LiDAR-Anchored Splat-and-Distill: Sensor-Grounded 3D-Aware Distillation for Vision Foundation Encoders

Anonymous authors
Paper under double-blind review

Abstract

Vision foundation encoders capture strong 2D semantics but lack the explicit 3D geometry essential for autonomous vehicle (AV) applications and embodied perception. While recent feed-forward "splat-and-distill" approaches lift camera pairs into Gaussian feature fields to distill 3D structure, they rely on inferred, scale-ambiguous geometry from frozen stereo networks, built for short-baseline, high-overlap scenes. In contrast, standard AVs are equipped with synchronized camera-and-LiDAR rigs that provide metric geometry directly. We therefore take a geometry-first approach and introduce LiDAR-Anchored Splat-and-Distill, a framework leveraging this native sensor setup to distill 3D-aware features from measured rather than inferred geometry. Adapting feed-forward distillation to AV environments raises two challenges. First, the views of a surround rig differ substantially in viewpoint, so geometry cannot be inferred from them; anchoring 3D Gaussian feature fields directly to LiDAR surfels removes any need for geometry estimation and lets distillation operate across arbitrary viewpoint changes. Second, LiDAR is sparse, confined to a narrow horizontal band, and subject to occlusion, so the lifted-and-rendered features leave most of every target view without a distillation target; a coverage-masked semantic-anchor loss preserves feature quality in those regions. An auxiliary metric-depth head supervised by the same LiDAR complements the distillation objective. We train on in-house multi-camera-and-LiDAR driving logs and evaluate zero-shot on KITTI, which uses a different sensor rig, across two backbones, DINOv2-Base and C-RADIOv4-Huge. Our method improves depth estimation and cross-view correspondence over the off-the-shelf encoders while also improving 2D semantic segmentation. Against the public SnD, FiT3D, and MEF checkpoints under the same protocol, it gives the largest depth gain and is the strongest on KITTI-pairs, a rotation-stratified driving correspondence benchmark we introduce. Our resulting encoder requires no LiDAR and adds no computational overhead at test time: measured geometry serves only as training-time scaffolding for a general-purpose 2D encoder.

1 Introduction

Vision foundation encoders, whether trained with self-supervised objectives (the DINO lineage; Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025), segmentation teachers (SAM; Kirillov et al., 2023), language supervision (CLIP and Perception Encoder; Radford et al., 2021; Bolya et al., 2025), or agglomerative distillation of several such teachers (RADIO; Ranzinger et al., 2024; Heinrich et al., 2025), capture strong 2D semantics but lack explicit 3D geometry. Probing studies show that their features encode a surprising amount of 3D structure (El Banani et al., 2024; Man et al., 2024; Chen et al., 2025), yet they still fall short of purpose-built geometric models (Yang et al., 2024b; Piccinelli et al., 2024; Wang et al., 2025) in metric depth and multi-view consistency under large viewpoint change. The gap follows from how they are trained: on single 2D images with no cross-view consistency constraint.

These encoders are meant to be general purpose, serving many downstream tasks from frozen features (Oquab et al., 2023; Siméoni et al., 2025; Bolya et al., 2025). In autonomous driving those tasks include depth, occupancy, and bird's-eye-view estimation alongside detection and segmentation (Philion & Fidler, 2020; Li et al., 2022; Hu et al., 2023), so a general encoder must deliver 3D quality without sacrificing 2D.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

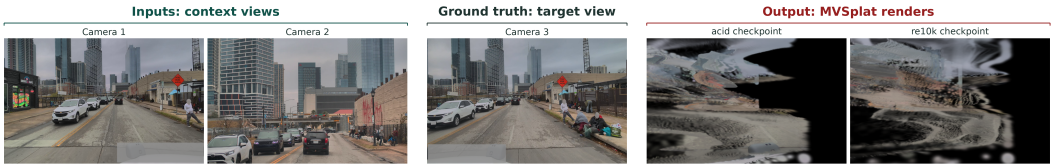

Figure 1: **Feed-forward stereo geometry fails on a surround-view rig.** To create the distillation target, SnD (Shavin & Benaim, 2026) uses frozen MVSplat (Chen et al., 2024) to infer 3D Gaussians from an image pair and render a held-out view. Given two context views from our surround-view rig to render a held-out target view, neither the ACID-trained nor the RE10K-trained public checkpoint produces a recognizable image: stereo matching degrades under the large viewpoint differences of a surround rig, so the recipe has no usable geometry here.

Recent work closes part of this gap by distilling 3D awareness into a 2D encoder. FiT3D (Yue et al., 2024) fits a per-scene 3D representation and supervises the encoder with features rendered from it, at the cost of a slow per-scene optimization loop; MEF (You et al., 2025) avoids rendering altogether by enforcing feature agreement at known 3D correspondences. Splat-and-Distill (SnD) (Shavin & Benaim, 2026) removes the per-scene cost with a feed-forward recipe: a frozen stereo network (MVSplat; Chen et al., 2024) predicts 3D Gaussian geometry from a close two-view image pair, teacher features from both context views are fused onto that inferred geometry, and the fused field is rendered into a third, held-out view as the distillation target. The recipe rests on inferred, scale-ambiguous geometry from a network built for static, high-overlap scenes. This has two consequences we document: the recipe cannot be run on a surround-view rig, where the stereo network fails outright (Figure 1), and even SnD's released encoder, evaluated on KITTI, falls below its off-the-shelf initialization once the viewpoint change between views is large (Section 4.3).

In contrast, standard AVs carry synchronized surround-camera and LiDAR rigs with accurate ego-motion (Geiger et al., 2012; Caesar et al., 2020; Sun et al., 2020), so scene geometry can be measured rather than inferred, and metric rather than scale-ambiguous. We introduce **LiDAR-Anchored Splat-and-Distill**, a framework that builds the feature field on this measured geometry and distills 3D-aware features without any geometry estimation. Adapting feed-forward distillation to AV environments introduces two challenges the indoor setting of prior work never posed:

- **Large viewpoint difference.** A surround-view AV rig is built for coverage, not stereo: adjacent cameras share little field of view and differ substantially in orientation and position. Feed-forward stereo networks such as MVSplat, built for short-baseline pairs, produce degenerate renders on every pair from our rig we tested (Figure 1) and encoders distilled from such pairs do not improve under large viewpoint change (Section 4.3). Because our pipeline lifts features onto measured geometry rather than matching features across views, the viewpoint difference between context and target views is not a limiting factor.
- **LiDAR sparsity and occlusion.** Our distillation target is produced by lifting context features onto LiDAR surfels and rendering them into the target view, so a target patch receives signal only where a LiDAR return exists and is seen by a context camera. LiDAR returns are sparse, decay with range, and are confined to a narrow horizontal band. Occlusion, back-face rejection, and limited overlap between adjacent cameras further reduce coverage, leaving roughly three quarters of every target view without a distillation target (Appendix A.5). Left unaddressed, this supervision gap degrades the encoder's feature quality. We regularize these unobserved regions with a coverage-masked semantic-anchor loss.

We further add an auxiliary metric-depth head that complements the distillation objective. All of this is training-time scaffolding. Adaptation uses bounded-rank LoRA (Hu et al., 2022) merged at export, so the trained encoder is architecturally identical to its off-the-shelf counterpart and uses no LiDAR at inference. Our core contributions are:

1. **A sensor-anchored, metric generalization of feed-forward splat-and-distill** that replaces estimated stereo geometry with measured LiDAR geometry, extending the recipe to multi-camera outdoor rigs with large viewpoint differences between views. Across two backbones it improves depth and correspondence while also improving a linear segmentation probe.

2. **A coverage-driven diagnosis and fix for representation drift** in sparse feed-forward splat-and-distill. We quantify how little of a target view fused context features actually cover (Section 3.5) and introduce a masked semantic-anchor loss to resolve the resulting performance regression. The principle applies to any sparse splat-and-distill setup beyond autonomous driving.
3. **Direct depth supervision as a complement to distillation.** Depth supervision by itself yields accurate depth but degrades cross-view correspondence below the off-the-shelf encoder on both backbones. When combined with distillation, the auxiliary depth head enhances both depth and correspondence, showing that the two signals reinforce each other rather than substituting.
4. **Evidence that indoor and object-centric 3D distillation does not transfer to driving.** Under an identical protocol on KITTI-pairs, our rotation-stratified driving correspondence benchmark, the public SnD, FiT3D, and MEF checkpoints all fall below the off-the-shelf encoder at 10 px and lose 6 to 10 points under large rotation, whereas ours holds (Section 4.2). Ours is the only checkpoint that improves correspondence at both evaluated thresholds, and it also gives the largest depth gain. SnD remains stronger on semantics, so the two recipes are complementary.

2 Related Work

Vision foundation models. Self-supervised encoders (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025), segmentation-supervised SAM (Kirillov et al., 2023), and agglomerative distillation models (Ranzinger et al., 2024; Heinrich et al., 2025) learn strong semantic representations from single-view 2D images. Probing studies find real but limited 3D structure in their features: strong single-view depth and normals, weaker multi-view consistency, and no metric scale (El Banani et al., 2024; Man et al., 2024; Chen et al., 2025). None of these imposes a multi-view consistency constraint, which is the gap we address.

3D-aware 2D encoders. Methods that instill 3D awareness into a pretrained 2D encoder differ chiefly in where their geometry comes from. FiT3D (Yue et al., 2024) optimizes a Gaussian feature field per scene (Zhou et al., 2024) and supervises the encoder with features rendered from it, and SPA (Zhu et al., 2025) pretrains with per-scene rendering losses. MEF (You et al., 2025) needs no field and enforces feature agreement at known correspondences. DUNE (Sariyildiz et al., 2025) distills from 2D and 3D teachers with no reconstruction at all. Splat-and-Distill (Shavin & Benaim, 2026) replaces per-scene optimization with a single forward pass: a frozen MVSplat (Chen et al., 2024) predicts Gaussians from an image pair and the distillation target is rendered into a held-out view. In every rendering-based case the geometry is estimated from images, by per-scene optimization or by a feed-forward network (Charatan et al., 2024; Chen et al., 2024; Wang et al., 2024; 2025) trained on short-baseline, high-overlap scenes. We keep SnD's feed-forward, novel-view-target structure and replace the estimated geometry with geometry a LiDAR-and-camera rig measures directly, at metric scale, and without requiring overlap between context views (Figure 1).

LiDAR as a training signal for cameras. Camera-based driving work (Philion & Fidler, 2020; Li et al., 2022; Hu et al., 2023; Zhou et al., 2017; Godard et al., 2019) mostly uses LiDAR only for evaluation, while driving reconstructions (Yang et al., 2024a; Yan et al., 2024) use it to fit a per-scene model. Closest to us, cross-modal distillation (Li et al., 2023; Chen et al., 2023; Zhou et al., 2023) supervises a camera detector with LiDAR during training and discards it at inference, but the student is a task-specific BEV detector. Ours differs in both student and mechanism: the student is a general-purpose 2D encoder rather than a task-specific detector, and LiDAR is a geometric scaffold for lifting and rendering context-view features into a held-out view, rather than a labeling source.

3 Method

3.1 Overview

We introduce a **geometry-first** reformulation of feed-forward splat-and-distill learning (Figure 2). During training, LiDAR returns are lifted into a surfel field, populated with a frozen teacher's 2D features across context views, and rendered into a target view as a distillation signal for the student. At inference time, the entire geometric scaffolding is removed: the pipeline collapses to a single forward pass through the student network, producing a deployed model that serves as a seamless replacement for the original vision encoder.

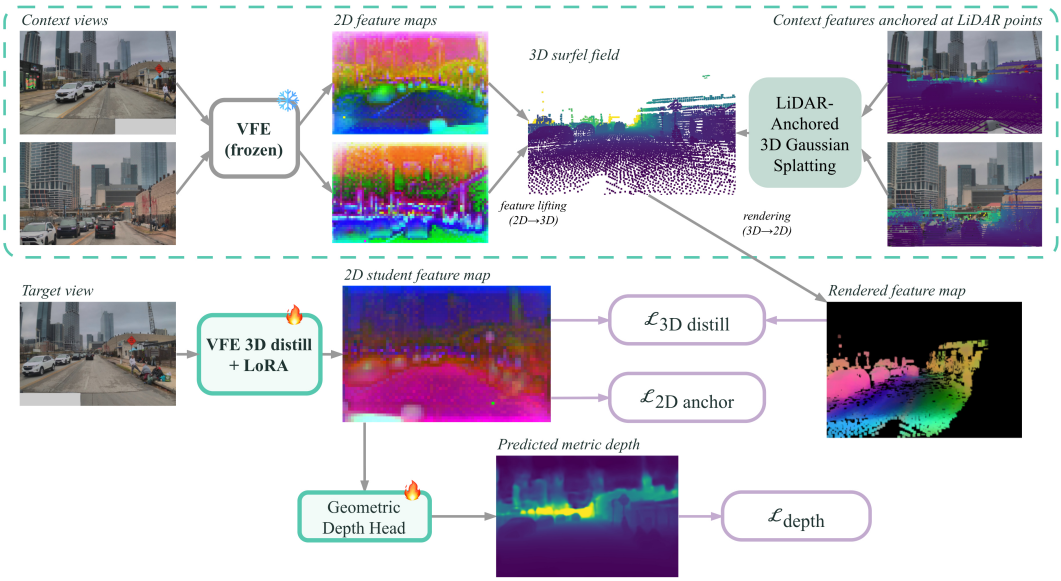

Figure 2: **Training-time pipeline.** LiDAR and calibration are lifted into a Gaussian surfel field (Section 3.2); frozen-teacher features from the context views are fused onto it and splatted into the held-out target view together with a render-confidence map (Section 3.3). The student is supervised by the splatted target where confidence is nonzero, by a frozen anchor elsewhere (Section 3.5), and by a LiDAR-supervised depth head (Section 3.4). VFE stands for Vision Foundation Encoder.

3.2 LiDAR-Anchored Geometry Lifting

Given one scene's synchronized camera images, LiDAR points in a common ego-vehicle frame, and per-camera intrinsics/extrinsics, we build a **Gaussian surfel field** (Dai et al., 2024): a set of thin, oriented disks rather than the volumetric blobs of radiance splatting, since each primitive here sits on a LiDAR-measured point and represents a local surface patch.

Range binning and adaptive voxel downsampling. LiDAR point density falls off sharply with range. We bin points by distance to the ego origin and voxel-downsample each bin using a voxel that grows with range, with a cap on the total count. Each centroid becomes one surfel. Because a surfel's image footprint shrinks with range, this keeps near and far surfels at a similar footprint without discarding distant structure.

Local-surface fit. For each centroid we run principal-component analysis on its nearest raw LiDAR points. The direction of minimal variance defines the surfel normal, oriented toward the ego origin. The remaining two axes span the disk, with radii equal to the standard deviations along them and a thickness fixed to a small fraction of the smaller radius, so each surfel is a flat disk that models a local surface patch.

3.3 Multi-View Fusion and Splatting

Context and target views. Each training sample is one timestamp of a scene and one combination of context cameras and a separate target camera, drawn uniformly from a fixed set. Because surfel geometry comes from LiDAR rather than from matching between cameras, context views do not need to overlap: each context camera paints whichever surfels it observes. The only requirement is that at least one context camera sees the surfels visible from the target view. Since a target patch is supervised only where a LiDAR surfel exists and is painted by some context camera, LiDAR sparsity, together with the limited field of view shared by adjacent cameras, means only a portion of each target view receives distillation coverage (Appendix A.5).

Feature fusion onto the field. A frozen teacher extracts a patch-grid feature map for each context view. Each surfel is projected into every context view and kept only if it sits within the camera frustum, lies within a small depth margin of the front-most surfel at that pixel, and faces toward the camera based on its surface normal. Each valid projection bilinearly samples the context feature with weighting that favors face-on viewing, short range, and distance from the image border. These

weighted features are averaged across observing context views and L2 normalized to yield a single fused feature per surfel.

Splatting to the target view. The fused field is rendered into the target view with an elliptical-weighted-average Gaussian splatter: each surfel's 3D covariance is projected through the pinhole Jacobian into a 2D footprint, a depth pre-pass resolves occlusion, and surviving surfels are scatter-accumulated per pixel into a fused target feature map and a per-pixel render-confidence map, which is the total accumulated splatting weight at a pixel and exactly zero wherever no surfel's footprint reaches. Both maps are rendered at image resolution and bilinearly resized to the student's patch grid.

3.4 Auxiliary Depth Head

We add a lightweight DPT decoder (Ranftl et al., 2021) on the student's final-layer features, trained to predict metric depth against the LiDAR's reprojected, occlusion-resolved depth as ground truth. This gives the student a direct geometric signal on top of whatever the fused-feature target implicitly encodes, and allows us to evaluate (Section 4.4) whether feature distillation alone is enough or whether direct metric supervision provides complementary value.

3.5 The Coverage Problem and the Masked Anchor Loss

Unlike a learned stereo network that predicts dense geometry across all input pixels, LiDAR returns points only where it happened to sample, and fusion additionally discards surfels due to occlusion and back-face rejection. As a result, a large majority of any target view's patch grid has zero render-confidence: no context view contributed any feature to those pixels at all. Appendix A.5 reports this fraction and shows it is a structural property of LiDAR coverage and rig geometry, not an artifact of any one scene.

A semantic-preservation loss applied to every pixel would compete with the distillation loss wherever both have a target. We therefore restrict the anchor loss to the zero-coverage region: the frozen anchor holds the student's features in place only where the splat provides no signal, and distillation alone governs the rest. We use a hard mask for simplicity, though soft weighting by render confidence remains a viable alternative. Without this term, distillation degrades the geometric probes below the off-the-shelf encoder, whereas adding the anchor restores feature quality (Section 4.4, Appendix A.1 records a rejected alternative).

3.6 Losses

The total training loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{distill}} + \lambda_a \, \mathcal{L}_{\text{anchor}} + \lambda_d \, \mathcal{L}_{\text{depth}}, \tag{1}$$

Let $S(x)$ denote the student feature at patch x, $T(x)$ the splatted target feature, $c(x)$ the render confidence, and $A(x)$ the frozen anchor feature. We define $\mathcal{M} = \{x : c(x) \leq \epsilon\}$ as the set of zero-coverage patches. The individual loss components are given by:

$$\mathcal{L}_{\text{distill}} = \frac{\sum_x c(x) \left(1 - \cos\langle S(x), T(x)\rangle\right)}{\sum_x c(x)}, \tag{2}$$

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\mathcal{M}|} \sum_{x \in \mathcal{M}} \left(1 - \cos\langle S(x), A(x)\rangle\right), \tag{3}$$

$$\mathcal{L}_{\text{depth}} = 10 \cdot \text{SILog}\big(D(S), \hat{D}\big) + 0.5 \cdot \text{grad}\big(D(S), \hat{D}\big), \tag{4}$$

with $D(S)$ the depth head's prediction, $\hat{D}$ the LiDAR depth reprojected and z-buffered into the target view, SILog the scale-invariant log-depth error, and grad a multi-scale log-depth gradient term, both evaluated only at pixels with a LiDAR return.

3.7 Bounded-Rank Adaptation

We adapt the student with LoRA (Hu et al., 2022) on the attention projections of every transformer block (exact modules per backbone in Appendix A.4) rather than full fine-tuning, and merge the

updates into the base weights at export, so the deployed checkpoint has the same architecture and inference cost as the off-the-shelf backbone. Bounding the rank also preserves pretrained semantics: with full fine-tuning the anchor loss alone is insufficient to prevent feature drift (Section 4.4.3).

4 EXPERIMENTS

4.1 SETUP

We train on a single in-house corpus of synchronized multi-camera and LiDAR driving logs with ego-poses and calibration; implementation and training details are in Appendix A.4. KITTI is used only for evaluation, so every reported number is zero-shot. We start from two public encoders with different pretraining recipes, DINOv2-Base (Oquab et al., 2023) and C-RADIOv4-Huge (Heinrich et al., 2025) and for each report the off-the-shelf checkpoint and the checkpoint after our distillation ("+ ours"), both evaluated frozen under the same protocol (Section 4.2). The comparison is before/after within each backbone. All runs use LoRA rank 8 except in the capacity ablation (Section 4.4.3).

4.2 EVALUATION PROTOCOL

Prior 3D-awareness probing work, Probe3D (El Banani et al., 2024) and FiT3D (Yue et al., 2024), evaluates on object-centric and indoor benchmarks. We evaluate on KITTI (Geiger et al., 2013), which matches our target domain: outdoor driving captured by a calibrated multi-camera and LiDAR rig. Its rig differs from our training rig in camera placement, LiDAR, and geography, and the encoder never sees KITTI during training. We probe the two gaps Section 1 identifies, metric depth and multi-view consistency under large viewpoint change, through monocular depth and cross-view correspondence, and we probe 2D retention through semantic segmentation.

Frozen backbone, linear readout. Every model, including every ablation variant, is evaluated frozen with only a linear head on top, so a probe can succeed only based on what is linearly accessible in the features. This is also how a general-purpose encoder is used: frozen, on a single image, with no LiDAR or multi-view input. Probe heads follow FiT3D (Appendix A.4); correspondence needs no head and uses nearest-neighbor matching.

Geometric correspondence (KITTI-pairs). Existing correspondence probes (NAVI and ScanNet-pairs) are object-centric or indoor, and KITTI's own stereo pair offers a single, fixed 0.54 m viewpoint change, so neither can show how correspondence degrades as viewpoint change grows. We therefore construct **KITTI-pairs**, a wide-baseline correspondence benchmark on KITTI raw sequences that follows the ScanNet-pairs construction: frame pairs are sampled from posed video at gaps of 0.1 to 2 s, with additional pairs drawn to balance relative rotation, and correspondence ground truth is derived from known relative pose and depth, giving 2,863 pairs from 102 drives (Appendix A.6). Features are matched by nearest neighbor and each match is scored by its reprojection error under the known relative pose. We report recall at 5 and 10 pixels, and bin recall@10px by relative rotation and translation in Appendix A.2. Pairs come from one camera at two times, so the probe tests consistency under egomotion, and since training uses only same-timestamp cross-camera views, it also tests generalization.

Depth. We use the Eigen split (Eigen et al., 2014) of KITTI raw's left color camera with the official semi-dense ground truth (Uhrig et al., 2017), the Garg crop (Garg et al., 2016) and a depth range of $[0.001, 80]$ m, leaving 16,690 training and 606 test images with ground truth. We report AbsRel and δ_1 over LiDAR-covered pixels.

Semantic segmentation. KITTI's semantic benchmark provides 200 labeled images in the 19-class Cityscapes taxonomy with no official validation partition, so we hold out the last 20 images in sorted-filename order, train on the remaining 180, using FiT3D's NYUv2-13 probe recipe, and then evaluate on the separate 200-image test set. Because both 3D probes take their ground truth from LiDAR, segmentation is the one probe that reaches the zero-coverage region the masked anchor loss regularizes (Section 3.5). mIoU is the primary metric and we report mAcc and aAcc alongside it.

Public 3D distillation baselines. We evaluated the released DINOv2-Base checkpoints of SnD (Shavin & Benaim, 2026), FiT3D (Yue et al., 2024), and MEF (You et al., 2025) frozen under the identical protocol. SnD releases two checkpoints, recommended for correspondence and

for dense tasks respectively; the former is better on every probe here, so we report it in Table 1 and both in Appendix A.2. All three are trained on indoor or object-centric data with inferred or per-scene-optimized geometry, and none releases a C-RADIOv4 checkpoint, so the comparison is single-backbone and measures what each recipe delivers in the driving domain rather than a controlled comparison of recipes under matched data and compute. Our central claim still rests on the before/after comparison within each backbone.

4.3 Main Results

Table 1: KITTI results, frozen backbone, identical probe protocol. "+ ours" is the off-the-shelf encoder after our distillation; FiT3D, MEF, and SnD are public DINOv2-Base checkpoints (Section 4.2). "10–90°" is recall@10px on the largest-rotation pairs (Appendix A.2). Best per backbone and column in bold.

Backbone	Variant	Correspondence (recall)			Depth		Segmentation		
		@5px↑	@10px↑	@10px, 10–90° ↑	AbsRel↓	δ_1 ↑	mIoU↑	mAcc↑	aAcc↑
DINOv2-Base	Off-the-shelf	60.59	73.07	**47.81**	0.099	0.910	60.37	73.75	92.94
	FiT3D (Yue et al., 2024)	58.14	70.02	40.18	0.134	0.848	53.16	66.67	91.83
	MEF (You et al., 2025)	58.61	69.95	38.19	0.099	0.911	60.12	73.99	91.94
	SnD (Shavin & Benaim, 2026)	61.43	72.58	41.89	0.096	0.915	**67.43**	**75.96**	**93.24**
	+ ours	**61.46**	**73.34**	**47.81**	**0.088**	**0.931**	63.64	73.19	92.62
C-RADIOv4-Huge	Off-the-shelf	62.32	72.45	**45.76**	0.126	0.868	66.53	75.09	**93.16**
	+ ours	**63.00**	**73.09**	45.10	**0.119**	**0.889**	**67.24**	**75.66**	93.00

Before/after. Our distillation improves geometric awareness on both backbones (Table 1). Correspondence recall gains 0.7 to 0.9 points at 5 px and 0.3 to 0.6 at 10 px, and stays within a point of off-the-shelf on the largest-rotation pairs. Depth improves more: AbsRel drops 11% on DINOv2-Base and 6% on C-RADIOv4-Huge, and δ_1 gains 2.1 points on each. The gains are uniform in direction across two encoders of different architecture and pretraining, suggesting they come from the sensor-grounded target rather than a backbone property. mIoU improves by 3.3 and 0.7 points while aAcc drops by 0.3 and 0.2: LiDAR returns concentrate on near-field structure, so distillation sharpens the small classes that mIoU weights equally, while sky and distant background fall under the masked anchor, which preserves rather than improves (Section 3.5).

Against public 3D-distillation checkpoints. On DINOv2-Base, ours is the only checkpoint that improves correspondence at both thresholds. SnD matches it at 5 px but falls below off-the-shelf at 10 px; FiT3D and MEF are 2 to 3 points below at both. The binned results (Appendix A.2) locate the loss: SnD gains below 10° of rotation and loses 5.9 points in the 10–90° bin, as expected from distillation targets rendered from close, high-overlap pairs; FiT3D and MEF lose in every bin, most at 10–90° (7.6 and 9.6 points); ours holds off-the-shelf recall there. On metric depth, ours gives the largest gain, 11% AbsRel against 3% for SnD, with MEF unchanged and FiT3D well below off-the-shelf. SnD is clearly stronger on semantics however (+7.1 mIoU to our +3.3), plausibly because dense inferred-geometry targets sharpen 2D features more than our sparse LiDAR-covered ones. The two recipes improve different things and, since both leave the architecture unchanged, could be combined, which we leave to future work.

4.4 Ablations

4.4.1 Loss Components

Table 2 isolates the two auxiliary losses. Distillation alone improves mIoU on both backbones, plausibly because matching features fused from other views sharpens the semantics of the covered regions, but it does not improve geometry. Depth falls below the off-the-shelf encoder (AbsRel 0.099 to 0.122 on DINOv2-Base, 0.126 to 0.165 on C-RADIOv4-Huge), as does correspondence. Each auxiliary term repairs part of this. The anchor pulls depth back toward the off-the-shelf level, consistent with a loss that pins uncovered regions to the original features, and the depth head pushes it further through direct metric supervision. On DINOv2-Base either term alone recovers the baseline. On C-RADIOv4-Huge neither does and only their combination surpasses it. The anchor, not the depth head, is what restores correspondence: anchor alone gives the best recall on C-RADIOv4-Huge and is within 0.2 points of the full objective on DINOv2-Base, because matches are drawn from the whole image, most of which receives no distillation signal, and the anchor is the only

Table 2: Loss ablation on KITTI, LoRA rank 8, $\mathcal{L}_{\text{distill}}$ fixed. Off-the-shelf is the reference and excluded from ranking; **bold** is best and underline second best per backbone and column.

Backbone	Variant	$\mathcal{L}_{\text{distill}}$	$\mathcal{L}_{\text{anchor}}$	$\mathcal{L}_{\text{depth}}$	Depth AbsRel↓	Depth δ_1 ↑	Correspondence r@5↑	Correspondence r@10↑	Seg. mIoU↑
DINOv2-Base	Off-the-shelf	–	–	–	0.099	0.910	60.59	73.07	60.37
	Distill only	✓	✗	✗	0.122	0.864	59.78	72.00	62.23
	+ anchor	✓	✓	✗	0.098	0.911	61.27	73.18	62.36
	+ depth	✓	✗	✓	0.089	**0.934**	60.85	72.56	62.49
	+ both (ours)	✓	✓	✓	**0.088**	0.931	**61.46**	**73.34**	**63.64**
C-RADIOv4-Huge	Off-the-shelf	–	–	–	0.126	0.868	62.32	72.45	66.53
	Distill only	✓	✗	✗	0.165	0.801	60.30	71.06	66.88
	+ anchor	✓	✓	✗	0.134	0.842	**63.76**	**73.92**	66.53
	+ depth	✓	✗	✓	0.128	0.867	61.63	72.00	**70.18**
	+ both (ours)	✓	✓	✓	**0.119**	**0.889**	63.00	73.09	67.24

term that constrains those regions. On mIoU the backbones disagree about the anchor, which Section 4.4.3 traces to adaptation capacity. No single term is best everywhere, but the full objective is the only configuration that improves on the off-the-shelf encoder in every metric on both backbones, ranking first or second in every cell.

4.4.2 Distillation versus direct depth supervision

Table 3: Is distillation necessary? "Depth head only" trains the rank-8 student with $\mathcal{L}_{\text{depth}}$ alone; "Ours" is the full loss. Bold is the best adapted row per backbone and column.

Backbone	Variant	Depth AbsRel↓	Depth δ_1 ↑	Corr. r@5↑	Corr. r@10↑	Seg. mIoU↑
DINOv2-Base	Off-the-shelf	0.099	0.910	60.59	73.07	60.37
	Depth head only	**0.076**	**0.951**	60.86	72.21	61.30
	Ours	0.088	0.931	**61.46**	**73.34**	**63.64**
C-RADIOv4-Huge	Off-the-shelf	0.126	0.868	62.32	72.45	66.53
	Depth head only	**0.092**	**0.934**	61.51	71.34	**68.06**
	Ours	0.119	0.889	**63.00**	**73.09**	67.24

Table 3 tests whether the splatted target is needed at all. Depth-only training gives the best depth on both backbones, as expected since the depth probe measures the quantity that loss optimizes, but it does not yield a more broadly 3D-aware representation. It lowers correspondence at 10 px below the off-the-shelf encoder on both backbones, and on C-RADIOv4-Huge at 5 px as well, whereas the full objective improves it at both thresholds on both and beats depth-only supervision in every viewpoint bin but one (Appendix A.2). On segmentation both variants beat the off-the-shelf encoder; the full objective leads by 2.3 mIoU on DINOv2-Base and depth-only by 0.8 on C-RADIOv4-Huge, the same backbone split as Table 2, which Section 4.4.3 explains. A depth head is thus an effective auxiliary signal but an insufficient primary one: it makes depth readily decodable from the features without making them geometrically consistent across views.

4.4.3 Adaptation Capacity

Figure 3 varies the LoRA rank and compares against full fine-tuning. Every LoRA rank improves on the off-the-shelf encoder in every metric on both backbones, but the backbones respond to adaptation capacity in opposite ways. On C-RADIOv4-Huge it helps: AbsRel falls from 0.119 at rank 8 to 0.104 at rank 32 and mIoU rises from 67.24 to 68.66, with correspondence flat. On DINOv2-Base, depth and correspondence barely move while mIoU drops from 63.64 to 61.48. A plausible reason is scale: rank 8 perturbs a 32-block ViT-H far less than a 12-block ViT-B, so the larger model still has headroom where the smaller one has used its budget. Full fine-tuning at the LoRA learning rate of 10^{-4} collapsed every probe on both backbones; at 10^{-5} it improves correspondence slightly on DINOv2-Base but drops it just below the off-the-shelf encoder on C-RADIOv4-Huge, lowers mIoU below off-the-shelf on both backbones, and degrades depth on C-RADIOv4-Huge (Table 5). The anchor loss alone therefore does not hold pretrained semantics once the whole backbone is free to

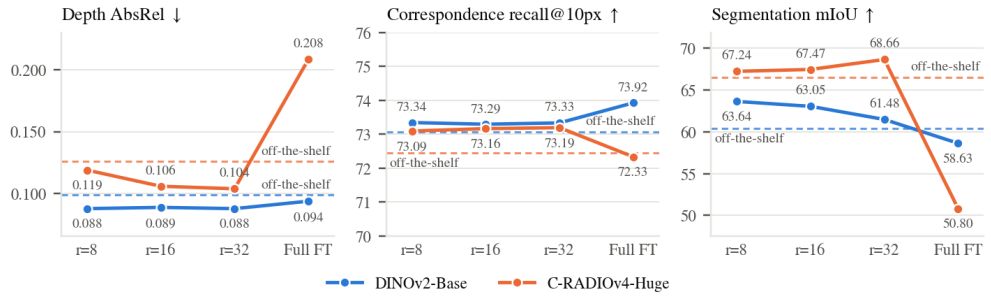

Figure 3: **Adaptation capacity on KITTI.** Full loss, rank varied up to full fine-tuning.

move; the low-rank constraint is also needed. We use rank 8 throughout as the setting at which DINOv2-Base keeps its full mIoU gain, at the cost of some headroom on C-RADIOv4-Huge.

The rank trend also explains the anchor's backbone-dependent effect on mIoU in Sections 4.4.1 and 4.4.2. The anchor holds uncovered regions at their original features. DINOv2-Base loses mIoU as adaptation capacity grows, so the constraint helps; C-RADIOv4-Huge gains mIoU with capacity within the LoRA range, so the same constraint costs it.

5 Discussion and Limitations

Evaluation scope. We evaluate the representation, not downstream systems. Frozen-backbone probes are the established protocol for 3D awareness (El Banani et al., 2024; Yue et al., 2024) and can be reproduced on public data, but how the gains transfer to downstream AV tasks like detection, occupancy, and BEV estimation is a question this paper does not answer.

Baseline comparison scope. The public SnD, FiT3D, and MEF checkpoints share only DINOv2-Base as a backbone and were trained on different data, so Table 1 measures what each released encoder delivers in the driving domain, not a controlled comparison of recipes.

Dependence on LiDAR density. Our coverage numbers (Appendix A.5) are specific to this rig; much denser or sparser LiDAR would change the size of the zero-coverage region, and with it how much the anchor loss matters. The masking mechanism itself is read off per-pixel render confidence and does not depend on density or vertical field of view.

The frozen teacher bounds the semantics. The student gains multi-view consistency over the teacher's semantics but cannot acquire semantics the teacher lacks.

6 Conclusion

We generalize feed-forward splat-and-distill from a short-baseline, indoor setting with estimated geometry to a dynamic, multi-camera outdoor setting with measured geometry. Replacing the stereo network with LiDAR removes the scale ambiguity and the need for stereo overlap, but leaves most of the target view without a distillation target, and distillation alone then degrades the encoder's depth below its off-the-shelf level. A semantic-anchor loss masked to the uncovered region and an auxiliary metric-depth head repair this in complementary ways, and on the larger backbone only their combination surpasses the baseline. Trained on in-house driving logs and evaluated zero-shot on KITTI, the method improves depth and cross-view correspondence for both DINOv2-Base and C-RADIOv4-Huge while also improving a linear segmentation probe, so the 3D gain does not come at the expense of 2D quality. Against the public SnD, FiT3D, and MEF checkpoints it gives the largest depth gain and is the only method that improves correspondence at both thresholds on KITTI-pairs, our driving correspondence benchmark spanning small to large viewpoint changes. All geometry is training-time scaffolding: the deployed encoder is architecturally unchanged, runs on a single image, and uses no LiDAR, which is what a general-purpose vision encoder for both semantic and geometric perception requires.

AI USE STATEMENT

Generative AI tools (a large language model, used through an interactive coding assistant) were used to help draft and organize the text of this paper from the authors' technical notes and design documents, and to help produce LaTeX formatting. Generative AI tools were not used to generate experimental results, run experiments, or fabricate data; all reported and to-be-reported numerical results come from the authors' own experiment tracking. All AI-assisted text has been reviewed and edited by the authors, who take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work trains on an in-house corpus of driving logs collected under the collecting organization's own data-governance and privacy policies; no personally identifying information is used or released as part of this paper, and all public evaluation datasets are used under their respective licenses and terms.

REPRODUCIBILITY STATEMENT

The training corpus is proprietary, and we do not release code or checkpoints, so the training runs themselves cannot be reproduced externally. To make everything else reproducible, the method is fully specified in Section 3: geometry lifting (Section 3.2), fusion and splatting (Section 3.3), the auxiliary depth head (Section 3.4), the coverage-masked anchor loss (Section 3.5), the full loss (Section 3.6), and the adaptation regime, with hyperparameters in Appendix A.4. The evaluation (Section 4.2) uses only public data and the public Probe3D and FiT3D probe codebases with documented modifications, and KITTI-pairs is constructed from KITTI raw by a deterministic, seeded procedure specified in Appendix A.6, so the evaluation protocol can be applied to any encoder.

REFERENCES

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *CVPR*, 2024.

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2GS: Probing visual foundation models with Gaussian splatting. In *CVPR*, 2025.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D Gaussian splatting from sparse multi-view images. In *ECCV*, 2024.

Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. BEVDistill: Cross-modal BEV distillation for multi-view 3D object detection. In *ICLR*, 2023.

Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using Gaussian surfels. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *CVPR*, 2024.

Ravi Garg, Vijay B G Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, and Pavlo Molchanov. RADIO amplified: Improved baselines for agglomerative vision foundation models. *arXiv preprint arXiv:2412.07679*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3D: Probing visual foundation models for complex 3D scene understanding. In *NeurIPS*, 2024.

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.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *ECCV*, 2020.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, 2024.

Mert Bulent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Pau de Jorge, Diane Larlus, and Yannis Kalantidis. DUNE: Distilling a universal encoder from heterogeneous 2D and 3D teachers. In *CVPR*, 2025.

David Shavin and Sagie Benaim. Splat and distill: Augmenting teachers with feed-forward 3D reconstruction for 3D-aware distillation of 2D vision foundation models. *arXiv preprint arXiv:2602.06032*, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien 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 *CVPR*, 2020.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *3DV*, 2017.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024.

Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street Gaussians: Modeling dynamic urban scenes with Gaussian splatting. In *ECCV*, 2024.

Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. EmerNeRF: Emergent spatial-temporal scene decomposition via self-supervision. In *ICLR*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024b.

Yang You, Yixin Li, Congyue Deng, Yue Wang, and Leonidas Guibas. Multiview equivariance improves 3D correspondence understanding with minimal feature finetuning. In *ICLR*, 2025.

Yuanwen Yue, Sayaka Kobayashi, Xingchen Zhu, Francis Engelmann, Marc Pollefeys, and Songyou Tang. Improving 2D feature representations by 3D-aware fine-tuning. In *ECCV*, 2024.

Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. UniDistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *CVPR*, 2023.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D Gaussian splatting to enable distilled feature fields. In *CVPR*, 2024.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. SPA: 3D spatial-awareness enables effective embodied representation. In *ICLR*, 2025.

A Appendix

A.1 Negative Results and Design Alternatives

We record two design alternatives that were implemented, evaluated, and rejected in favor of the choices in Section 3, since they clarify why the adopted design is not the only one that seemed plausible a priori.

EMA teacher. Section 3.1 reports a frozen teacher as the adopted design. An EMA teacher was implemented first, following the convention of prior feed-forward splat-and-distill work, with the DINO-style momentum schedule (Caron et al., 2021) (cosine ramp from 0.996, ending at 0.9999 in our implementation). Because each EMA step keeps only a fraction m_t of the previous teacher, the share of the original checkpoint still present after t steps is $\prod_{i \le t} m_i$. With this schedule it fell below 10^{-3} within about one epoch, after which the teacher was a lagged average of the student's own weights rather than an independent target, and training collapsed. A controlled ablation with teacher, student, and anchor initialized from the same checkpoint reproduced the collapse, isolating the EMA update rule itself as the cause.

Teacher fallback for zero-coverage regions. Section 3.5 reports a masked anchor loss as the adopted fix for coverage sparsity. An earlier alternative – filling zero-coverage pixels with a fixed-confidence, same-view teacher forward pass directly, rather than masking the anchor loss to that region – was implemented and found to regress the correspondence probe in a controlled comparison; it is superseded by the masked anchor loss.

A.2 Correspondence Recall by Viewpoint Difference

Table 4 breaks the KITTI correspondence probe's recall@10px down by the viewpoint difference between the two views of each pair, binned separately by angular and translational difference. It expands the aggregate recall reported in Tables 1 and 3.

Public checkpoints. SnD's no-blending checkpoint leads every bin up to 10° and the 0–1 m bin, then falls 5.9 points below the off-the-shelf encoder at 10–90°; the blending checkpoint, MEF, and FiT3D fall 7.6 to 9.6 in the same bin, and MEF also loses 4.9 at 1–5 m. Ours is the only adapted DINOv2-Base encoder that holds off-the-shelf recall at 10–90°. Beyond 5 m of translation all adapted encoders, ours included, are below off-the-shelf; ours loses the least. Of SnD's two released checkpoints, the no-blending one is the stronger baseline: it beats the blending checkpoint on segmentation (67.4 vs 66.3 mIoU) and in every correspondence bin but 20–100 m, and is within 0.002 AbsRel on depth. Table 1 therefore reports it.

Depth-only versus full objective. The full objective outperforms depth-only supervision in every bin on C-RADIOv4-Huge and in all but the 0–1 m bin on DINOv2-Base, where the two are within 0.1 points. The margins are largest at 1–10° (2.3 to 3.9 points) and smallest where both variants are near ceiling (0–1°, 0–1 m) or near off-the-shelf (10–90°). Against the off-the-shelf encoder, the gains concentrate at moderate viewpoint change: C-RADIOv4-Huge improves at 1–10°, 1–5 m, and 20–100 m, and DINOv2-Base at 1–5 m, while both hold within 0.7 points at 10–90°. Beyond 5 m of translation DINOv2-Base falls 1.3 and 1.4 points below off-the-shelf; depth-only supervision loses 4.2 and 4.1 points in the same bins, so cross-view distillation limits rather than causes the drop.

A.3 Adaptation Capacity: Full Results

Table 5 lists the values plotted in Figure 3, together with δ_1 and recall@5px, which the figure omits because they track AbsRel and recall@10px respectively. Full fine-tuning is reported at two learning rates: the 10^{-4} used by all LoRA runs, at which the backbone drifts from its initialization within the first epoch and every probe collapses, and 10^{-5}. The figure shows the 10^{-5} run.

A.4 Training Configuration

Table 6 lists the resolved configuration of the two runs reported as ours (C-RADIOv4-Huge and DINOv2-Base, LoRA rank 8). All ablation runs share these values except for the factor being ablated. Student, teacher, and anchor are all initialized from the public off-the-shelf weights of the respective backbone.

Table 4: KITTI correspondence recall@10px binned by viewpoint difference between the two views of a pair. "All" is the aggregate value reported in Tables 1 and 3. FiT3D, MEF, and the two SnD checkpoints are the public DINOv2-Base releases (Section 4.2). "Depth head only" trains the LoRA rank-8 student with $\mathcal{L}_{\text{depth}}$ alone; "Ours" is the full objective. **Bold** is best per backbone and column.

Backbone	Variant	All	Angular difference (°)				Translational difference (m)			
			0–1	1–3	3–10	10–90	0–1	1–5	5–20	20–100
DINOv2-Base	Off-the-shelf	73.07	86.60	58.96	56.56	**47.81**	97.06	68.93	**46.65**	**26.69**
	FiT3D	70.02	85.44	55.06	52.76	40.18	96.42	68.14	38.32	22.78
	MEF	69.95	86.61	58.03	54.46	38.19	97.31	64.04	41.78	23.76
	SnD (blending)	71.09	87.45	58.49	55.43	39.27	97.74	67.70	41.22	24.66
	SnD (no blending)	72.58	**87.88**	**59.52**	**57.36**	41.89	**98.13**	70.27	42.98	23.54
	Depth head only	72.21	85.92	55.50	54.22	46.97	97.46	70.14	42.43	22.57
	Ours	**73.34**	86.72	58.43	56.56	**47.81**	97.36	**70.79**	45.31	25.29
C-RADIOv4-Huge	Off-the-shelf	72.45	86.47	55.27	54.67	**45.76**	98.60	70.86	**41.45**	18.20
	Depth head only	71.34	85.85	53.46	52.55	44.51	98.27	68.99	39.86	17.60
	Ours	**73.09**	**86.82**	**56.68**	**56.44**	45.10	**98.63**	**72.78**	41.25	**19.01**

Table 5: Adaptation capacity on KITTI. All rows use the full loss ($\mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{anchor}} + \mathcal{L}_{\text{depth}}$) and vary only the trainable-parameter budget. LoRA alpha is twice the rank. The off-the-shelf row is the no-adaptation reference. Best per backbone and column in bold.

Backbone	Adaptation	Depth		Correspondence		Seg.
		AbsRel↓	δ_1 ↑	r@5↑	r@10↑	mIoU↑
DINOv2-Base	Off-the-shelf	0.099	0.910	60.59	73.07	60.37
	LoRA, rank 8 (ours)	**0.088**	0.931	61.46	73.34	**63.64**
	LoRA, rank 16	0.089	0.932	61.49	73.29	63.05
	LoRA, rank 32	**0.088**	**0.933**	61.57	73.33	61.48
	Full fine-tune, lr 10^{-5}	0.094	0.918	**62.22**	**73.92**	58.63
	Full fine-tune, lr 10^{-4}	0.196	0.749	54.58	66.50	27.86
C-RADIOv4-Huge	Off-the-shelf	0.126	0.868	62.32	72.45	66.53
	LoRA, rank 8 (ours)	0.119	0.889	63.00	73.09	67.24
	LoRA, rank 16	0.106	**0.912**	63.03	73.16	67.47
	LoRA, rank 32	**0.104**	0.908	**63.11**	**73.19**	**68.66**
	Full fine-tune, lr 10^{-5}	0.208	0.709	61.89	72.33	50.80
	Full fine-tune, lr 10^{-4}	0.411	0.394	39.48	49.40	22.71

A.5 How Much of the Target View Receives Distillation Signal?

The distillation loss only supervises target-view patches with projected LiDAR-anchored features whose rendered confidence is nonzero, i.e. patches where a LiDAR surfel exists and is painted by at least one context view. We measure the size of that region on 120 frames, spread evenly over the 40 drives of the training set, building the surfel field from each frame's LiDAR (Section 3.2) and splatting a constant placeholder feature into the target view exactly as during training (Section 3.3); render confidence depends on geometry alone, so this yields the same confidence map as real features would. On average, 76.4% of the target view receives no distillation signal (standard deviation 2.7 points), so the fraction is a property of the sensor rig rather than of particular scenes.

Two factors produce this gap. First, and most importantly, LiDAR sparsity: the LiDAR sweeps a narrow lateral band, so the upper 30–35% of every camera image lies outside its vertical field of view and can never receive a surfel regardless of camera placement; within the band, returns are sparse, thin with range, and are removed by occlusion and back-face rejection. The second factor is the limited field of view shared by adjacent cameras, which leaves some surfels visible from the target but not painted by a context camera. Enlarging the context pool to the two neighbouring frames as well (three frames in total, dynamic points excluded) lowers the unsupervised fraction only from 76.4% to 70.6%, confirming that camera overlap is the smaller contributor. The masked anchor loss (Section 3.5) exists to supervise exactly this uncovered region by tying it to the frozen off-the-shelf encoder

Table 6: Training configuration for the reported runs and evaluation probe heads.

Data and view sampling	
Training corpus	In-house labeled driving set
Timesteps per segment	1; dynamic LiDAR points retained
Context/target pool	Two groups of 3 adjacent cameras; each camera serves as the target once with the other two as context (6 combinations, uniformly sampled)
Global batch	64 segments (8×8 GPUs)
Geometry lifting (Section 3.2)	
Range shells	$[0, 15)$, $[15, 30)$, $[30, 60)$, $[60, 120)$ m
Voxel size per shell	$\mathrm{clip}(0.05 + 0.01\,\bar{r},\ 0.05,\ 1.0)$ m
PCA neighbourhood	$k = 20$ raw points
In-plane extent	$\sqrt{\lambda}$ clipped to $[0.02, 0.40]$ m
Thickness	$\max(0.10\,s_{\text{minor}},\ 0.02)$ m
Surfel budget	300k per segment
Fusion and splatting (Section 3.3)	
Facing weight	$(\mathbf{n} \cdot \mathbf{v})^2$
Occlusion tolerance	0.5 m
Border ramp	16 px
Splat footprint	3σ truncation, 4 px half-window
Losses (Section 3.6)	
$\lambda_{\text{distill}}, \lambda_a, \lambda_d$	1.0, 1.0, 0.5
Depth head (Section 3.4)	
Decoder	DPT on final-layer features, hidden 512
Output	sigmoid-bounded, $[0.001, 80]$ m
Learning rate	4×10^{-4} ($4\times$ backbone LR)
Optimization	
Optimizer	AdamW, LR 10^{-4}, weight decay 10^{-6}
Schedule	3.6% linear warmup, cosine to 10^{-6}
Epochs	16 ($\approx$8.7k steps)
Precision / grad. clip	bf16 / norm 1.0
Adaptation (Section 3.7)	
LoRA rank / alpha / dropout	8 / 16 / 0
Target modules	C-RADIOv4-Huge: QKV + output proj.; DINOv2: query + value
Evaluation probes (Section 4.2), following FiT3D	
Input features	Final-block patch tokens, frozen backbone
Depth head	Linear 1×1 to 256 uniform bins, soft-argmax
Segmentation head	BatchNorm + linear 1×1 classifier, 19 classes
Correspondence	Mutual nearest-neighbour cosine matching with ratio test, top 1000 matches, no trained parameters
Hardware	
Node / wall-clock	$8\ \times$ A100-80GB, DDP; 37 h (C-RADIOv4-Huge), 31 h (DINOv2-Base)

A.6 KITTI-pairs Construction

KITTI-pairs follows the ScanNet-pairs recipe (El Banani et al., 2024): sample frame pairs from posed video and derive correspondence ground truth from known relative pose and depth. This section records the choices specific to KITTI.

Poses. Camera poses are derived from each drive's OxTS GPS/IMU log following the KITTI raw devkit: latitude and longitude are Mercator-projected to a local metric frame with the scale fixed by the first frame's latitude, altitude is used directly, and roll, pitch, and yaw compose the IMU rotation. IMU poses are converted to left-camera poses through the per-date calibration chain (IMU to Velodyne, Velodyne to camera 0, rectification, camera 0 to camera 2). We use OxTS poses rather

than the SLAM-refined poses of the KITTI odometry benchmark because the latter cover only 11 sequences; OxTS poses are accurate to well under a meter and a degree between frames a few seconds apart, which is sufficient for constructing an evaluation set.

Pair sampling. Within each drive on all five capture dates, we sample pairs at frame gaps of $\{1, 3, 5, 10, 20\}$ frames (0.1 to 2.0 s at 10 Hz), up to three pairs per gap per drive, with a fixed random seed; both frames must have LiDAR depth ground truth. Because straight driving dominates KITTI, this base set is concentrated at small rotations, so we add pairs with gaps up to 20 frames sampled evenly across rotation ranges from 1° to 90°, balanced across drives. This yields 2,863 pairs from 102 drives. Relative rotation has mean 8.6° and median 1.5°, with 42% of pairs under 1°; relative translation has mean 6.2 m and reaches 51.9 m (Figure 4). We evaluate across time rather than across KITTI's stereo pair because the stereo pair offers a single small, fixed viewpoint change (0.54 m, no rotation), which cannot show how correspondence degrades as the viewpoint change grows. It also tests generalization, since training uses only different cameras at the same timestamp, never pairs across time.

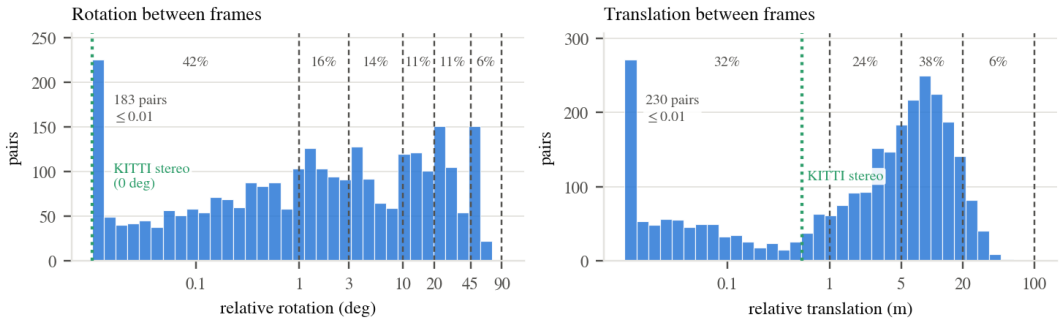

Figure 4: Viewpoint change across the 2,863 KITTI-pairs, log scale. Dashed lines mark the rotation ranges used in sampling and the translation bins of Section A.2, with the share of pairs per range above; the leftmost bar collects pairs with no measurable motion. The green line marks the KITTI stereo pair (0.54 m baseline, rectified), the only same-time cross-camera pair KITTI offers.

Ground-truth filtering. Reprojecting LiDAR depth through the relative pose assumes a static, unoccluded scene, which driving video violates, so we remove unreliable points from both frames with three filters that use only geometry and KITTI's own annotations. (i) *Occlusion:* a point in frame t is kept only if a LiDAR return within 2 px of its projection into t' agrees with its depth to within 10%; an occluded point fails because the return there is closer to the camera. The same test is applied from t' to t. (ii) *Depth edges:* a point is dropped if a LiDAR neighbor within 4 px, in either frame, differs in depth by more than 15%. At such boundaries a match one pixel off lands on a different surface, so the reprojection error is not meaningful. (iii) *Moving objects:* the depth checks above compare depth, not identity, so a point survives them if another object takes the same place at similar depth. Where KITTI provides 3D tracklets (37 drives, 1,139 pairs, all from 2011_09_26), an object counts as moving if its box center shifts by more than 0.5 m between t and t' after removing ego motion, or if it enters or leaves the scene; we drop points on moving objects in either frame and static points they cover in the other frame. Parked and queued vehicles are kept. On the tracklet subset, points on moving objects make up 1.5% of raw ground truth and 0.8% after filters (i) and (ii), which bounds the leakage on unlabeled drives, while 79% of points on static labeled objects survive. Overall, filtering keeps 42% of LiDAR points in frame t and 69% in t'. Figure 5 shows filtered correspondences for five pairs spanning the sampled frame gaps.

Evaluation. For each pair, the frozen backbone extracts dense features from both images. Each image's filtered LiDAR depth is unprojected to a 3D point cloud with the drive's intrinsics, features are matched by nearest neighbor with a ratio test to yield up to 1,000 correspondences per pair, and each correspondence is scored by transforming its 3D point through the known relative pose and measuring the 2D reprojection error against the matched location. Recall is the fraction of correspondences, pooled over all pairs, with error under the pixel threshold. Because most pairs have small rotation, we use rotation bins of $[0, 1, 3, 10, 90]^\circ$ rather than ScanNet-pairs' $[0, 15, 30, 60, 180]^\circ$, which would place most pairs in one bin, and add translation bins of $[0, 1, 5, 20, 100]$ m. Both binnings partition the same set of correspondences from a single evaluation run.



Figure 5: Ground-truth correspondences for five KITTI-pairs at increasing frame gaps (0.1 to 2.0 s, top to bottom). Left: frame t; right: frame t', annotated with the relative translation and rotation from the OxTS poses. Each line links a LiDAR point in frame t to its location in t' under the known relative pose, after the ground-truth filtering described above.

A.7 Additional Qualitative Results

Figures 6–9 show zero-shot KITTI examples for both backbones, off-the-shelf and after our training. No LiDAR is used at inference; the depth and segmentation panels use the probe heads evaluated in Table 1.

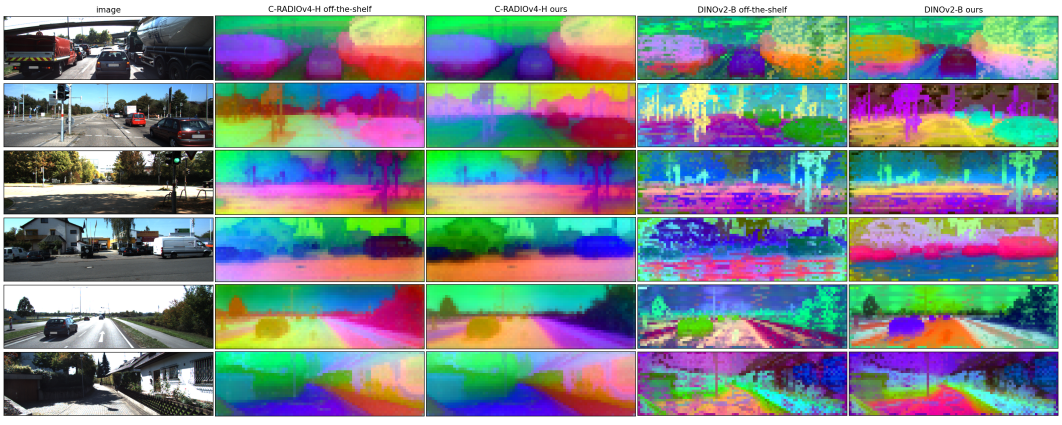

Figure 6: **Patch features on KITTI.** First three principal components of the patch tokens mapped to RGB. After training, features vary smoothly along the road surface and follow object extents, and the high-frequency speckle of off-the-shelf DINOv2-Base is largely removed.

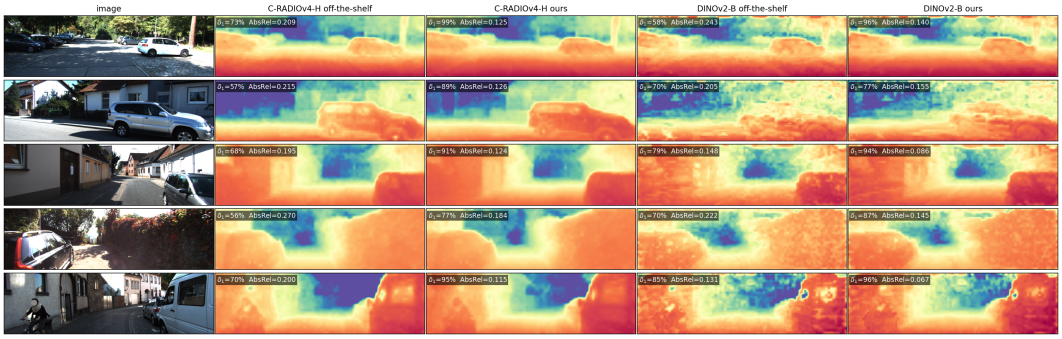

Figure 7: **Depth probe on KITTI.** Probe predictions with per-image δ_1 and AbsRel. Our features give smoother, spatially consistent depth within surfaces and remove the patch-level noise of the off-the-shelf predictions, most visibly for DINOv2-Base; δ_1 and AbsRel improve on every shown image for both backbones.

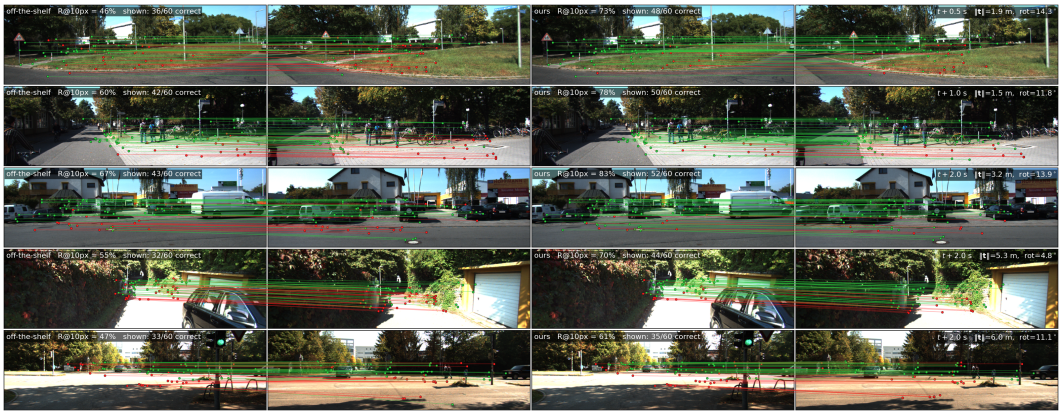

Figure 8: **Correspondence on KITTI pairs (C-RADIOv4-Huge).** Each row is one pair from the test set; the right panel gives the time gap, translation, and rotation. Sixty of the same query points are matched by nearest-neighbour feature lookup; green lines are within 10 px of the ground truth, red lines are not. Off-the-shelf matches drift most on the road surface, where our features recover them. Four of the five pairs fall in the 10–90° rotation bin, the hardest bin for every encoder (Table 4).

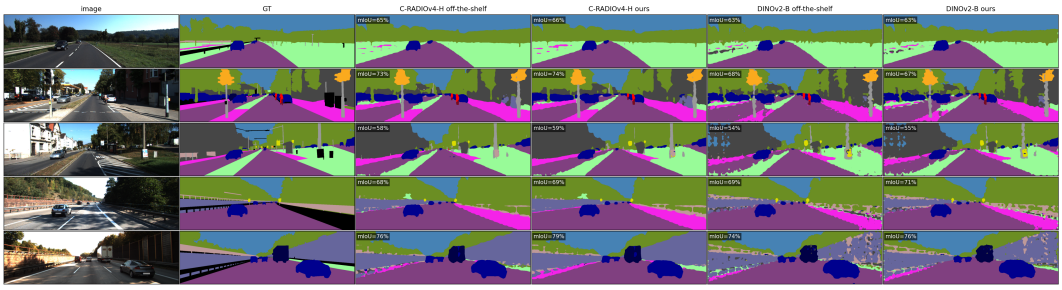

Figure 9: **Semantic segmentation on KITTI.** Ground truth and probe predictions with per-image mIoU. Predictions before and after training are nearly identical, with per-image mIoU within 3 points.