term feature propagation module, which enforces temporal consistency and propagates long-term motion cues (cf. Sec. 4.2), and a **cross-sequence memory alignment** mechanism, which improves robustness by aligning long-term features across diverse driving contexts (cf. Sec. 4.3).

4.1. Cross-View Aggregation

Motivation. Autonomous vehicles utilize multiple cameras to achieve a surround-view perception of their environment, ensuring comprehensive scene coverage. However, due to the inherent overlap in neighboring cameras’ fields of view, redundant yet complementary visual information is captured across multiple perspectives. When LiDAR points are projected onto these camera views, they may be mapped to multiple pixel locations, resulting in inconsistent feature representations. This multi-view ambiguity introduces conflicts during feature alignment, leading to optimization instability and degraded learning efficiency.

Cross-View Aggregation. To address this challenge, we propose a cross-view aggregation mechanism that unifies feature representations across different camera perspectives. Specifically, for each LiDAR point appearing in multiple views, we extract the corresponding pixel-aligned features $\{\mathbf{f}_{(i,j)}^t\}_{j=1}^V$ from V cameras and aggregate them with an averaging operation to obtain the unified feature representation \mathcal{F}_u^t . This operation produces a unified feature representation that mitigates inconsistencies while preserving complementary information from different perspectives. Compared to alternative fusion strategies, such as max or attention-based aggregation, our mean aggregation provides

a balanced fusion of cross-view information while naturally adapting to variations in sensor configurations, ensuring stable and efficient multi-view feature integration.

Role in Our Framework. The cross-view aggregation module enhances the spatial coherence of LiDAR features by unifying multi-view representations, mitigating optimization conflicts, and ensuring stable and efficient training. This strengthens the model’s capacity to learn robust and invariant representations, providing a more consistent and informative feature space for the memory bank.

4.2. Long-Term Feature Propagation

Motivation. Capturing long-range temporal dependencies is crucial for robust autonomous driving perception [42, 48, 77]. However, directly extracting features from historical frames at every timestep incurs significant computational and memory overhead. To tackle this challenge, we propose a long-term feature propagation strategy that efficiently retains and aggregates informative representations from past frames. By ensuring temporal alignment and reducing redundant computations, this approach enables LiDAR features to capture rich motion-aware contextual information while maintaining computational efficiency.

Long-Term Feature Alignment & Propagation. At each timestamp t , we maintain a structured memory bank that stores compact yet informative unified image features from the past k frames, denoted as $\{\mathcal{F}_u^{t-k}, \dots, \mathcal{F}_u^{t-1}\}$. To ensure temporal alignment, we first apply ego-motion transformation, warping historical features into the current frame’s coordinate space through temporal calibration [77]. This

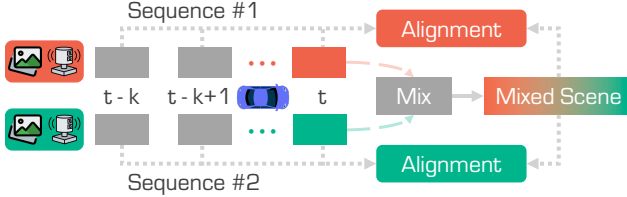


Figure 3. Illustration of **cross-sequence memory alignment**. Existing mixed strategies [32, 82] are utilized to blend two LiDAR scenes from distinct sequences, generating out-of-distribution variations. This process encourages feature alignment across sequences by leveraging memory banks from the original sequences.

yields a set of frame-aligned features $\{\hat{\mathcal{F}}_u^{t-k}, \dots, \hat{\mathcal{F}}_u^{t-1}\}$, effectively compensating for vehicle motion and viewpoint variations. The aligned features are then efficiently retrieved and aggregated with the current frame’s features \mathcal{F}_u^t via average pooling, resulting in a temporally enriched representation \mathcal{F}_d^t . Finally, \mathcal{F}_d^t is distilled into LiDAR representations \mathcal{F}_p^t to facilitate the pretraining objective, as depicted in Eq. (3). To ensure computational efficiency, the memory bank employs a first-in, first-out (FIFO) update strategy, continuously storing the latest \mathcal{F}_u^t while discarding obsolete frames, thereby preserving essential long-term information without incurring excessive memory overhead.

Role in Our Framework. Long-term feature propagation serves as a cornerstone of LiMA, enabling the model to capture temporal dynamics efficiently while avoiding redundant computations. By integrating motion-aware contexture information across frames, our approach significantly enhances the spatial-temporal consistency of LiDAR feature learning. Unlike conventional methods that process each frame independently, LiMA exploits historical context to improve generalization within driving sequences.

4.3. Cross-Sequence Memory Alignment

Motivation. While our framework effectively retains historical context within a single driving sequence, its reliance on intra-sequence consistency constrains its ability to generalize across diverse real-world scenarios. Autonomous driving environments undergo substantial variations in weather, illumination, and road structures, necessitating a mechanism that enforces feature alignment not only within a sequence but also across different driving contexts.

Cross-Sequence Adaptation. To address this challenge, we propose a cross-sequence memory alignment strategy designed to enhance feature consistency and improve generalization across diverse driving scenarios. As illustrated in Fig. 3, given two LiDAR point clouds \mathcal{P}_1 and \mathcal{P}_2 from distinct sequences, we leverage established mixing strategies, such as LaserMix [32] and PolarMix [82], to generate a synthetic mixed scene \mathcal{P}_m . This mixed representation introduces controlled out-of-distribution variations, facilitating alignment across heterogeneous driving conditions. To reinforce feature consistency, the mixed features are op-

timized to maintain structural coherence with the original sequence memory banks, ensuring robust adaptation across both spatial and temporal domains.

Role in Our Framework. The proposed cross-sequence memory alignment functions as a mixed-training strategy designed to bridge domain gaps across driving sequences, maintain long-term feature coherence, and enhance generalization in dynamic real-world settings. By integrating cross-sequence consistency into the learning process, LiMA significantly improves the robustness and transferability of LiDAR-based representations, enabling more reliable downstream perception across diverse environments.

5. Experiments

5.1. Experimental Settings

Datasets. Following standard practice [5, 47, 66], we use *nuScenes* [4] for model pretraining. For downstream evaluation, we evaluate our approach on a diverse set of datasets, including *nuScenes* [4, 17], *SemanticKITTI* [2], *ScribbleKITTI* [76], *Waymo Open* [72], *RELLIS-3D* [28], *SemanticPOSS* [57], *SemanticSTF* [85], *SynLiDAR* [83], *DAPS-3D* [29], *Synth4D* [65], and *nuScenes-C* [31].

Implementation Details. Our experiments are conducted using *MMDetection3D* [12]. We adopt *MinkUNet* [11] for LiDAR semantic segmentation and *VoxelNet* [105] for 3D object detection. The 2D backbone is *ViT* [13] (small, base, large), pretrained with *DINOv2* [56]. Pretraining is performed for 50 epochs on 8 GPUs with a batch size of 2 per GPU, using the AdamW optimizer [51] (initial learning rate 0.01) and the OneCycle scheduler [70]. To capture long-term temporal dependencies, we use 6 consecutive frames. During downstream fine-tuning, task-specific heads are trained with a learning rate $10\times$ higher than the backbone. All evaluations are performed without test-time augmentations for fair comparison. We report mean Intersection-over-Union (mIoU) for segmentation, mean Corruption Error (mCE) and mean Resilience Rate (mRR) for robustness, and mean Average Precision (mAP) and *nuScenes* Detection Score (NDS) for detection.

5.2. Comparative Study

In-Domain Fine-Tuning. We evaluate LiMA against state-of-the-art pretraining methods using both Linear Probing (LP) and few-shot fine-tuning on the *nuScenes* dataset [17]. As shown in Tab. 1, LiMA demonstrates significant improvements under limited annotation settings (e.g., 1%, 5%, and 10%). Specifically, when distilled from ViT-S to ViT-L, LiMA achieves mIoU scores of 54.76%, 56.65%, and 56.67% in the LP, yielding gains of 5.08%, 4.75%, and 4.90% over the baseline [61]. Moreover, LiMA consistently outperforms prior methods across all fine-tuning settings, achieving a 2% to 3% mIoU improvement. These re-

Table 1. Comparisons of state-of-the-art LiDAR pretraining methods pretrained on *nuScenes* [4] and fine-tuned on *nuScenes* [17], *SemanticKITTI* [2], and *Waymo Open* [72] datasets, respectively, with specific data portions. LP denotes linear probing with frozen backbones. All scores are given in percentage (%). The Best and 2nd Best scores under each group are highlighted in Green and Red.

Method	Venue	Backbone (2D)	Backbone (3D)	Frames	LP	nuScenes					KITTI 1%	Waymo 1%
						1%	5%	10%	25%	Full		
Random	-	-	-	-	8.10	30.30	47.84	56.15	65.48	74.66	39.50	39.41
SLiDR [66]	CVPR'22	ResNet-50 [21]	MinkUNet-34 [11]	1	38.80	38.30	52.49	59.84	66.91	74.79	44.60	47.12
ST-SLiDR [52]	CVPR'23			1	40.48	40.75	54.69	60.75	67.70	75.14	44.72	44.93
TriCC [58]	CVPR'23			2	38.00	41.20	54.10	60.40	67.60	75.60	45.90	-
Seal [47]	NeurIPS'23			2	44.95	45.84	55.64	62.97	68.41	75.60	46.63	49.34
CSC [5]	CVPR'24			1	46.00	47.00	57.00	63.30	68.60	75.70	47.20	-
HVDistill [102]	IJCV'24			1	39.50	42.70	56.60	62.90	69.30	76.60	49.70	-
SLiDR [66]	CVPR'22	ViT-S [13]	MinkUNet-34 [11]	1	44.70	41.16	53.65	61.47	66.71	74.20	44.67	47.57
Seal [47]	NeurIPS'23			2	45.16	44.27	55.13	62.46	67.64	75.58	46.51	48.67
SuperFlow [94]	ECCV'24			3	46.44	47.81	59.44	64.47	69.20	76.54	47.97	49.94
ScaLR [61]	CVPR'24			1	49.66	45.89	56.52	61.07	65.79	73.39	46.06	47.67
LiMA	Ours			6	54.76	48.75	60.83	65.41	69.31	76.94	49.28	50.23
SLiDR [66]	CVPR'22			ViT-B [13]	MinkUNet-34 [11]	1	45.35	41.64	55.83	62.68	67.61	74.98
Seal [47]	NeurIPS'23	2	46.59			45.98	57.15	62.79	68.18	75.41	47.24	48.91
SuperFlow [94]	ECCV'24	3	47.66			48.09	59.66	64.52	69.79	76.57	48.40	50.20
ScaLR [61]	CVPR'24	1	51.90			48.90	57.69	62.88	66.85	74.15	47.77	49.38
LiMA	Ours	6	56.65			51.29	61.11	65.62	70.43	76.91	50.44	51.35
SLiDR [66]	CVPR'22	ViT-L [13]	MinkUNet-34 [11]			1	45.70	42.77	57.45	63.20	68.13	75.51
Seal [47]	NeurIPS'23			2	46.81	46.27	58.14	63.27	68.67	75.66	47.55	50.02
SuperFlow [94]	ECCV'24			3	48.01	49.95	60.72	65.09	70.01	77.19	49.07	50.67
ScaLR [61]	CVPR'24			1	51.77	49.13	58.36	62.75	66.80	74.16	48.64	49.72
LiMA	Ours			6	56.67	53.22	62.46	66.00	70.59	77.23	52.29	51.19

Table 2. Domain generalization study of different LiDAR pretraining methods pretrained on the *nuScenes* [4] dataset and fine-tuned on a collection of seven different LiDAR semantic segmentation datasets [28, 29, 57, 65, 76, 83, 85], respectively, with specific data portions. All scores are given in percentage (%). The Best and 2nd Best scores from each metric are highlighted in Green and Red.

Method	Venue	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
SLiDR [66]	CVPR'22	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 [47]	NeurIPS'23	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 [94]	ECCV'24	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
ScaLR [61]	CVPR'24	40.64	52.39	52.53	55.57	53.65	56.86	54.06	55.96	44.42	51.96	81.92	85.58	64.36	67.44
LiMA	Ours	45.90	55.13	55.62	57.15	55.05	57.81	55.45	56.70	46.66	52.32	83.11	86.63	66.04	70.19

sults underscore the effectiveness of long-term distillation in capturing temporal dynamics and maintaining view consistency, leading to more robust feature representations.

Cross-Domain Generalization. To assess the scalability of LiMA, we conduct a comprehensive evaluation across nine diverse 3D semantic segmentation datasets, each representing distinct driving scenarios. As shown in Tab. 1 and Tab. 2, LiMA consistently outperforms baseline methods across all settings. These results highlight our strong feature representation capabilities, demonstrating its adaptability to varying data distributions and its ability to maintain high performance across heterogeneous environments.

Out-of-Distribution Robustness. Evaluating a model’s resilience to out-of-distribution scenarios is critical for assessing its robustness. As shown in Tab. 3, we conduct a comprehensive robustness evaluation on the *nuScenes-C* dataset from Robo3D [31]. LiMA consistently outperforms recent pretraining methods across most corruption types, highlighting its enhanced ability to preserve performance

under adverse conditions. This robustness is particularly important for real-world applications, where models must adapt to diverse and unpredictable environments.

Fine-Tuning for 3D Object Detection. To evaluate the generalization capability of LiMA, we fine-tune our pre-trained model within the SECOND [99] and CenterPoint [100] detection frameworks. As shown in Tab. 4, LiMA yields significant improvements in detection accuracy over state-of-the-art methods. These results demonstrate the effectiveness of LiMA in transferring learned features across different downstream tasks, further validating its robustness and versatility in diverse and challenging scenarios.

Qualitative Results. Fig. 4 visualizes the similarity between a query point and the pretrained 2D image backbone, as well as other LiDAR points. The results demonstrate that long-term information ensures semantic coherence during pretraining, leading to more effective feature learning. Additionally, qualitative segmentation and detection results, presented in Fig. 5 and Fig. 6, respectively, demonstrate that

Table 3. **Out-of-distribution robustness assessment** of LiDAR pretraining methods under corruptions and sensor failures in the *nuScenes-C* dataset from the *Robo3D* benchmark [31]. **Full** denotes fine-tuning with full labels. **LP** denotes linear probing with frozen backbones. All mCE, mRR, and mIoU scores are given in percentage (%). The **Best** and **2nd Best** scores are highlighted in **Green** and **Red**.

#	Method	Venue	mCE ↓	mRR ↑	Fog ↑	Rain ↑	Snow ↑	Blur ↑	Beam ↑	Cross ↑	Echo ↑	Sensor ↑	Average ↑
Full	Random	-	112.20	72.57	62.96	70.65	55.48	51.71	62.01	31.56	59.64	39.41	54.18
Full	SLiDR [66]	CVPR'22	106.08	75.99	65.41	72.31	56.01	56.07	62.87	41.94	61.16	38.90	56.83
	Seal [47]	NeurIPS'23	92.63	83.08	72.66	74.31	66.22	66.14	65.96	57.44	59.87	39.85	62.81
	SuperFlow [94]	ECCV'24	91.67	83.17	70.32	75.77	65.41	61.05	68.09	60.02	58.36	50.41	63.68
	ScaLR [61]	CVPR'24	99.36	80.67	68.43	72.15	66.99	60.07	67.35	44.02	58.98	40.53	59.82
	LiMA	Ours	91.43	82.57	71.24	73.38	67.33	66.73	66.71	47.66	61.72	48.65	62.93
LP	SLiDR [66]	CVPR'22	179.38	77.18	34.88	38.09	32.64	26.44	33.73	20.81	31.54	21.44	29.95
	Seal [47]	NeurIPS'23	166.18	75.38	37.33	42.77	29.93	37.73	40.32	20.31	37.73	24.94	33.88
	SuperFlow [94]	ECCV'24	161.78	75.52	37.59	43.42	37.60	39.57	41.40	23.64	38.03	26.69	35.99
	ScaLR [61]	CVPR'24	150.45	78.24	45.27	50.42	41.75	40.69	44.53	29.93	41.03	31.22	40.61
	LiMA	Ours	137.23	79.30	51.52	54.90	45.63	50.55	49.67	27.24	45.76	34.09	44.92

Table 4. **Comparisons of state-of-the-art LiDAR pretraining methods** pretrained and fine-tuned on the *nuScenes* [4] dataset with specific data portions. All detection methods employ CenterPoint [100] or SECOND [99] as the 3D object detection method.

Method	Venue	nuScenes					
		5%		10%		20%	
		mAP	NDS	mAP	NDS	mAP	NDS
Backbone: VoxelNet [105] + CenterPoint [100]							
Random	-	38.0	44.3	46.9	55.5	50.2	59.7
PointContrast [87]	ECCV'20	39.8	45.1	47.7	56.0	-	-
GCC-3D [41]	ICCV'21	41.1	46.8	48.4	56.7	-	-
SLiDR [66]	CVPR'22	43.3	52.4	47.5	56.8	50.4	59.9
TriCC [58]	CVPR'23	44.6	54.4	48.9	58.1	50.9	60.3
CSC [5]	CVPR'24	45.3	54.2	49.3	58.3	51.9	61.3
ScaLR [61]	CVPR'24	44.3	53.3	48.2	57.1	50.7	60.8
LiMA	Ours	46.5	56.4	50.1	59.6	52.3	62.3
Backbone: VoxelNet [105] + SECOND [99]							
Random	-	35.8	45.9	39.0	51.2	43.1	55.7
SLiDR [66]	CVPR'22	36.6	48.1	39.8	52.1	44.2	56.3
TriCC [58]	CVPR'23	37.8	50.0	41.4	53.5	45.5	57.7
CSC [5]	CVPR'24	38.2	49.4	42.5	54.8	45.6	58.1
ScaLR [61]	CVPR'24	37.3	48.7	41.4	53.5	45.5	58.6
LiMA	Ours	39.4	50.1	43.2	55.3	46.0	59.5

Table 5. Ablation study of **each component** in LiMA. All variants use ViT-B [13] for distillation. **CVA**: Cross-view aggregation. **LTFP**: Long-term feature propagation. **CSMA**: Cross-sequence memory alignment. All scores are given in percentage (%).

#	CVA	LTFP	CSMA	nuScenes			KITTI
	(Sec. 4.1)	(Sec. 4.2)	(Sec. 4.3)	LP	1%	5%	1%
(a)	✗	✗	✗	51.90	48.90	57.69	47.77
(b)	✓	✗	✗	53.72	49.52	58.34	48.75
(c)	✓	✓	✗	55.26	50.43	60.03	49.31
(d)	✓	✓	✓	56.65	51.29	61.11	50.44

LiMA produces more precise predictions, particularly for challenging object classes and dynamic scenes. By effectively integrating spatial and temporal cues, LiMA mitigates misclassifications and enhances object localization, further validating its robustness in real-world scenarios.

5.3. Ablation Study

Component Analysis. Tab. 5 presents an ablation study evaluating the three key components of LiMA, each con-

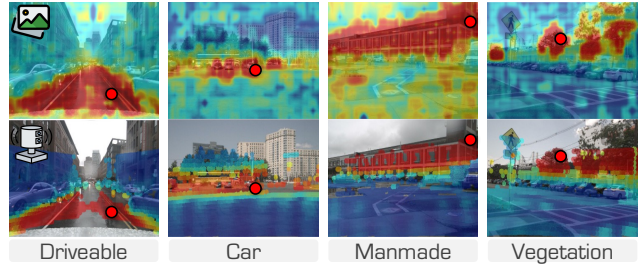


Figure 4. **Cosine similarity** between a query point (marked as red dot) and: (1) image features, and (2) LiDAR features projected onto the image. Colors range from red (indicating high similarity) to blue (indicating low similarity). Best viewed in colors.

Table 6. Ablation study on **aggregation methods** in the cross-view aggregation module. All scores are given in percentage (%).

#	Aggregation Strategy	nuScenes			KITTI	Waymo
		LP	1%	5%	1%	1%
(a)	None	54.64	49.34	59.46	48.02	50.76
(b)	Maximum	55.93	50.02	60.23	49.43	50.91
(c)	Average	56.65	51.29	61.11	50.44	51.35
(d)	Attention Module	55.02	49.87	59.92	49.32	50.23

tributing to enhanced LiDAR representation learning. First, cross-view aggregation improves fine-tuning mIoU by 0.8% and LP by 1.82% by unifying overlapping regions across viewpoints. This ensures consistent point-wise representations and mitigates “optimization conflicts” arising from inconsistent feature learning across views. Second, the long-term feature propagation module facilitates motion pattern learning by leveraging historical image features, leading to an improvement of over 1% in fine-tuning performance and a 1.54% gain in LP. This highlights the importance of incorporating past observations to refine feature representations and enhance temporal consistency. Finally, to improve robustness and adaptability in complex, diverse environments, we introduce the cross-sequence memory alignment. This strategy achieves a 0.9% mIoU gain in fine-tuning and a 1.39% improvement in LP, demonstrating its effectiveness in enabling better generalization across driving scenarios.

Cross-View Feature Aggregation Strategy. Cross-view

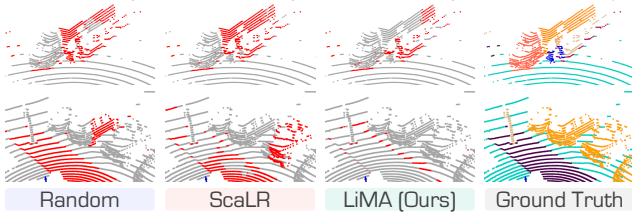


Figure 5. **Qualitative assessments** of state-of-the-art methods, pretrained on *nuScenes* [4] and fine-tuned on *nuScenes* [17] with 1% annotations. The error maps depict **correct** and **incorrect** predictions in **gray** and **red**, respectively. Best viewed in colors.

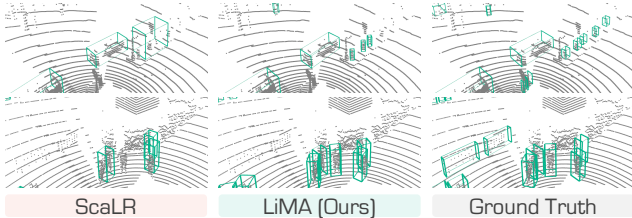


Figure 6. **Visual comparisons** from the 3D object detection task, where methods are pretrained on *nuScenes* [4] and fine-tuned on *nuScenes* [4] with 5% annotations. Best viewed in colors.

Table 7. Ablation study on the **effect of varying frames** for long-term information aggregation. Training time and memory usage are measured with a batch size of 2 on eight A800 GPUs.

Method	Frames	Training Time (Hours)	Memory (GB)	nuScenes LP	nuScenes 1%	KITTI 1%
ScaLR [61]	1	~ 10.1	12.43	51.90	48.90	47.77
LiMA (Ours)	2	~ 14.7	18.23	53.34	49.14	48.27
	3	~ 15.3	20.67	54.52	49.75	48.92
	4	~ 16.1	23.19	55.65	50.29	49.44
	5	~ 17.0	26.59	56.03	50.95	50.32
	6	~ 17.9	29.07	56.65	51.29	50.44
	7	~ 18.7	33.55	55.37	50.91	51.00
	8	~ 19.5	36.27	54.97	49.36	51.78
Seal [47]	2	~ 27.3	20.92	46.59	45.98	47.24
SuperFlow [94]	3	~ 30.7	23.65	47.66	48.09	48.40

feature aggregation is a crucial module for unifying multi-view features. In this ablation study, we evaluate different aggregation methods, including max-pooling, average-pooling, and attention-based approaches as shown in Tab. 6. The results indicate that average-pooling achieves the best performance, as it effectively balances complementary visual features across different views. In contrast, max-pooling captures prominent features but suppresses finer details, leading to suboptimal results. Meanwhile, the attention-based method introduces additional learnable parameters, increasing optimization complexity.

Effects of Memory Bank Frames. We conduct an ablation study to evaluate the impact of memory bank size on long-term information aggregation, as shown in Tab. 7. Increasing the number of stored frames generally improves performance by providing a richer temporal context compared to the spatial alignment baseline [61]. However, beyond an optimal threshold, excessive historical frames introduce

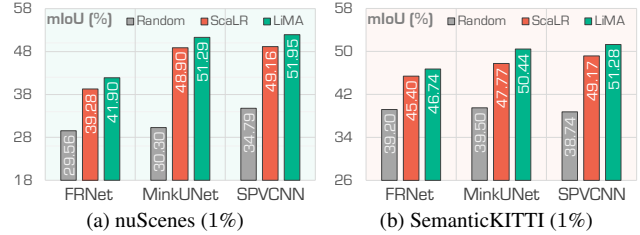


Figure 7. Ablation study on **different backbones** for downstream tasks. The backbones are initialized with random weights, ScaLR [61], and LiMA, respectively, and fine-tuned on the *nuScenes* [17] and *SemanticKITTI* [2] datasets using 1% annotations.

feature misalignment due to calibration errors and temporal drift, leading to inconsistencies in the aggregated representations. This underscores a trade-off between leveraging long-term dependencies and maintaining feature coherence, as accumulating too many past frames may introduce noise from moving scenes rather than beneficial information.

Efficacy Analysis. As shown in Tab. 7, we observe the following: (1) Compared to prior temporal contrastive methods [47, 94], LiMA achieves significantly higher pretraining efficiency, requiring no more than 20 hours even when propagating across 8 frames, whereas prior contrastive methods exceed 24 hours of computation with limited frames. (2) Despite extended temporal modeling, long-term propagation does not substantially increase pretraining time or memory consumption. This efficiency gain stems from the memory bank mechanism, which efficiently stores and retrieves informative features, effectively eliminating redundant computations and optimizing resource utilization.

Representations from 3D Backbones. To evaluate the adaptability of LiMA across different 3D architectures, we replace the 3D backbone with SPVCNN [73] and FRNet [95]. The results in Fig. 7 demonstrate that LiMA consistently outperforms the baseline across diverse representations, underscoring its robustness and flexibility. This highlights LiMA’s ability to effectively integrate with various 3D backbones while maintaining superior performance.

6. Conclusion

This work proposes image-to-LiDAR Memory Aggregation (LiMA), a new data pretraining framework that enhances the robustness and adaptability of LiDAR-based perception systems in dynamic scenarios. LiMA effectively captures temporal dynamics and diversifies training data through three key designs: cross-view aggregation, long-term feature propagation, and cross-sequence memory alignment. These modules enable scalable and effective pretraining. Extensive experiments across various tasks and datasets demonstrate the superiority of LiMA, while evaluations on different 3D backbones highlight its strong generalization capabilities. Our findings position LiMA as a promising approach for developing powerful 3D foundation models.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grants U24B20155 and U21B2044, and the Key Research and Development Program of Jiangsu Province under Grant BE2023016-3. This work was also supported in part by the National Natural Science Foundation of China under Grant 62402245, the Natural Science Foundation of Jiangsu Province under Grant BK20240644, and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 24KJB520023.

References

- [1] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5240–5250, 2023. 2
- [2] 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 *IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 5, 6, 8
- [3] Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. In *International Conference on Learning Representations*, 2025. 1
- [4] 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 multi-modal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 5, 6, 7, 8
- [5] Haoming Chen, Zhizhong Zhang, Yanyun Qu, Ruixin Zhang, Xin Tan, and Yuan Xie. Building a strong pre-training baseline for universal 3d large-scale perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19925–19935, 2024. 5, 6, 7
- [6] Qi Chen, Sourabh Vora, and Oscar Beijbom. Polarstream: Streaming object detection and segmentation with polar pillars. In *Advances in Neural Information Processing Systems*, pages 26871–26883, 2021. 2
- [7] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, Xinge Zhu, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, pages 75896–75910, 2023. 2
- [8] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023. 2
- [9] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *European Conference on Computer Vision*, pages 543–560, 2022. 2
- [10] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. Af2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12547–12556, 2021. 2
- [11] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2, 5, 6
- [12] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d>, 2020. 5
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5, 6, 7
- [14] Ben Fei, Weidong Yang, Liwen Liu, Tianyue Luo, Rui Zhang, Yixuan Li, and Ying He. Self-supervised learning for pre-training 3d point clouds: A survey. *arXiv preprint arXiv:2305.04691*, 2023. 1
- [15] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025. 2
- [16] Martin A Fischler and Robert C Bolles. 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. 3
- [17] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022. 5, 6, 8
- [18] Biao Gao, Yancheng Pan, Chengkun Li, Sibao Geng, and Huijing Zhao. 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. 1
- [19] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2020. 1
- [20] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your hd map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, pages 22441–22482, 2024. 1
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [22] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic

- shifting network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13090–13099, 2021. 2
- [23] Fangzhou Hong, Lingdong Kong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Unified 3d and 4d panoptic segmentation via dynamic shifting networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3480–3495, 2024. 2
- [24] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020. 2
- [25] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549, 2022. 1
- [26] Di Huang, Sida Peng, Tong He, Honghui Yang, Xiaowei Zhou, and Wanli Ouyang. Ponder: Point cloud pre-training via neural rendering. In *IEEE/CVF International Conference on Computer Vision*, pages 16089–16098, 2023. 2
- [27] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 2
- [28] Peng Jiang, Philip Osteen, Maggie Wigness, and Srikanth Saripalli. Rellis-3d dataset: Data, benchmarks and analysis. In *IEEE International Conference on Robotics and Automation*, pages 1110–1116, 2021. 5, 6
- [29] 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. 5, 6
- [30] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023. 2
- [31] 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 *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023. 5, 6, 7
- [32] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023. 2, 5
- [33] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottureau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, pages 21298–21342, 2023. 2
- [34] Lingdong Kong, Xiang Xu, Jun Cen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Calib3d: Calibrating model preferences for reliable 3d scene understanding. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1965–1978, 2025. 1
- [35] Lingdong Kong, Xiang Xu, Youquan Liu, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Largead: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025. 2
- [36] 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*, 47(5):3748–3765, 2025. 2
- [37] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 2
- [38] Li Li, Hubert PH Shum, and Toby P Breckon. Less is more: Reducing task and model complexity for 3d point cloud semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9361–9371, 2023. 2
- [39] Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3707–3717, 2025. 2
- [40] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your lidar placement optimized for 3d scene understanding? In *Advances in Neural Information Processing Systems*, pages 34980–35017, 2024. 1
- [41] Hanxue Liang, Chenhan Jiang, Dapeng Feng, Xin Chen, Hang Xu, Xiaodan Liang, Wei Zhang, Zhenguo Li, and Luc Van Gool. Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 3293–3302, 2021. 7
- [42] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 2, 4
- [43] Xinhao Liu, Moonjun Gong, Qi Fang, Haoyu Xie, Yiming Li, Hang Zhao, and Chen Feng. Lidar-based 4d occupancy completion and forecasting. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 11102–11109, 2024. 1
- [44] Yan Liu, Qingyong Hu, Yinjie Lei, Kai Xu, Jonathan Li, and Yulan Guo. Box2seg: Learning semantics of 3d point clouds with box-level supervision. *arXiv preprint arXiv:2201.02963*, 2022. 2
- [45] Yunze Liu, Junyu Chen, Zekai Zhang, Jingwei Huang, and Li Yi. Leaf: Learning frames for 4d point cloud sequence understanding. In *IEEE/CVF International Conference on Computer Vision*, pages 604–613, 2023. 2
- [46] Youquan Liu, Runnan Chen, Xin Li, Lingdong Kong, Yuchen Yang, Zhaoyang Xia, Yeqi Bai, Xinge Zhu, Yuexin

- Ma, Yikang Li, et al. Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In *IEEE/CVF International Conference on Computer Vision*, pages 21662–21673, 2023. 2
- [47] 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. In *Advances in Neural Information Processing Systems*, pages 37193–37229, 2023. 1, 2, 3, 5, 6, 7, 8
- [48] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petr2: A unified framework for 3d perception from multi-camera images. In *IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 2, 4
- [49] Youquan Liu, Lingdong Kong, Xiaoyang Wu, Runnan Chen, Xin Li, Liang Pan, Ziwei Liu, and Yuexin Ma. Multi-space alignments towards universal lidar segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14661, 2024. 1
- [50] 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. 3
- [51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [52] Anas Mahmoud, Jordan SK Hu, Tianshu Kuai, Ali Harakeh, Liam Paull, and Steven L Waslander. Self-supervised image-to-point distillation via semantically tolerant contrastive loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7102–7110, 2023. 1, 2, 3, 6
- [53] Qinghao Meng, Wenguan Wang, Tianfei Zhou, Jianbing Shen, Yunde Jia, and Luc Van Gool. Towards a weakly supervised framework for 3d point cloud object detection and annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4454–4468, 2021. 2
- [54] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4213–4220, 2019. 2
- [55] Lucas Nunes, Louis Wiesmann, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Temporal consistent 3d lidar representation learning for semantic perception in autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5228, 2023. 2
- [56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5
- [57] Yancheng Pan, Biao Gao, Jilin Mei, Sibao Geng, Chengkun Li, and Huijing Zhao. Semanticpos: A point cloud dataset with large quantity of dynamic instances. In *IEEE Intelligent Vehicles Symposium*, pages 687–693, 2020. 5, 6
- [58] Bo Pang, Hongchi Xia, and Cewu Lu. Unsupervised 3d point cloud representation learning by triangle constrained contrast for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5229–5239, 2023. 1, 3, 6, 7
- [59] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision*, pages 604–621, 2022. 2
- [60] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, Yujing Sun, Tai Wang, Xinge Zhu, and Yuexin Ma. Learning to adapt sam for segmenting cross-domain point clouds. In *European Conference on Computer Vision*, pages 54–71, 2024. 2
- [61] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21519–21529, 2024. 1, 2, 3, 5, 6, 7, 8
- [62] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 2
- [63] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5105–5114, 2017. 2
- [64] Andrey Rudenko, Luigi Palmieri, and Kai O Arras. Joint long-term prediction of human motion using a planning-based social force approach. In *IEEE International Conference on Robotics and Automation*, pages 4571–4577, 2018. 2
- [65] Cristiano Saltori, Evgeny Krivosheev, Stéphane Lathuilière, Nicu Sebe, Fabio Galasso, Giuseppe Fiameni, Elisa Ricci, and Fabio Poiesi. Gipso: Geometrically informed propagation for online adaptation in 3d lidar segmentation. In *European Conference on Computer Vision*, pages 567–585, 2022. 5, 6
- [66] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022. 1, 2, 3, 5, 6, 7
- [67] Corentin Sautier, Gilles Puy, Alexandre Boulch, Renaud Marlet, and Vincent Lepetit. Bevcontrast: Self-supervision in bev space for automotive lidar point clouds. In *International Conference on 3D Vision*, pages 559–568, 2024. 2
- [68] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointcnn: 3d object proposal generation and detection from point cloud. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019. 2
- [69] Hui Shuai, Xiang Xu, and Qingshan Liu. Backward attentive fusing network with local aggregation classifier for 3d point cloud semantic segmentation. *IEEE Transactions on Image Processing*, 30:4973–4984, 2021. 2
- [70] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learn-*

- ing for Multi-Domain Operations Applications*, pages 369–386, 2019. 5
- [71] Jiahao Sun, Chunmei Qing, Xiang Xu, Lingdong Kong, Youquan Liu, Li Li, Chenming Zhu, Jingwei Zhang, Zeqi Xiao, Runnan Chen, et al. An empirical study of training state-of-the-art lidar segmentation models. *arXiv preprint arXiv:2405.14870*, 2024. 1
- [72] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 5, 6
- [73] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, pages 685–702, 2020. 2, 8
- [74] Xiaolin Tang, Kai Yang, Hong Wang, Jiahang Wu, Yechen Qin, Wenhao Yu, and Dongpu Cao. Prediction-uncertainty-aware decision-making for autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, 7(4):849–862, 2022. 2
- [75] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *IEEE/CVF International Conference on Computer Vision*, pages 6411–6420, 2019. 2
- [76] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2697–2707, 2022. 2, 5, 6
- [77] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 2, 4
- [78] Song Wang, Xiaolu Liu, Lingdong Kong, Jianyun Xu, Chunyong Hu, Gongfan Fang, Wentong Li, Jianke Zhu, and Xinchao Wang. Pointlora: Low-rank adaptation with token selection for point cloud learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6605–6615, 2025. 1
- [79] Xuzhi Wang, Wei Feng, Lingdong Kong, and Liang Wan. Nuc-net: Non-uniform cylindrical partition network for efficient lidar semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2
- [80] Xiaopei Wu, Liang Peng, Liang Xie, Yuenan Hou, Binbin Lin, Xiaoshui Huang, Haifeng Liu, Deng Cai, and Wanli Ouyang. Semi-supervised 3d object detection with patchteacher and pillarmix. In *AAAI Conference on Artificial Intelligence*, pages 6153–6161, 2024. 2
- [81] Yanhao Wu, Tong Zhang, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Spatiotemporal self-supervised learning for point clouds in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5251–5260, 2023. 2
- [82] Aoran Xiao, Jiaying Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. Polarmix: A general data augmentation technique for lidar point clouds. In *Advances in Neural Information Processing Systems*, pages 11035–11048, 2022. 5
- [83] Aoran Xiao, Jiaying Huang, Dayan Guan, Fangneng Zhan, and Shijian Lu. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *AAAI Conference on Artificial Intelligence*, pages 2795–2803, 2022. 5, 6
- [84] Aoran Xiao, Jiaying Huang, Dayan Guan, Xiaoqin Zhang, Shijian Lu, and Ling Shao. Unsupervised point cloud representation learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11321–11339, 2023. 1
- [85] Aoran Xiao, Jiaying 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9382–9392, 2023. 5, 6
- [86] Aoran Xiao, Xiaoqin Zhang, Ling Shao, and Shijian Lu. A survey of label-efficient deep learning for 3d point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9139–9160, 2024. 1
- [87] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision*, pages 574–591, 2020. 2, 7
- [88] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*, 2025. 1
- [89] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird’s eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025. 2
- [90] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeeze-seg3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, pages 1–19, 2020. 2
- [91] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 16024–16033, 2021. 2
- [92] Jingyi Xu, Weidong Yang, Lingdong Kong, Youquan Liu, Rui Zhang, Qingyuan Zhou, and Ben Fei. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *arXiv preprint arXiv:2403.10001*, 2024. 2
- [93] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels.

- In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13706–13715, 2020. [2](#)
- [94] 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 *European Conference on Computer Vision*, pages 58–80, 2024. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#)
- [95] Xiang Xu, Lingdong Kong, Hui Shuai, and Qingshan Liu. Frnet: Frustum-range networks for scalable lidar segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025. [2](#), [8](#)
- [96] Xiang Xu, Lingdong Kong, Hui Shuai, Liang Pan, Ziwei Liu, and Qingshan Liu. Limoe: Mixture of lidar representation learners from automotive scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27368–27379, 2025. [2](#)
- [97] Xiang Xu, Lingdong Kong, Hui Shuai, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, and Qingshan Liu. Superflow++: Enhanced spatiotemporal consistency for cross-modal data pretraining. *arXiv preprint arXiv:2503.19912*, 2025. [2](#)
- [98] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, pages 677–695, 2022. [2](#)
- [99] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. [6](#), [7](#)
- [100] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11784–11793, 2021. [6](#), [7](#)
- [101] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022. [2](#)
- [102] Sha Zhang, Jiajun Deng, Lei Bai, Houqiang Li, Wanli Ouyang, and Yanyong Zhang. Hvdistill: Transferring knowledge from images to point clouds via unsupervised hybrid-view distillation. *International Journal of Computer Vision*, pages 1–15, 2024. [1](#), [2](#), [6](#)
- [103] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2020. [2](#)
- [104] Xiao Zheng, Xiaoshui Huang, Guofeng Mei, Yuenan Hou, Zhaoyang Lyu, Bo Dai, Wanli Ouyang, and Yongshun Gong. Point cloud pre-training with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22935–22945, 2024. [2](#)
- [105] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018. [5](#), [7](#)
- [106] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13194–13203, 2021. [2](#)
- [107] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9939–9948, 2021. [2](#)
- [108] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021. [2](#)