HilDA: Hierarchical Distillation with Diffusion for Advancing Self-Supervised LiDAR Pre-training

Maciej Wozniak*[,1], Jesper Ericsson*[1,3], Hariprasath Govindarajan[2,4], Truls Nyberg[1,3], Thomas Gustafsson[3], Patric Jensfelt[1], and Olov Andersson[1]

[1] KTH Royal Institute of Technology, Sweden
{maciejw, jesperik, patric, olovand}@kth.se
[2] Linköping University, Sweden
hariprasath.govindarajan@liu.se
[3] TRATON AB, Sweden
{truls.nyberg, thomas.gustafsson}@scania.com
[4] Qualcomm Auto Ltd Sweden Filial
*Equal contribution.

Abstract. Leveraging Vision Foundation Models (VFMs) for camera-to-LiDAR knowledge distillation offers a promising solution to the scarcity of annotated data needed to represent the immense geometric and kinematic diversity of real-world autonomous driving (AD). However, current approaches typically treat VFMs as black-box teachers, relying exclusively on frame-wise feature similarity. Consequently, they do not fully exploit the teacher's layer-wise semantic structure and global context, as well as the rich spatiotemporal information inherent in LiDAR sequences. We propose **HilDA**, a self-supervised pre-training framework for LiDAR backbones that better captures the semantic *what* and geometric *where* needed for driving tasks. HilDA combines *hierarchical distillation* comprising multi-layer distillation for progressive semantic alignment and global context distillation for scene-level semantics, with a *temporal occupancy diffusion* objective promoting spatiotemporal consistency. Models pre-trained with HilDA achieve state-of-the-art results on cross-modal distillation benchmarks and outperform models trained via prior distillation approaches on 3D object detection, scene flow, and semantic occupancy prediction. Code available at: https://maxiuw.github.io/hilda.

Keywords: Autonomous Driving · SSL · Knowledge Distillation

1 Introduction

Robust spatial perception in dynamic and challenging environments is essential for autonomous driving safety. While LiDAR provides high-fidelity spatial data, training 3D models typically demands massive, labeled datasets. However, the main limitation is not only the annotation cost but also the inability of manual labeling to cover the combinatorial diversity of real-world geometry and motion [21, 55]. To mitigate this challenge, focus has shifted towards Knowledge

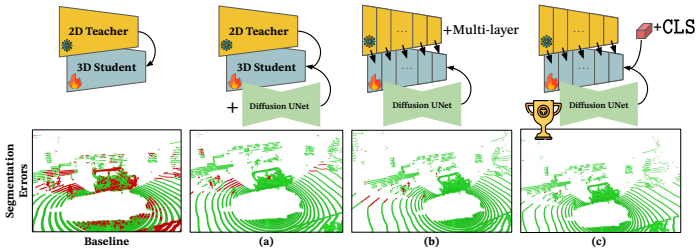

Fig. 1: Effect of HilDA components on semantic accuracy. From baseline (left), we can see error reduction by adding (a) temporal occupancy diffusion, (b) multi-layer distillation, and (c) global context distillation. Best viewed zoomed in 🔍.

Distillation (KD) from Vision Foundation Models (VFMs) into self-supervised LiDAR backbones as a pre-training step [24, 49, 63, 88, 90]. By using VFMs as a teacher, recent work has successfully equipped 3D networks with dense semantic features robust to domain shifts and adverse weather conditions [49, 50, 68].

While these methods established a solid foundation for cross-modal distillation, we observe that they only distill the features from the final layer of the VFM [24, 49, 50, 63, 68, 86, 88, 90]. This approach limits the ability to fully harness the rich representational knowledge that recent work suggests is distributed throughout the VFM's layers [6]. In vision encoders, the multi-layer progression of features across layers captures increasingly abstract and semantically rich information [2]. Because of this, restricting distillation to the final layer may cause two important drawbacks. First, the progressive semantic refinement across layers is potentially lost, limiting the transfer of intermediate representations, which are often better suited for various downstream tasks [6, 20, 34, 95]. Second, the final-layer approach overlooks the global context captured by the `CLS` token [19]. Recent studies on image-based VFMs demonstrate that this token encodes scene-specific information, including the prominence of objects and their relational cues [17, 29, 65]. Instead of distilling only the last layer features, we investigate distilling the hierarchy of how the VFM learns semantics by using both multi-layer features and the global `CLS` token. We term this approach *hierarchical distillation*.

Finally, while VFMs trained on images excel at semantic recognition, they inherently lack explicit information about 3D scene geometry and temporal cues, leaving these aspects underconstrained during pre-training. This issue has been noted in prior cross-modal distillation and pre-training works [24, 52, 104], which introduce an auxiliary occupancy prediction objective to improve downstream performance. However, none of these methods explores the pairing of discriminative and generative priors for spatiotemporal coherence. Recent findings in 2D image space have shown that diffusion-based denoising can serve as geometry-oriented supervision [27,31,51,98,106], encouraging models to internalize spatial manifolds and surface hierarchies through denoising. By combining the semantic *what* from VFMs and the generative *where* from diffusion, we aim to create a representation that is both semantically informed and structurally

robust [99], particularly in the dynamic and occluded environments typical of autonomous driving [11]. Concretely, we introduce a temporal occupancy prediction task with diffusion as an auxiliary pre-training objective and show that this design yields representations that surpass prior methods on cross-modal distillation benchmarks and improve downstream performance on 3D tasks requiring semantic and spatiotemporal reasoning.

To address both the aspects of final-layer distillation and the lack of explicit 3D spatiotemporal supervision, we introduce **HilDA**, a self-supervised pre-training framework that combines hierarchical distillation with temporal occupancy diffusion. Our main contributions are summarized as follows:

- *Hierarchical Distillation:* We introduce a novel cross-modal distillation approach that goes beyond standard final-layer distillation by capturing the semantic evolution of the VFM. This is achieved through a dual mechanism: first, a *multi-layer distillation* strategy utilizes intermediate layers to teach the student LiDAR model how to progressively construct features; second, a *global context distillation* aligns the VFM's `CLS` token with a novel, learnable 3D global context token to promote holistic scene-level representations.
- *Temporal Occupancy Diffusion*: We introduce an auxiliary self-supervised diffusion objective that casts future occupancy prediction as conditional generation, encouraging predictive geometric and motion encoding while addressing the spatiotemporal limits of pure semantic distillation.

Figure 1 visualizes the effect of progressively incorporating temporal occupancy diffusion and hierarchical distillation. Compared to standard final-layer distillation, adding consecutive components yields visibly reduced segmentation errors, indicating increasingly structured and semantically aligned LiDAR representations. Extensive experiments on autonomous driving benchmarks in Sec. 4 demonstrate that HilDA achieves state-of-the-art (SOTA) performance in camera-LiDAR cross-modal distillation. Whether evaluated via linear probing, few-shot transfer, or as an initialization for full supervision, our approach consistently outperforms existing methods. Furthermore, our distillation strategy produces features with enhanced multi-task transferability. Beyond 3D semantic segmentation, our model outperforms previous SOTA distillation approaches in 3D object detection, scene flow, and semantic occupancy prediction.

2 Related Work

Cross-Modal Distillation Methods. In VFMs, the pre-training objective strongly shapes the learned representations. For example, global alignment enhances aggregated feature semantics [9, 25, 64], while dense prediction localizes features and promotes boundary adherence [39, 107]. When transferring such features, early camera–LiDAR distillation methods largely relied on contrastive objectives, which can be sensitive to noisy point–pixel correspondences, as highlighted by [24, 63]. Several methods [49, 68, 88, 103] therefore used grouping or semantic priors (*e.g.*, SAM [39] or OpenSeed [97] pseudo-masks). While helpful,

these priors add overhead, and contrastive distillation can still exhibit "self-conflict" by sampling semantically similar regions as negatives [49, 63, 68, 88, 103]. Recent work instead favors direct feature alignment [24, 63, 86, 90], using *e.g.*, expressive 2D-to-3D transfer [24] or scaling teacher/student models and data [63].

Because VFM layers differ in semantic abstraction [6, 72, 92], the progressively learned feature hierarchy is distributed across intermediate layers rather than confined to the final output [71,93]. This hierarchy contains signals about "how" to learn the features. Yet both contrastive and direct-alignment pipelines supervise only the final-layer [24, 49, 50, 63, 68, 86, 88, 90] and discard the teacher `CLS` token [19], limiting global alignment despite its shown benefit in the 2D setting [29, 65]. Rather than relying on computationally heavy grouping priors or data scaling, we leverage the VFM's hierarchical progression via multi-layer distillation and global `CLS` token distillation to strengthen cross-modal alignment.

Geometric and Temporal Auxiliaries. Image-based VFMs transfer semantics but typically lack explicit 3D geometric or temporal supervision. Many camera–LiDAR distillation methods prioritize cross-modal feature alignment [86], or semantic-temporal consistency objectives [88, 90]. Recent approaches [24, 104] and representation learning works [1, 18, 52] increasingly use occupancy prediction to impose spatiotemporal priors. While effective, discriminative occupancy prediction trained with per-query/voxel binary cross-entropy supervision models local marginals rather than the joint scene distribution. Shared receptive fields may implicitly promote consistency, but this objective provides no explicit incentive for globally coherent configurations, limiting its ability to capture complex 3D structure [79]. In practice, occupancy is correlated across space and time, forming continuous structures [81] and maintaining object permanence [54]. This motivates the use of generative objectives that can model occupancy jointly and impose scene-level spatiotemporal structure through iterative refinement [26].

Diffusion models [30, 66], analogous to denoising autoencoders trained across noise levels [13, 84, 91], define a coarse-to-fine refinement process over dense structured targets [75]. Applied to spatiotemporal denoising, they can learn to represent stochastic future evolution by encoding global scene dynamics [74]. In autonomous driving, diffusion is used for spatiotemporal prediction and refinement, including LiDAR completion [60], world-model rollouts [76, 100, 105], BEV/scene-flow refinement [47, 94], and as action decoders [22, 45, 78]. While the diffusion objective alone can overemphasize local detail and weaken semantic discrimination, pairing it with robust discriminative teachers (*e.g.*, DINOv2 [61]) is shown to preserve semantics while improving geometric priors [44, 99]. Motivated by this complementarity, we use diffusion as a pre-training objective [31], denoising future occupancy from past context across noise levels to inject multi-scale geometric and motion cues into the LiDAR encoder.

3 Method

In this section, we define the pre-training formulation and present the components comprising our framework **HilDA**. Figure 2 shows an architectural overview.

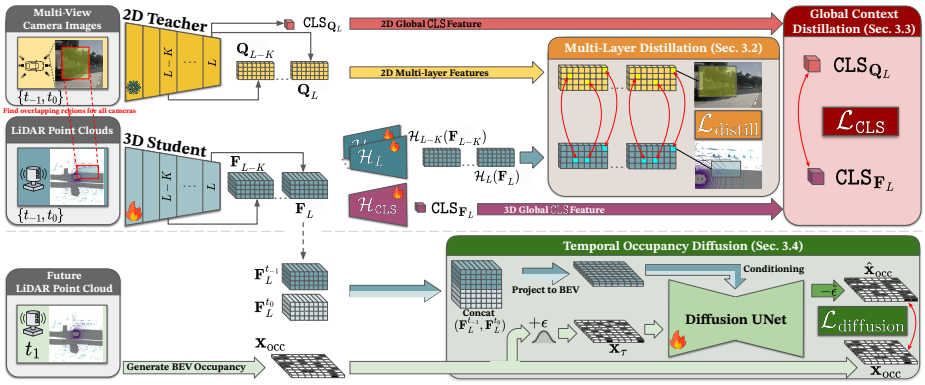

Fig. 2: Overview of HilDA. With LiDAR sweeps and synchronized multi-view images, HilDA learns a 3D backbone using three self-supervised objectives: **(1)** multi-layer point–pixel distillation from a frozen VFM, **(2)** global context distillation via `CLS` tokens (t_{-1} and t_0 are processed independently through the distillation modules), and **(3)** temporal occupancy diffusion conditioned on the student features ($\mathbf{F}_L^{t_{-1}}, \mathbf{F}_L^{t_0}$). Features are extracted from t_{-1} and t_0, while t_1 constructs the future BEV occupancy target.

3.1 Problem Definition and Motivation

The primary objective of cross-modal knowledge distillation is to pre-train a LiDAR student network by transferring rich and dense semantic knowledge from a VFM while also learning geometric features exclusive to the LiDAR modality. For this task, let the input be a point cloud $\mathcal{P} = \{\mathbf{p}_i \mid i = 1, \ldots, N\}$ of N points, where each point $\mathbf{p}_i \in \mathbb{R}^4$ denotes the 3D spatial coordinates (x, y, z) and intensity. Let V be the number of synchronized cameras forming a surround view $\mathcal{I} = \{\mathbf{I}_c \mid c = 1, \ldots, V\}$, where $\mathbf{I}_c \in \mathbb{R}^{H \times W \times 3}$ is the RGB-image from camera c.

The challenge lies in the fundamental discrepancy between modalities: LiDAR data is geometrically precise but semantically sparse and texture-less, whereas camera images are semantically dense but lack explicit depth or metric 3D structure. We aim to pre-train a LiDAR backbone $S_\theta(\cdot)$ in a self-supervised manner to inherit visual semantics without requiring manual annotations, while simultaneously learning spatiotemporal features through a generative auxiliary task. The outcome of our self-supervised pre-training is a backbone S_θ that can produce robust feature representations useful for various perception tasks (note, for inference we remove all other components used in pre-training). HilDA processes temporal sequences of three frames $\{t_{-1}, t_0, t_1\}$. Feature extraction is restricted to the past and present frames, $\{t_{-1}, t_0\}$, while the future frame t_1 serves exclusively as the target for occupancy prediction. We omit the temporal index in the two following Secs. 3.2 and 3.3 for notational simplicity.

3.2 Multi-Layer Distillation

To effectively transfer semantic knowledge, we employ a dense-to-sparse distillation strategy. This process relies strictly on geometric calibration, avoiding the need for semantic priors. We find corresponding points and pixels by projecting the point coordinates (x_i, y_i, z_i) onto the image plane $(u_i, v_i)_c$ of camera c using calibration information. The following sensor calibration parameters are used: $\begin{bmatrix} u_i & v_i & 1 \end{bmatrix}^T = \rho(i) = \frac{1}{z_i} \times \Gamma_I \times \Gamma_{c \leftarrow \text{LiDAR}} \times \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^T$, where Γ_I denotes the camera-intrinsic matrix and $\Gamma_{c \leftarrow \text{LiDAR}}$ is the transformation matrix from the LiDAR sensor to each camera. This enables a mapping between points and pixels for each camera image. Let $\mathcal{M}_c$ be the set of valid point–pixel correspondences for image $\mathbf{I}_c$. We define a pair $(i, j) \in \mathcal{M}_c$ if the projection of the 3D point $\mathbf{p}_i$ onto camera c maps to the pixel with index j. Points outside $\mathcal{M}_c$ are excluded from multi-layer distillation but still contribute to the losses in Secs. 3.3 and 3.4.

We distill features from multiple teacher layers into the corresponding student layers to capture not only the teacher's final representation, but also how features form across the VFM hierarchy. By default, we use the last two layers. Let L be the index of the final output layer for both teacher and student[5] (the two networks may have different total depths). We define the layer index $\ell \in \{L-K, \ldots, L\}$. We denote the student point-feature tensor at layer ℓ by $\mathbf{F}_\ell \in \mathbb{R}^{N \times C_\ell}$, and the feature of point i by $\mathbf{f}_{i,\ell} = \mathbf{F}_\ell[i] \in \mathbb{R}^{C_\ell}$. Let $\mathbf{Q}_{\ell,c} \in \mathbb{R}^{H \times W \times D_\ell}$ be the feature map at the corresponding teacher layer for camera c, bi-linearly up-sampled to the image resolution $H \times W$. From this, we can obtain the teacher feature at pixel j as $\mathbf{q}_{j,\ell,c} = \mathbf{Q}_{\ell,c}[j] \in \mathbb{R}^{D_\ell}$. We apply a lightweight, layer-specific MLP $\mathcal{H}_\ell : \mathbb{R}^{C_\ell} \to \mathbb{R}^{D_\ell}$ to align the teacher and student feature dimensions (see Fig. 2). Cosine distance is then minimized over point-pixel pairs across layers and images:

$$\mathcal{L}_{\text{distill}} = \frac{1}{V(K+1)} \sum_{\ell=L-K}^{L} \sum_{c=1}^{V} \frac{1}{|\mathcal{M}_c|} \sum_{(i,j) \in \mathcal{M}_c} \left(1 - \frac{\mathcal{H}_\ell(\mathbf{f}_{i,\ell}) \cdot \mathbf{q}_{j,\ell,c}}{\|\mathcal{H}_\ell(\mathbf{f}_{i,\ell})\|_2 \, \|\mathbf{q}_{j,\ell,c}\|_2} \right). \quad (1)$$

By maximizing similarity across layers, the 3D student learns to progressively encode visual concepts using an information hierarchy similar to the teacher.

3.3 Global Context Distillation

Local point-pixel distillation ensures fine-grained alignment but may miss the holistic context of a scene (*e.g.*, distinguishing a highway environment from a residential area). ViT-based VFMs inherently encode global context in their `CLS` token. Thus, as part of our novel hierarchical distillation strategy, we complement the local multi-layer distillation loss in Eq. (1) with global context distillation.

To construct a unified visual scene descriptor $\texttt{CLS}_{\mathbf{Q}_L}$, we extract the `CLS` tokens from the teacher's final layer L for all V camera images and apply a

[5] In the case of MinkUnet34 [15], it consists of "planes" that have multiple layers. We refer to these planes as "student layers". For the teacher, we select adjacent ViT transformer blocks and refer to these as "teacher layers". Thus, each selected feature layer from the student and teacher includes a non-trivial stage/block-level transformation rather than a single atomic layer.

max-pooling across them. Correspondingly, for the LiDAR scene $\mathcal{P}$, we design a global context "token" $\texttt{CLS}_{\mathbf{F}_L}$ by applying a dedicated MLP projection head $\mathcal{H}_{\texttt{CLS}}(\cdot)$ to the features from the final layer L of the student network, followed by global max-pooling over all point features in the scene. Max pooling highlights the most active features. We then align the global representation of the student with the teacher by minimizing the Mean Squared Error (MSE):

$$\mathcal{L}_{\texttt{CLS}} = \|\texttt{CLS}_{\mathbf{F}_L} - \texttt{CLS}_{\mathbf{Q}_L}\|_2^2. \tag{2}$$

This objective acts as a scene-level regularizer, encouraging the 3D backbone to aggregate global context by matching the student's pooled representation to the VFM's `CLS` embedding. This emphasizes salient view- and point-level activations to align global semantic cues between the student and the teacher. Section A and Tab. 9 ablates different loss formulations for Eq. (2) as well as different pooling methods for constructing $\texttt{CLS}_{\mathbf{Q}_L}$ and $\texttt{CLS}_{\mathbf{F}_L}$.

3.4 Spatiotemporal Self-Supervised Learning as an Auxiliary Task

Distillation provides semantic context but does not explicitly model spatiotemporal scene dynamics. We therefore add a label-free auxiliary task that predicts future BEV occupancy with a conditional diffusion model, encouraging the 3D student S_θ to learn predictive spatial representations [31, 62]. Given two LiDAR sweeps $\mathcal{P}^{t_{-1}}$ and $\mathcal{P}^{t_0}$, S_θ encodes them into features $\mathbf{F}_L^{t_{-1}}$ and $\mathbf{F}_L^{t_0}$, which are collapsed into a shared dense BEV history feature map $\mathbf{C}_{\text{history}}$. The future sweep $\mathcal{P}^{t_1}$ is transformed into the t_0 frame and projected to a ground-removed BEV occupancy target $\mathbf{x}_{\text{occ}} \in \{0,1\}^{H_{\text{BEV}} \times W_{\text{BEV}}}$. We formulate future occupancy prediction as conditional DDPM denoising [30]. At diffusion step τ, the target occupancy is corrupted as $\mathbf{x}_\tau = \sqrt{\bar{\alpha}_\tau}\mathbf{x}_{\text{occ}} + \sqrt{1-\bar{\alpha}_\tau}\epsilon$, with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_\tau = \prod_{s=1}^{\tau}(1-\beta_s)$. A denoising network (2D UNet [67]) then predicts $\epsilon_\theta(\mathbf{x}_\tau, \tau, \mathbf{C}_{\text{history}})$, as shown in Fig. 2. Conditioning is implemented via channel-wise concatenation of $\mathbf{x}_\tau$ and $\mathbf{C}_{\text{history}}$, following [43]. We train it with the hybrid objective

$$\mathcal{L}_{\text{diffusion}} = \underbrace{\mathbb{E}_{\tau,\epsilon,\mathbf{C}_{\text{history}}}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{x}_\tau, \tau, \mathbf{C}_{\text{history}})\right\|_2^2\right]}_{\text{Noise Prediction}} + \lambda \underbrace{\left\|\mathbf{x}_{\text{occ}} - \hat{\mathbf{x}}_{\text{occ}}\right\|_2^2}_{\text{Occ. Reconstr.}} \tag{3}$$

where the first term supervises noise prediction, while the second, weighted by λ, provides a complementary reconstruction signal [58] which improves conditioning at low τ (see Sec. A.4 for ablation). The reconstructed occupancy $\hat{\mathbf{x}}_{\text{occ}}$ is recovered from the noisy sample by inverting the forward process using the predicted noise. Training across varying τ exposes the model to denoising tasks at different noise levels, encouraging a coarse-to-fine behavior from global structure to local detail. Full implementation details and target construction are provided in Sec. C.

3.5 Final Pre-Training Objective

HilDA is trained end-to-end on three objectives, encouraging the 3D student to learn detailed semantics via distillation, while jointly learning scene geometry

and dynamics via diffusion-based temporal occupancy prediction:

$$\mathcal{L}_{\text{total}} = \omega_{ds}\mathcal{L}_{\text{distill}} + \omega_{gl}\mathcal{L}_{\texttt{CLS}} + \omega_{df}\mathcal{L}_{\text{diffusion}}. \tag{4}$$

4 Experiments

4.1 Experimental Setup

We distill from DINOv2 [61] to MinkUnet34 [15], the standard backbone in prior camera–LiDAR distillation work. Our method also supports other 3D backbones, *e.g.*, PTv3 [83] (supplementary Sec. A). Each teacher–student configuration is pre-trained *once* on the nuScenes training split [8] using synchronized RGB–LiDAR data and calibration only; no task labels are used. The resulting backbone is transferred to *all* downstream benchmarks *without* target-dataset re-pre-training. Table 1 varies the DINOv2 teacher size, while Sec. 4.3 uses the same ViT-B/MinkUnet34 checkpoint for all transfer tasks. We evaluate segmentation on nuScenes [8], SemanticKITTI [4], Waymo [70], ScribbleKITTI [73], RELLIS-3D [32], SemanticSTF [85], and DAPS-3D [40]; robustness on nuScenes-C [41]; 3D detection on KITTI [23] and nuScenes; semantic occupancy on nuScenes; and scene flow on Argoverse 2 [82]. Linear probing (LP) freezes the pre-trained backbone and trains only a linear head, whereas data-scarce/full fine-tuning updates the full network using the indicated label fraction. For semantic occupancy, we freeze the backbone and train a lightweight decoder. Metrics are mIoU for segmentation/occupancy, mAP for detection, EPE-based metrics for scene flow, and mCE/mRR for robustness. We use author-reported results unless stated otherwise and provide details in supplementary Sec. C. HilDA† denotes HilDA without the diffusion auxiliary pre-training loss; it is excluded from **best**/$\underline{2^{nd}\text{-best}}$ ranking to emphasize the comparison of HilDA against prior SOTA.

4.2 Cross-Modal Distillation Benchmarks

In the next three tables, we compare our method with previous SOTA approaches on standard cross-modal distillation benchmarks for 3D semantic segmentation.

Main Results. As shown in Tab. 1, **HilDA** achieves SOTA performance across all scenarios (except for one - second best). The most significant gains are observed in the data-scarce scenarios (1%–10%) where our method outperforms previous SOTA by a significant margin, including CleverDistiller [24], LiMoE [86], and ScaLR [63]. While LiMA [90] previously set a clear margin on linear probing (LP), our method surpasses it while also delivering stronger fine-tuning results. This suggests that our hierarchical distillation with a temporal diffusion strategy yields a richer representation than prior methods relying on *e.g.*, global contrastive learning or semantic priors, thereby reducing dependence on large-scale annotated data. Additionally, it substantially improves transfer to new domains like Semantic-KITTI [4] or the Waymo Open Dataset [70]. This pattern holds across different teacher sizes (ViT-S, ViT-B, and ViT-L). We also qualitatively observe this in Fig. 3, where our features yield notably lower segmentation error and work much better for long-tail cases (*e.g.*, a person on the top of a truck).

Table 1: 3D semantic segmentation results for cross-modal distillation methods pre-trained on nuScenes [8], and fine-tuned on nuScenes, SemanticKITTI [4], and Waymo [70]. All methods use the MinkUnet34 [15] backbone. LP denotes linear probing with a frozen backbone. "S.P." refers to usage of semantic priors.

Method	Venue	Teacher	S.P.	nuScenes						SKITTI	Waymo
				LP	1%	5%	10%	25%	Full	1%	1%
Random	-	-	✗	8.10	30.30	47.84	56.15	65.48	74.66	39.50	39.41
SLidR [68]	CVPR'22	ViT-S	✓	44.70	41.16	53.65	61.47	66.71	74.20	44.67	47.57
Seal [49]	NIPS'23	ViT-S	✓	45.16	44.27	55.13	62.46	67.64	75.58	46.51	48.67
SuperFlow [88]	ECCV'24	ViT-S	✓	46.44	47.81	59.44	64.47	69.20	76.54	47.97	49.94
LiMoE [87]	CVPR'25	ViT-S	✓	48.20	49.60	60.54	65.65	71.39	77.27	49.53	51.42
CleverDistiller [24]	BMVC'25	ViT-S	✗	49.81	56.90	64.55	65.92	70.11	77.61	50.59	50.99
SuperFlow++ [89]	TPAMI'25	ViT-S	✓	48.57	49.07	60.57	65.21	70.05	76.92	49.27	51.25
LiMA [90]	ICCV'25	ViT-S	✗	54.76	48.75	60.83	65.41	69.31	76.94	49.28	50.23
HilDA †	-	ViT-S	✗	55.13	57.81	66.26	67.32	70.54	77.17	50.80	50.87
HilDA	-	ViT-S	✗	**56.29**	**59.46**	**68.57**	**70.22**	**73.94**	**78.15**	**52.88**	**52.16**
SLidR [68]	CVPR'22	ViT-B	✓	45.35	41.64	55.83	62.68	67.61	74.98	45.50	48.32
Seal [49]	NIPS'23	ViT-B	✓	46.59	45.98	57.15	62.79	68.18	75.41	47.24	48.91
SuperFlow [88]	ECCV'24	ViT-B	✓	47.66	48.09	59.66	64.52	69.79	76.57	48.40	50.20
ScaLR [63]	CVPR'24	ViT-B	✗	41.80	55.83	63.46	65.24	68.70	74.76	45.59	49.60
LiMoE [87]	CVPR'25	ViT-B	✓	49.07	50.23	61.51	66.17	71.56	77.81	50.30	51.77
CleverDistiller [24]	BMVC'25	ViT-B	✗	51.89	59.80	66.44	67.65	69.53	78.49	51.48	53.56
SuperFlow++ [89]	TPAMI'25	ViT-B	✓	48.86	49.56	60.75	65.46	70.19	77.29	49.90	51.65
LiMA [90]	ICCV'25	ViT-B	✗	56.65	51.29	61.11	65.62	70.43	76.91	50.44	51.35
HilDA †	-	ViT-B	✗	56.51	61.01	67.17	70.13	72.83	77.53	51.58	52.11
HilDA	-	ViT-B	✗	**58.95**	**62.71**	**70.19**	**71.00**	**73.68**	**79.12**	**53.44**	**53.89**
SLidR [68]	CVPR'22	ViT-L	✓	45.70	42.77	57.45	63.20	68.13	75.51	47.01	48.60
Seal [49]	NIPS'23	ViT-L	✓	46.81	46.27	58.14	63.27	68.67	75.66	47.55	50.02
SuperFlow [88]	ECCV'24	ViT-L	✓	48.01	49.95	60.72	65.09	70.01	77.19	49.07	50.67
ScaLR [63]	CVPR'24	ViT-L	✗	40.12	55.78	63.28	64.76	68.19	75.09	44.85	50.34
LiMoE [87]	CVPR'25	ViT-L	✓	49.35	51.41	62.07	66.64	71.59	77.85	50.69	51.93
CleverDistiller [24]	BMVC'25	ViT-L	✗	52.45	60.64	67.03	67.29	70.45	78.29	52.28	**54.83**
SuperFlow++ [89]	TPAMI'25	ViT-L	✓	49.78	50.92	61.83	66.30	71.07	77.63	50.33	52.12
LiMA [90]	ICCV'25	ViT-L	✗	56.67	53.22	62.46	66.00	70.59	77.23	52.29	51.19
HilDA †	-	ViT-L	✗	56.91	61.60	67.57	70.71	72.68	78.47	51.99	52.64
HilDA	-	ViT-L	✗	**60.06**	**62.82**	**70.58**	**72.55**	**75.06**	**79.44**	**54.48**	54.41

Domain Generalization. In Tab. 2, we evaluate the transferability of our model, pre-trained on nuScenes, to four diverse datasets. The results show that our learned representations are transferable to various sensor configurations and urban layouts. On ScriKITTI [73], despite extremely sparse scribble-level annotations, our method maintains top performance, demonstrating its ability to extract meaningful features from minimal and spatially disconnected supervision. On Rellis-3D [32], which is recorded in an unstructured off-road environment, our model outperforms strong baselines such as LiMoE [86]. This indicates that our approach captures fundamental geometric and semantic properties that generalize across drastically different terrains. Consistent gains on SemSTF [85] and DAPS-3D [40] further confirm that our pre-training produces robust, dataset-agnostic features serving as strong universal initializations.

Robustness. In Tab. 3 (and supplementary Sec. A for full fine-tuning), we evaluate feature robustness on the corrupted nuScenes-C [41] benchmark. HilDA exhibits superior resilience compared to previous methods, achieving the lowest mean Corruption Error (mCE) and highest mean Relative Robustness (mRR).

It performs particularly well under *Cross-Sensor* and *Snowy* corruptions, which involve heavy interference and sparsity, clearly outperforming previous SOTA methods in these scenarios. These results confirm that our combination of

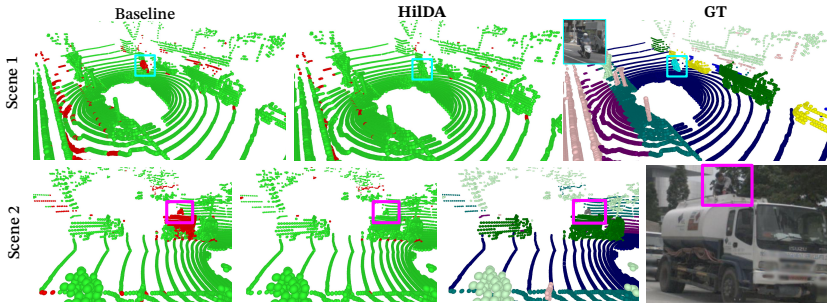

Fig. 3: Qualitative comparison of segmentation performance of HilDA and ScaLR. HilDA makes fewer errors and correctly segments rare cases like a scooter driver (Scene 1) or a person on top of a truck (Scene 2).

Table 2: Domain generalization benchmarks. Models pre-trained on nuScenes and fine-tuned on listed datasets. Metrics in mIoU.

Method	ScriKITTI		Rellis-3D		SemSTF		DAPS-3D	
	1%	10%	1%	10%	50%	100%	50%	100%
Random	23.81	47.60	38.46	53.60	48.03	48.15	74.32	79.38
SLidR	39.60	50.45	49.75	54.57	52.01	54.35	81.00	85.40
Seal	40.64	52.77	51.09	55.03	53.46	55.36	81.88	85.90
ScaLR	36.45	49.16	47.91	48.86	52.10	54.40	81.92	85.58
SuperFlow	42.70	54.00	52.83	55.71	54.72	56.57	82.43	86.21
LiMoE	43.95	55.96	53.74	56.67	55.60	57.31	83.24	86.68
CleverD	44.03	56.70	58.35	60.92	53.99	55.66	83.06	87.95
LiMA	45.90	55.13	55.62	57.15	55.45	56.70	83.11	86.63
HilDA†	48.26	59.02	56.77	59.04	55.86	58.11	84.26	87.18
HilDA	**48.39**	**58.22**	**59.49**	**62.88**	**56.72**	**57.83**	**85.93**	**89.08**

hierarchical distillation and temporal occupancy prediction produces reliable 3D representations even under severe sensor degradation.

4.3 Transferability to Spatiotemporal Tasks

We further assess transfer to tasks requiring spatial and temporal reasoning: 3D object detection (spatial), semantic occupancy (spatiotemporal), and scene flow (spatiotemporal). All experiments use the same ViT-B/MinkUnet34 backbone after pre-training. We re-trained ScaLR, SuperFlow, and CleverDistiller (SF/CD) using author settings and verified that their segmentation performance matches the reported results[6]; details are in supplementary Sec. C.

3D Object Detection. We evaluate PointRCNN [69] with HilDA on KITTI [23] and nuScenes [8]. As shown in Tab. 4, HilDA consistently outperforms prior distillation baselines across datasets. Notably, we find that our model demonstrates

[6] Checkpoints/complete code for LiMoE/LiMA is unavailable, authors reported lost access. CD (code provided by CD's authors)/ScaLR/SF code was used for obtaining the checkpoints.

Table 3: Comparison on robustness under weather corruption and sensor failure on nuScenes-C [41] benchmark using linear probing. Metrics given in percentage (%).

Method	**mCE** ↓	**mRR** ↑	mIoU ↑								
			Fog	Rain	Snow	Blur	Beam	Cross	Echo	Sensor	**Mean**
SLidR [68]	179.38	77.18	34.88	38.09	32.64	26.44	33.73	20.81	31.54	21.44	29.95
Seal [49]	166.18	75.38	37.33	42.77	29.93	37.73	40.32	20.31	37.73	24.94	33.88
SuperFlow [88]	161.78	75.52	37.59	43.42	37.60	39.57	41.40	23.64	38.03	26.69	35.99
Scalr [63]	173.18	78.91	37.55	37.96	36.29	33.64	33.06	23.01	33.62	23.70	33.59
LiMoE [86]	155.77	78.23	40.35	45.28	39.14	42.10	44.21	27.33	39.20	29.49	38.39
CleverD [24]	151.21	79.76	43.96	46.91	41.20	41.05	42.15	45.67	41.30	28.85	41.39
LiMA [90]	137.23	79.30	51.52	54.90	45.63	50.55	49.67	27.24	45.76	34.09	44.92
HilDA†	134.63	80.70	50.41	55.86	49.63	49.95	45.25	50.18	44.09	34.44	47.58
HilDA	**124.27**	**88.20**	**55.99**	**57.97**	**54.29**	**51.16**	**51.30**	**56.29**	**47.91**	**40.31**	**52.00**

robust detections in complex scenarios, such as at long-range and under heavy occlusion (see Fig. 4). This suggests that hierarchical distillation with temporal diffusion yields semantically consistent and geometrically grounded features.

Table 4: Evaluation on 3D Object Detection (3DOD) benchmarks.

Method	KITTI (mAP)			nuSc (mAP)		
	5%	10%	20%	5%	10%	20%
Random	56.1	59.1	61.6	38.1	43.5	45.2
PPKT	57.8	60.1	61.2	-	-	-
SLidR	57.8	61.4	62.4	-	-	-
ScaLR	56.2	62.3	66.5	46.1	50.3	55.1
SF	59.3	62.7	64.2	45.9	51.1	54.5
CD	59.8	66.6	67.1	47.9	51.9	54.3
HilDA†	60.1	66.3	69.4	47.0	52.2	55.7
HilDA	**61.4**	**67.3**	**71.0**	**50.4**	**54.9**	**57.9**

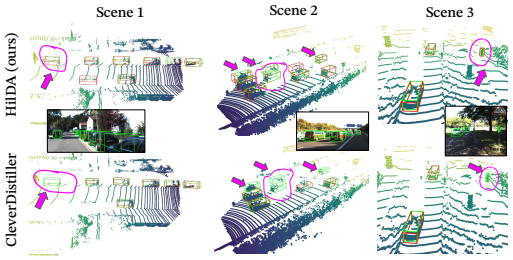

Fig. 4: Qualitative comparison of 3DOD. Green boxes are GT, and red are predictions.

Semantic Occupancy. We evaluate 3D/4D semantic occupancy following Occ4cast [48], with added class labels. The pre-trained backbone is frozen and only a lightweight decoder is trained. Inputs are t_{-1} and t_0 ($\Delta t = 0.5$s). We predict occupancy at t_0 for 3D, and at t_0 and t_1 for 4D, and report frame-averaged metrics. Table 5 aggregates IoU into *dynamic*, *static*, and *surface* classes (see Sec. A for a class-wise breakdown across time horizons and a detailed discussion on the effects of including the diffusion auxiliary task. Section B shows qualitative results). HilDA outperforms all baselines in 3D and 4D. The largest gains from including the diffusion-based loss during pre-training are on dynamic/object-centric classes where future occupancy directly supervises object permanence and short-term motion. Surface gains are smaller because the ground-removed pre-training BEV target emphasizes free/occupied boundaries and removes vertical cues useful for within-region semantic distinctions (*e.g.*, terrain *vs.* sidewalk). Figure 5 shows that HilDA maintains the highest mIoU across a horizon of up to 5 seconds.

Table 5: Comparison of frozen pre-trained backbones on nuScenes semantic occupancy prediction. Frame-averaged metrics.

	Method	mIoU	Dyna.	Stat.	Surf.
3D	Random	10.6	3.7	10.6	24.4
	ScaLR	16.2	9.5	16.4	29.4
	SF	12.5	5.2	12.1	25.7
	CD	16.5	10.0	16.5	29.4
	HilDA†	19.0	13.3	19.0	30.2
	HilDA	**20.0**	**14.5**	**20.2**	**30.4**
4D	Random	10.3	3.1	10.7	24.4
	ScaLR	14.9	7.1	15.9	29.5
	SF	11.9	4.9	12.2	25.6
	CD	15.5	8.0	16.5	29.6
	HilDA†	17.3	10.1	18.5	30.4
	HilDA	**18.4**	**11.4**	**20.2**	**30.5**

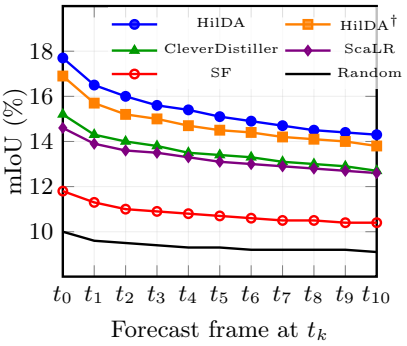

Fig. 5: Semantic mIoU over forecast horizon. Decoder trained from t_0 through t_{10}.

Table 6: Scene flow on ArgoverseV2 [82]. Ego Motion refers to the apparent motion of points caused solely by the movement of ego itself.

Methods	3-way EPE				Norm. EPE	
	Mean ↓	FD↓	FS↓	BS↓	Dyn↓	Stat↓
Ego Motion	0.181	0.534	0.010	0.000	1.000	0.007
SSF [37]	0.028	0.058	0.018	0.009	0.267	0.013
+ScaLR [63]	0.026	0.056	0.016	0.007	0.234	0.012
+SF [89]	0.025	0.053	0.015	0.007	0.198	0.012
+CD [24]	0.024	0.049	0.016	0.008	0.210	**0.011**
+**HilDA†**	0.024	0.051	0.015	0.007	0.181	0.011
+**HilDA**	**0.021**	**0.044**	**0.014**	**0.006**	**0.146**	**0.011**

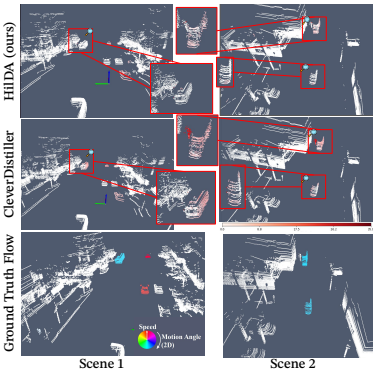

Fig. 6: Qualitative scene flow comparison.

Scene Flow. We integrate MinkUnet34 into SSF [37] and evaluate scene flow on Argoverse 2 [82], jointly testing task transfer and domain generalization. Replacing the randomly initialized backbone with HilDA pre-trained weights improves all metrics in Tab. 6, especially dynamic object estimation. Figure 6 shows lower speed/direction errors (indicated by fewer red regions) and cleaner motion maps compared to baseline, indicating that HilDA features better capture temporal structure and enable better motion estimation where baselines fail.

4.4 Ablations

We ablate the different components and design choices of our method, using 3D semantic segmentation as the evaluation task. See the supplementary material in Sec. A for an extensive set of additional ablations.

Component Analysis. In Tab. 7, ablation of our proposed components are presented. Adding diffusion (a) or hierarchical distillation (c) both significantly

Table 7: Ablation of pre-training loss components. ScaLR [63] with MLP projection head is (base).

#	Diff	Distill	CLS	nuScenes LP	nuScenes 1%	SK 1%	Waymo 1%
(base)	✗	✗	✗	46.36	55.01	50.15	49.08
(a)	✓	✗	✗	50.43	56.97	49.59	49.84
(b)	✗	✓	✗	53.77	57.03	50.63	50.71
(c)	✗	✓	✓	55.13	57.81	50.80	50.87
(d)	✓	✓	✗	55.53	59.04	52.41	51.89
(e)	✓	✓	✓	**56.29**	**59.46**	**52.88**	**52.16**

improves the baseline, while applying both components (e) yields the highest overall performance, demonstrating the synergy between hierarchical distillation and temporal occupancy diffusion.

Multi-Layer Distillation Design. We ablate how to transfer the VFM hierarchy for calibrated point–pixel supervision. While multi-layer distillation has been studied in 2D [12], camera-to-LiDAR distillation introduces an additional geometric point–pixel correspondence constraint. Earlier MinkUnet34 student layers aggregate coarser, irregular 3D neighborhoods whose sparse features contain voxels that may project to different semantic regions. Additionally, overly early VFM teacher layers may not match the abstraction level of late LiDAR features. HilDA therefore uses what we term *Separate* late-layer matching, aligning neighboring blocks (*e.g.*, $\mathbf{Q}_L \to \mathbf{F}_L$ and $\mathbf{Q}_{L-1} \to \mathbf{F}_{L-1}$).

Table 8 shows that this design outperforms both final- and penultimate-layer distillation. Extending it to earlier layers reduces performance, but still surpasses single-layer. This supports the view that intermediate teacher features are useful when their abstraction level and geometric support remain compatible with the student features. The Separate strategy also outperforms *KR-style* aggregation [12], *Aggregate* matching, and *non-uniform* matching. In KR-style aggregation, sliding windows of teacher layers supervise successive student layers, $\{\mathbf{Q}_{L-k}\}_{k=j}^{K-1} \to \mathbf{F}_{L-j}$ with $j = 0, 1$ for a K-layer aggregation. Aggregate matching instead concatenates the selected teacher-layer features into a single target, $\{\mathbf{Q}_{L-k}\}_{k=0}^{K-1} \to \mathbf{F}_L$, whereas non-uniform matching skips intermediate layer pairs, *e.g.*, $(\mathbf{Q}_L \to \mathbf{F}_L, \mathbf{Q}_{L-2} \to \mathbf{F}_{L-2})$. Thus, the gain comes from preserving compatible late-layer correspondences rather than simply adding more teacher features. Table 9 further shows that cosine distance is the strongest point–pixel alignment loss. Since ViT teacher and MinkUnet34 student features have different architectures and feature-norm distributions, cosine supervises semantic direction rather than cross-architecture scale matching. For a more extensive analysis, see Sec. A.4.

Global Context Distillation Design. Table 9 shows that max pooling is the strongest aggregation strategy for constructing $\texttt{CLS}_{\mathbf{F}_L}$, outperforming learnable pooling and per-view/frustum student CLS tokens. This suggests that scene-level alignment benefits from surfacing the most salient responses, and that *global* CLS distillation best complements the local multi-layer distillation, making the

Table 8: Ablation of multi-layer distillation designs. All rows use only $\mathcal{L}_{\text{distill}}$ during pre-training. HilDA corresponds to "Separate, Last 2".

Design	nuSc LP
Final layer only	46.4
Penultimate only	49.9
Separate, Last 2	**53.8**
Separate, Last 3	52.0
Separate, Last 4	51.6
Aggregate, Last 3	53.0
3-layer KR-style [12]	51.4
Non-uniform matching	52.4

Table 9: Ablation of several design choices. Each block changes one component while keeping the HilDA setups (cosine, max pooling, diffusion) fixed.

Comp.	Variant	nuSc LP
$\mathcal{L}_{\text{distill}}$	ℓ_2	53.8
	Kullback-Leibler	52.1
	Cosine distance	**56.3**
CLS pool.	Learnable pool.	55.9
	Per-img. $\texttt{CLS}_{\mathbf{F}_L}$	55.5
	Max pooling	**56.3**
Occ. dec.	Simple decoder	49.1
	ALSO	54.7
	Diffusion	**56.3**

hierarchical distillation design of HilDA the most effective configuration. A more extensive analysis is found in Sec. A.4.

Diffusion* vs. *Occupancy Decoders. To isolate temporal occupancy diffusion, we replace it with deterministic occupancy decoders: a simple decoder [48] and ALSO [7]. Table 9 shows that diffusion performs best, supporting our use of a joint generative scene objective rather than independent per-voxel occupancy prediction which may miss global structural coherence. Through iterative denoising, diffusion provides a label-free spatiotemporal pre-training signal that better captures complex 3D-scene structure and multi-modal evolution through coarse-to-fine refinement. As shown in Secs. 4.2 and 4.3, this auxiliary yields gains across all evaluated tasks, complementing hierarchical distillation.

5 Conclusion

We introduced HilDA, a self-supervised LiDAR pre-training framework that leverages the hierarchical information in VFMs to address limitations of prior camera-to-LiDAR distillation. With hierarchical distillation, combining multi-layer and global context distillation, HilDA captures how semantic features evolve across VFM layers, yielding stronger representations. HilDA also uses temporal occupancy diffusion as an auxiliary task, encouraging features to encode geometric structure and temporal dynamics. The result is an informative 3D representation that captures both the semantic *what* and the geometric *where*. Across segmentation, detection, semantic occupancy, robustness, and scene flow, HilDA improves over prior camera–LiDAR distillation methods. A limitation of point–pixel distillation is sensitivity to LiDAR–camera misalignment. HilDA however partly mitigates this with alignment-free global context distillation and temporal diffusion. Future work could explore how to optimize layer-to-layer distillation through learned matching strategies and dynamic loss weighting, and extend diffusion beyond BEV (*e.g.*, via latent diffusion [66]) to better preserve 3D geometry and temporal evolution.

Acknowledgements

We are grateful to Mohammad Nazari, Ajinkya Khoche, Qingwen Zhang, and John Folkesson for insightful discussions and helpful feedback on the proposed method. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. This work has been performed with support from Sweden's Innovation Agency (VINNOVA) through the Strategic Vehicle Research and Innovation Programme (FFI), grant no. 2024-03640, and from TRATON AB. The computations and data handling were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

1. Agro, B., Sykora, Q., Casas, S., Gilles, T., Urtasun, R.: UnO: Unsupervised occupancy fields for perception and forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
2. Ahn, S., Lee, D., Park, J.: Adaptability of vision foundation models for 3d medical image segmentation. IEEE Open Journal of Signal Processing (2026). https://doi.org/10.1109/OJSP.2025.3650437
3. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denoising diffusion models in discrete state-spaces. In: NeurIPS. vol. 34, pp. 17981–17993 (2021)
4. 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: Proceedings of the International Conference on Computer Vision (2019)
5. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
6. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Bangalath, H., Wang, J., Monteiro, M., Xu, H., Dong, S., Ravi, N., Li, S.W., Dollár, P., Feichtenhofer, C.: Perception encoder: The best visual embeddings are not at the output of the network. In: Advances in Neural Information Processing Systems (2025)
7. Boulch, A., Sautier, C., Michele, B., Puy, G., Marlet, R.: ALSO: Automotive LiDAR self-supervision by occupancy estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision. pp. 9650–9660 (2021)
10. Chan, P.H., Li, B., Baris, G., Sadiq, Q., Donzella, V.: The inconvenient truth of ground truth errors in automotive datasets and DNN-based detection. Data-Centric Engineering p. e34 (2024)

11. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: Challenges and frontiers. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024). https://doi.org/10.1109/TPAMI.2024.3435937
12. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5008–5017 (2021)
13. Chen, X., Liu, Z., Xie, S., He, K.: Deconstructing denoising diffusion models for self-supervised learning. In: Proceedings of the International Conference on Learning Representations (ICLR) (2025)
14. Chodosh, N., Ramanan, D., Lucey, S.: Re-evaluating lidar scene flow. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 6005–6015 (January 2024)
15. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
16. Contributors, P.: Pointcept: A codebase for point cloud perception research. GitHub repository (2023)
17. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. Proceedings of the International Conference on Learning Representations (2024)
18. Diehl, C., Sykora, Q., Agro, B., Gilles, T., Casas, S., Urtasun, R.: Dio: Decomposable implicit 4d occupancy-flow world model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27456–27466 (2025)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (2021)
20. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3d awareness of visual foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21795–21806 (2024)
21. Fang, J., Zhou, D., Yan, F., Zhao, T., Zhang, F., Ma, Y., Wang, L., Yang, R.: Augmented lidar simulator for autonomous driving. IEEE Robotics and Automation Letters **5**(2), 1931–1938 (2020)
22. Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., Bai, X.: Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In: Proceedings of the International Conference on Computer Vision. pp. 24823–24834. IEEE (2025)
23. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research (2013). https://doi.org/10.1177/0278364913491297
24. Govindarajan, H., Wozniak, M., Klingner, M., Maurice, C., Kiran, B.R., Yogamani, S.: Cleverdistiller: Simple and spatially consistent cross-modal distillation. In: 36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025. BMVA (2025)
25. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: NeurIPS (2020)
26. Gu, S., Yin, W., Jin, B., Guo, X., Wang, J., Li, H., Zhang, Q., Long, X.: Dome: Taming diffusion model into high-fidelity controllable occupancy world model. arXiv preprint arXiv:2410.10429 (2024)

27. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., Chen, Y.C.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In: The Thirteenth International Conference on Learning Representations (2025)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016)
29. Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: Radiov2. 5: Improved baselines for agglomerative vision foundation models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22487–22497 (2025)
30. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33 (2020)
31. Hudson, D.A., Zoran, D., Malinowski, M., Lampinen, A.K., Jaegle, A., McClelland, J.L., Matthey, L., Hill, F., Lerchner, A.: Soda: Bottleneck diffusion models for representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23115–23127 (2024)
32. Jiang, P., Osteen, P., Wigness, M., Saripalli, S.: RELLIS-3D dataset: Data, benchmarks and analysis. In: 2021 IEEE International Conference on Robotics and Automation (ICRA) (2021). https://doi.org/10.1109/ICRA48506.2021.9561251
33. Justo Miro, A., af Klinteberg, L., Timus, B., Asefaw, A., Khoche, A., Gustafsson, T., Mansouri, S.S., Daneshtalab, M.: Correcting and Quantifying Systematic Errors in 3D Box Annotations for Autonomous Driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6724–6732 (March 2026)
34. Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: AnyLoc: Towards universal visual place recognition. IEEE Robotics and Automation Letters **9**(2), 1286–1293 (2024). https://doi.org/10.1109/LRA.2023.3343602
35. Khatri, I., Vedder, K., Peri, N., Ramanan, D., Hays, J.: I Can't Believe It's Not Scene Flow! In: Proceedings of the European Conference on Computer Vision (2024). https://doi.org/10.1007/978-3-031-72649-1_14
36. Khoche, A., Asefaw, A., González, A., Timus, B., Mansouri, S.S., Jensfelt, P.: Addressing data annotation challenges in multiple sensors: A solution for scania collected datasets. In: 2024 European Control Conference (ECC). pp. 1032–1038 (2024)
37. Khoche, A., Zhang, Q., Sánchez, L.P., Asefaw, A., Mansouri, S.S., Jensfelt, P.: SSF: Sparse long-range scene flow for autonomous driving. In: Proceedings of the IEEE International Conference on Robotics and Automation (2025). https://doi.org/10.1109/ICRA55743.2025.11128770
38. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
39. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the International Conference on Computer Vision (2023)
40. Klokov, A.A., Pak, D.U., Khorin, A., Yudin, D.A., Kochiev, L., Luchinskiy, V.D., Bezuglyj, V.D.: DAPS3D: Domain adaptive projective segmentation of 3D LiDAR point clouds. IEEE Access **11**, 79341–79356 (2023)
41. 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: Proceedings of the International Conference on Computer Vision (2023)

42. KTH-RPL, contributors: Opensceneflow: A codebase for point cloud scene flow estimation research. GitHub repository (2026), accessed 2026-01-09
43. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., Zhou, S., Zhang, L., Qi, X., Zhao, H., Yang, M., Zeng, W., Jin, X.: Uniscene: Unified occupancy-centric driving scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11971–11981 (June 2025)
44. Li, J., Saltori, C., Poiesi, F., Sebe, N.: Cross-modal and uncertainty-aware agglomeration for open-vocabulary 3d scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19390–19400 (June 2025)
45. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., Wang, X.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12037–12047. IEEE (2025)
46. Liebel, L., Körner, M.: Auxiliary tasks in multi-task learning. arXiv preprint arXiv:1805.06334 (2018)
47. Liu, J., Wang, G., Ye, W., Jiang, C., Han, J., Liu, Z., Zhang, G., Du, D., Wang, H.: Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15109–15119 (2024)
48. Liu, X., Gong, M., Fang, Q., Xie, H., Li, Y., Zhao, H., Feng, C.: LiDAR-based 4D occupancy completion and forecasting. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2024). https://doi.org/10.1109/IROS58592.2024.10801302
49. 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)
50. 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)
51. Liu, Y., Fu, J., Wu, Y., Wu, K., Li, P., Wu, J., Zhou, S., Xin, J.: Mind the gap: Aligning vision foundation models to image feature matching. In: Proceedings of the International Conference on Computer Vision (2025)
52. Ljungbergh, W., Lilja, A., Tonderski, A., Ling, A.L., Lindström, C., Verbeke, W., Fu, J., Petersson, C., Hammarstrand, L., Felsberg, M.: GASP: Unifying geometric and semantic self-supervised pre-training for autonomous driving. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (2026)
53. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
54. Ma, J., Chen, X., Huang, J., Xu, J., Luo, Z., Xu, J., Gu, W., Ai, R., Wang, H.: Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21486–21495 (2024)
55. Meng, Q., Wang, W., Zhou, T., Shen, J., Jia, Y., Van Gool, L.: 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 (2022). https://doi.org/10.1109/TPAMI.2021.3063611
56. MMDetection3D Contributors: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection (2020)

57. Murrugarra-Llerena, J., Kirsten, L., Jung, C.R.: Can we trust bounding box annotations for object detection? In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 4812–4821 (2022)
58. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8162–8171. PMLR (2021)
59. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. In: Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks (December 2021)
60. Nunes, L., Marcuzzi, R., Mersch, B., Behley, J., Stachniss, C.: Scaling diffusion models to real-world 3d lidar scene completion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14770–14780. IEEE (2024)
61. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., 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. Transactions on Machine Learning Research (2024)
62. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022). https://doi.org/10.1109/CVPR52688.2022.01036
63. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
64. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
65. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: Am-radio: Agglomerative vision foundation model reduce all domains into one. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12490–12500 (2024)
66. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
67. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, Cham (2015)
68. Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R.: Image-to-LiDAR self-supervised distillation for autonomous driving data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
69. Shi, S., Wang, X., Li, H.: PointRCNN: 3D object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
70. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
71. Tian, H., Xu, B., Li, S.: Distillation dynamics: Towards understanding feature-based distillation in vision transformers. Proceedings of the AAAI Conference on Artificial Intelligence (2026). https://doi.org/10.1609/aaai.v40i11.37913
72. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
73. Unal, O., Dai, D., Van Gool, L.: Scribble-supervised LiDAR semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2697–2707 (2022)
74. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In: NeurIPS (2022)
75. Wang, G., Wang, Z., Tang, P., Zheng, J., Ren, X., Feng, B., Ma, C.: Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. In: Proceedings of the European Conference on Computer Vision. pp. 95–112. Springer (2024)
76. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: Computer Vision – ECCV 2024 (2025). https://doi.org/10.1007/978-3-031-73195-2_4
77. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In: Proceedings of the International Conference on Computer Vision. pp. 17850–17859 (October 2023)
78. Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., Chen, Y., Diamond, J., Ding, Y., Ding, W., et al.: Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. arXiv preprint arXiv:2511.00088 (2025)
79. Wang, Y., Liu, Y., Yuan, T., Mao, Y., Liang, Y., Yang, X., Zhang, H., Zhao, H.: Diffusion-based generative models for 3D occupancy prediction in autonomous driving. In: 2025 IEEE International Conference on Robotics and Automation (ICRA) (2025). https://doi.org/10.1109/ICRA55743.2025.11128716
80. Wei, C., Mangalam, K., Huang, P.Y., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A.L., Feichtenhofer, C.: Diffusion models as masked autoencoders. In: Proceedings of the International Conference on Computer Vision. pp. 16238–16248 (2023)
81. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21729–21740 (2023)
82. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: NeurIPS (2021)
83. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point Transformer V3: Simpler, faster, stronger. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
84. Xiang, W., Yang, H., Huang, D., Wang, Y.: Denoising diffusion autoencoders are unified self-supervised learners. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)

85. Xiao, A., Huang, J., Xuan, W., Ren, R., Liu, K., Guan, D., El Saddik, A., Lu, S., Xing, E.P.: 3D semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
86. Xu, X., Kong, L., Shuai, H., Pan, L., Liu, Z., Liu, Q.: Limoe: Mixture of lidar representation learners from automotive scenes. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27368–27379 (2025)
87. Xu, X., Kong, L., Shuai, H., Pan, L., Liu, Z., Liu, Q.: LiMoE: Mixture of LiDAR representation learners from automotive scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
88. Xu, X., Kong, L., Shuai, H., Zhang, W., Pan, L., Chen, K., Liu, Z., Liu, Q.: 4D contrastive superflows are dense 3D representation learners. In: Proceedings of the European Conference on Computer Vision (2024)
89. Xu, X., Kong, L., Shuai, H., Zhang, W., Pan, L., Chen, K., Liu, Z., Liu, Q.: Enhanced spatiotemporal consistency for image-to-LiDAR data pretraining. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026). https://doi.org/10.1109/TPAMI.2025.3640589
90. Xu, X., Kong, L., Wang, S., Zhou, C., Liu, Q.: Beyond one shot, beyond one perspective: Cross-view and long-horizon distillation for better lidar representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25506–25518 (2025)
91. Yang, X., Wang, X.: Diffusion model as representation learner. In: Proceedings of the International Conference on Computer Vision (2023)
92. Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., Yuan, C.: Focal and global knowledge distillation for detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4643–4652 (2022)
93. Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C., Li, Y.: ViTKD: Feature-based knowledge distillation for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1379–1388 (2024)
94. Ye, X., Yaman, B., Cheng, S., Tao, F., Mallik, A., Ren, L.: Bevdiffuser: Plug-and-play diffusion model for bev denoising with ground-truth guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1495–1504 (2025)
95. Yoon, H., Jung, J., Kim, J., Choi, H., Shin, H., Lim, S., An, H., Kim, C., Han, J., Kim, D., Eom, C., Hong, S., Kim, S.: Visual representation alignment for multimodal large language models. In: ICLR 2026 Workshop on Multimodal Intelligence (2026)
96. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: Proceedings of the International Conference on Learning Representations (2025)
97. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: Proceedings of the International Conference on Computer Vision (2023)
98. Zhang, J., Herrmann, C., Hur, J., Chen, E., Jampani, V., Sun, D., Yang, M.H.: Telling left from right: Identifying geometry-aware semantic correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

99. Zhang, J., Herrmann, C., Hur, J., Polania Cabrera, L., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. Advances in Neural Information Processing Systems **36**, 45533–45547 (2023)
100. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., Cao, X., Yin, W.: Epona: Autoregressive diffusion world model for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
101. Zhang, Q., Khoche, A., Yang, Y., Ling, L., Mansouri, S.S., Andersson, O., Jensfelt, P.: HiMo: High-speed objects motion compensation in point clouds. IEEE Transactions on Robotics (2025). https://doi.org/10.1109/TRO.2025.3619042
102. Zhang, Q., Yang, Y., Fang, H., Geng, R., Jensfelt, P.: DeFlow: Decoder of scene flow network in autonomous driving. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 2105–2111 (2024). https://doi.org/10.1109/ICRA57147.2024.10610278
103. Zhang, Y., Hou, J.: Fine-grained image-to-LiDAR contrastive distillation with visual foundation models. Advances in Neural Information Processing Systems (2024)
104. Zhang, Y., Hou, J.: Is contrastive distillation enough for learning comprehensive 3D representations? International Journal of Computer Vision (2026). https://doi.org/10.1007/s11263-026-02879-z
105. Zhang, Y., Gong, S., Xiong, K., Ye, X., Li, X., Tan, X., Wang, F., Huang, J., Wu, H., Wang, H.: BEVWorld: A multimodal world simulator for autonomous driving via scene-level BEV latents. arXiv preprint arXiv:2407.05679 (2024)
106. Zheng, S., Bao, Z., Zhao, R., Hebert, M., Wang, Y.X.: Diff-2-in-1: Bridging generation and dense perception with diffusion models. In: Proceedings of the International Conference on Learning Representations (2025)
107. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT pre-training with online tokenizer. In: Proceedings of the International Conference on Learning Representations (2022)